Received: 17 August 2019 | Revised: 1 December 2020 | Accepted: 10 June 2021

DOI: 10.1111/sjos.12548

ORIGINAL ARTICLE

Scandinavian Journal of Statistics

Ultrahigh-dimensional generalized additive model: Unified theory and methods

Kaixu Yang | **Tapabrata Maiti**

Department of Statistics and Probability, Michigan State University, East Lansing, Michigan

Correspondence
Kaixu Yang, Department of Statistics and Probability Michigan State University, 619 Red Cedar Rd. Room 507, East Lansing, MI 48824, USA.
Email: yangkaix@msu.edu

Abstract

Generalized additive model is a powerful statistical learning and predictive modeling tool that has been applied in a wide range of applications. The need of high-dimensional additive modeling is eminent in the context of dealing with high throughput data such as genetics data analysis. In this article, we studied a two-step selection and estimation method for ultrahigh-dimensional generalized additive models. The first step applies group lasso on the expanded bases of the functions. With high probability this selects all nonzero functions without having too much over selection. The second step uses adaptive group lasso with any initial estimators, including the group lasso estimator, that satisfies some regular conditions. The adaptive group lasso estimator is shown to be selection consistent with improved convergence rates. Tuning parameter selection is also discussed and shown to select the true model consistently under generalized information criterion procedure. The theoretical properties are supported by extensive numerical study.

KEYWORDS

adaptive group lasso, generalized additive model, high-dimensional variable selection, selection consistency, tuning parameter selection

1 | INTRODUCTION

The main objective of this work is to establish theory-driven high-dimensional generalized additive modeling method with nonlinear links. The methodology includes convergence rate, variable selection consistency, and tuning parameter selection consistency. Additive models play important roles in nonparametric statistical modeling and machine learning. Although this important statistical learning tool has been used in many important applications and there are free software available for implementing these models along with their variations, to our surprise, there is no literature that has studied the high-dimensional generalized additive model (GAM) with nonidentity link systematically with theoretical foundation. Generalized additive modeling allows nonlinear relationship between a response variable and a set of predictor variables. This general setup includes the special case, namely, the generalized linear models (GLMs), by letting each additive component be a linear function. In general, let $(y_i, \boldsymbol{X}_i), i = 1, \ldots, n$ be independent observations, where y_is are response variables whose corresponding p-dimensional predictor vectors are $\boldsymbol{X}_i$s. A generalized additive model (Hastie & Tibshirani, 1986) is defined as

$$\mu_i = E(y_i|\boldsymbol{X}_i) = g^{-1}\left(\sum_{j=1}^{p_n} f_j(X_{ij})\right), \tag{1}$$

where $g(\cdot)$ is a link function, f_js are unspecified smooth functions, and X_{ij} is the jth component of vector $\boldsymbol{X}_i$. One of the functions could be a constant, which is the intercept term, but this is not necessary. The number of additive components is written as p_n, since it sometimes (usually in high-dimensional setup) increases as n increases. A simple case that many people have studied is $p_n = p$, where the number of additive components is fixed and usually less than the sample size n. The choice of link function is as simple as in GLMs, where people prefer to choose link functions that make the distribution of the response variables belong to the popular exponential family. A widely used generalized additive model has the identity link function $g(\mu) = \mu$, which gives the classical additive model

$$y_i = \sum_{j=1}^{p_n} f_j(X_{ij}) + \epsilon_i, \tag{2}$$

where ϵ_is are i.i.d random variables with mean 0 and finite variance σ^2.

On the other hand, high-dimensional data analysis has become a part of many modern days scientific applications. Often the number of predictors p_n is much larger than the number of observations n, which is usually written as $p_n \gg n$. One of the most interesting scale is p_n increases exponentially as n increases, that is, $\log p_n = O(n^\rho)$ for some constant $\rho > 0$. Fan and Lv (2011) called this as nonpolynomial dimensionality or ultrahigh dimensionality.

In this article, we consider the generalized additive model in a high-dimensional setup. To avoid identification problems, the functions are assumed to be sparse, that is, only a small proportion of the functions are nonzero and all others are exactly zero. A more generalized setup is that the number of nonzero functions, denoted s_n, also diverges as n increases. This case is also considered in this article.

Many others have worked on generalized additive models. Common approaches use basis expansion to deal with the nonparametric functions, and perform variable selection and

estimation methods on the bases. Meier et al. (2009) considered a simpler case (2), with a new sparsity-smoothness penalty and proved its oracle property. They also performed a simulation study under logit link with their new penalty; however, no theoretical support was provided. Huang et al. (2010) focused on the variable selection of (2) with fixed number of nonzero functions and identity link function using a two-step approach: first group lasso (Bakin, 1999; Yuan & Lin, 2006) on the bases to select the nonzero predictors and then use adaptive group lasso to estimate the bases coefficients. They then established the selection consistency and provided the rate of convergence of the estimation. Fan et al. (2011) proposed the nonparametric independence screening (NIS) method in screening the model (2). However, the selection consistency and the generalized link functions were not discussed. Marra and Wood (2011) discussed the practical variable selection in additive models, but not in the high-dimensional setup. Amato et al. (2016) reviewed several existing algorithms highlighting the connections between them, including the nonnegative garrote, COSSO, and adaptive shrinkage, and presented some computationally efficient algorithms for fitting the additive models. Nandy et al. (2017) extended the consistency and rate of convergence of Huang et al. (2010) to spatial additive models. Fan and Zhong (2018) studied the GAM with identity link under the endogeneity setting. All work above are based on model (2). For GAM, that is, model (1), Tutz and Binder (2006) studied fitting GAM and perform variable selection implicitly through likelihood based boosting. Later Liu et al. (2013) considered a two-step oracally efficient approach in generalized additive models in the low-dimensional setup, but no variable selection in the high-dimensional setup was done.

However, though widely used, no systematic theory about selection and estimation consistency or rate of convergence has been established for generalized additive models with nonidentity link functions in the high-dimensional setup.

In this article, we establish the theory part for generalized additive models with nonidentity link functions in high-dimensional setup. We develop a two-step selection approach, where in the first step we use group lasso to perform a screening, which, under mild assumptions, is able to select all nonzero functions and not overselect too much. In the second step, the adaptive group lasso procedure is used and is proved to select the true predictors consistently.

Another important practical issue in variable selection and penalized optimization problems is tuning parameter selection. Various cross-validation techniques have been used in practice for a long time. Information criteria such as Akaike information criterion (AIC), AICc, Bayesian information criterion (BIC), Mallow's C_p, and so on have been used to select "the best" model as well. Many equivalences among the tuning parameter selection methods have been shown in the Gaussian linear regression case. However, the consistency of these selection methods were not established. Later some variations of the information criteria such as modified BIC (Wang et al., 2009; Zhang & Siegmund, 2007) extended BIC (Chen & Chen, 2008) and generalized information criterion (GIC) (Fan & Tang, 2013) were proposed and shown to have good asymptotic properties in penalized linear models and penalized likelihoods. However, the results are not useful for grouped variables in additive models, for which basis expansion technique is usually used and thus brings grouped selection.

In this article, we generalize the result of GIC by Fan and Tang (2013) to group-penalized likelihood problems and show that under some common conditions and with a good choice of the parameter in GIC, we are able to select the tuning parameter that corresponds to the true model.

In Section 2, the model is specified and basic approach is discussed. Notations and basic assumptions are also introduced in this section. Section 3 gives the main results of the two-step

selection and estimation procedure. Section 4 develops the tuning parameter selection. Extensive simulation study and real data example are presented in Section 5 followed by a short discussion in Section 6. The proofs of all theorems are deferred to supplementary materials.

2 | MODEL

We consider the generalized additive model (1) with the link function corresponding to an exponential family distribution of the response. For each of the n independent observations, the density function is given as

$$f_{y_i}(y) = c(y) \exp\left[\frac{y\theta_i - b(\theta_i)}{\phi}\right], \quad 1 \le i \le n, \ \theta_i \in \mathbb{R}. \tag{3}$$

Without loss of generality, we assume that the dispersion parameter $0 < \phi < \infty$ is assumed to be a known constant. Specifically we assume $\phi = 1$. We consider a fixed design throughout this article, that is, the design matrix $\boldsymbol{X}$ is assumed to be fixed. However, we have shown in Appendix A that the same theory works for a random design under simple assumptions on the distribution of $\boldsymbol{X}$. The additive relationship assumes that the densities of y_is depend on $\boldsymbol{X}_i$s through the additive structure $\theta_i = \sum_{j=1}^{p_n} f_j(X_{ij})$. This is the canonical link. If we use other link functions, for example, $A(\cdot)$, the theory also works as long as the functions $A(\cdot)$ satisfies the Lipschitz conditions for some order. Let $b^{(k)}(\cdot)$ be the kth derivative of $b(\cdot)$, then by property of the exponential family, the expectation and variance matrix of $\boldsymbol{y} = (y_1, \ldots, y_n)^T$, under mild assumptions of $b(\cdot)$, is given by $\boldsymbol{\mu}(\boldsymbol{\theta})$ and $\phi\Sigma(\boldsymbol{\theta})$, where

$$\boldsymbol{\mu}(\boldsymbol{\theta}) = (b^{(1)}(\theta_1), \ldots, b^{(1)}(\theta_n))^T \quad \text{and} \quad \Sigma(\boldsymbol{\theta}) = \text{diag}\{b^{(2)}(\theta_1), \ldots, b^{(2)}(\theta_n)\}. \tag{4}$$

The log-likelihood (ignoring the term $c(y)$ which is not interesting to us in parameter estimation) can be written as

$$l = \sum_{i=1}^{n} \left[y_i \left(\sum_{j=1}^{p_n} f_j(X_{ij}) \right) - b\left(\sum_{j=1}^{p_n} f_j(X_{ij}) \right) \right]. \tag{5}$$

Assume that the additive components belong to the Sobolev space $W_2^d([a, b])$. According to Schumaker (1981), see pages 268–270, there exists B-spline approximation

$$f_{nj}(x) = \sum_{k=1}^{m_n} \beta_{jk}\phi_k(x), \quad 1 \le j \le p \tag{6}$$

with $m_n = K_n + l$, where K_n is the number of internal knots and $l \ge d$ is the order of the splines. Generally, it is recommended that $d = 2$ and $l = 4$, that is, cubic splines.

Using the approximation above, Huang et al. (2010) proved that f_{nj} well approximates f_j in the sense of rate of convergence that

$$\|f_j - f_{nj}\|_2^2 = \int_a^b (f_j(x) - f_{nj}(x))^2 dx = O(m_n^{-2d}). \tag{7}$$

Therefore, using the basis approximation, the log-likelihood can be written as

$$l = \sum_{i=1}^{n} \left[y_i \left(\sum_{j=1}^{p_n} \sum_{k=1}^{m_n} \beta_{jk}^0 \Phi_k(x_{ij}) \right) - b \left(\sum_{j=1}^{p_n} \sum_{k=1}^{m_n} \beta_{jk}^0 \Phi_k(x_{ij}) \right) \right] = \sum_{i=1}^{n} \left[y_i \left(\boldsymbol{\beta}^{0T} \Phi_i \right) - b \left(\boldsymbol{\beta}^{0T} \Phi_i \right) \right], \tag{8}$$

where $\boldsymbol{\beta}^0$ and Φ_i are the vector basis coefficients and bases defined below.

It is also worth noting that the number of bases m_n increases as n increases. This is necessary since Schumaker (1981) mentioned that one need to have sufficient partitions to well approximate f_j by f_{nj}. If we fix m_n, that is, let $m_n = m_0$, though in the later part we will show the approach to estimate the basis coefficients can have better rate of convergence, the approximation error between the additive components and the spline functions $\|f_j(x) - f_{nj}(x)\|_2 = [\int_a^b (f_j(x) - f_{nj}(x))^2 dx]^{1/2} = O(1)$ will increase and lead to inconsistent estimations. Therefore, m_n, or more precisely, K_n, need to increase with n.

Our selection and estimation approach will be based on the bases approximated log likelihood (8). Before starting the methodology, we list the notations and state the assumptions we need in this article.

2.1 | Notations

The design matrix is $\boldsymbol{X}_{(n\times p_n)} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$. The basis matrix is $\Phi_{(n\times m_n p_n)} = (\Phi_1, \ldots, \Phi_n)^T$, where $\Phi_i = (\phi_1(x_{i1}), \ldots, \phi_{m_n}(x_{i1}), \ldots, \phi_1(x_{ip_n}), \ldots, \phi_{m_n}(x_{ip_n}))^T$.

The true basis parameters are $\boldsymbol{\beta}^0 = (\beta_{11}^0, \ldots, \beta_{1m_n}^0, \ldots, \beta_{p_n1}^0, \ldots, \beta_{p_nm_n}^0)^T \in \mathbb{R}^{m_n p_n}$

We assume the functions $f_1, \ldots, f_{p_n}$ are sparse, then $\boldsymbol{\beta}^0$ is blockwise sparse, that is, the blocks ${\boldsymbol{\beta}^0}_1 = (\beta_{11}^0, \ldots, \beta_{1m_n}^0)^T, \ldots, {\boldsymbol{\beta}^0}_{p_n} = (\beta_{p_n1}^0, \ldots, \beta_{p_nm_n}^0)^T$ are sparse.

Let $\boldsymbol{\mu}_y$ be the expectation of $\boldsymbol{y}$ based on the true basis parameters and $\boldsymbol{\varepsilon} = \boldsymbol{y} - \boldsymbol{\mu}_y$.

Define the relationship $a_n \preccurlyeq b_n$ as there exists a finite constant c such that $a_n \le c b_n$.

For any function f define $\|f\|_2 = [\int_a^b f^2(x)dx]^{1/2}$, whenever the integral exists.

For any two collections of indices $S, \tilde{S} \subseteq \{1, \ldots, p_n\}$, the difference set is denoted $S - \tilde{S}$. The cardinality of S is denoted card(S). For any $\boldsymbol{\delta} \in \mathbb{R}^{m_n p_n}$, define $\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_{p_n}$ as its subblocks, where $\boldsymbol{\delta}_i \in \mathbb{R}^{m_n}$, and define the blockwise support

$$\text{supp}_B(\boldsymbol{\delta}) = \{j \in \{1, \ldots, p_n\}; \boldsymbol{\delta}_j \neq \boldsymbol{0}\}.$$

Define the blockwise cardinality $\text{card}_B(\boldsymbol{\delta}) = \text{card}(\text{supp}_B(\boldsymbol{\delta}))$.

For $S = \{s_1, \ldots, s_q\} \subseteq \{1, \ldots, p_n\}$, define subblock vector $\boldsymbol{\delta}_S = (\boldsymbol{\delta}_{s_1}^T, \ldots, \boldsymbol{\delta}_{s_q}^T)^T$.

The number of additive components is denoted p_n, which is possible to grow faster than the sample size n. Let $T = \text{supp}_B(\boldsymbol{\beta}^0)$ and T^c be the complementary set. Let $\text{card}(T) = s_n$, where s_n is allowed to diverge slower than n.

For each $U \subseteq \{1, \ldots, p_n\}$ with $\text{card}(U - T) \le m$ for some m, define

$$\mathcal{B}(U) = \{\boldsymbol{\delta} \in \mathbb{R}^{m_n p_n}; \text{supp}_B(\boldsymbol{\delta}) \subseteq U\},$$
$$\mathcal{B}(m) = \{\mathcal{B}(U); \text{for any } U \subseteq \{1, \ldots, p_n\}; \text{Card}(U - T) \le m\}.$$

Let q be an integer such that $q > s_n$ and $q = o(n)$. Define

$$\mathcal{B}_1 = \{\boldsymbol{\beta} \in \mathcal{B} : \text{card}_B(\boldsymbol{\beta}) \le q\},$$

where $\mathcal{B}$ is a sufficiently large, convex, and compact set in $\mathbb{R}^{p_n m_n}$.

2.2 | Assumptions

Assumption 1 (On design matrix). Using the normalized B-spline bases, the basis matrix Φ has each covariate vector $\Phi_j, j = 1, \ldots, m_n p_n$ bounded, that is, $\exists\ c_\Phi$ such that $\|\Phi_j\|_2 \leq \sqrt{n} c_\Phi, \forall j = 1, \ldots, m_n \times p_n$.

Assumption 2 (Restricted eigenvalues RE). For a given sequence N_n, there exist γ_0 and γ_1 such that

$$\gamma_0 \gamma_2^{2s_n} m_n^{-1} \leq \frac{\boldsymbol{\delta}^T \Phi^T \Phi \boldsymbol{\delta}}{n \|\boldsymbol{\delta}\|_2^2} \leq \gamma_1 m_n^{-1}, \tag{9}$$

where γ_2 is a positive constant such that $0 < \gamma_2 < 0.5$, for all $\boldsymbol{\delta} \in C$, where $\boldsymbol{\delta}^T = (\boldsymbol{\delta}_1^T, \ldots, \boldsymbol{\delta}_{p_n}^T)$ and

$$C = \left\{ \boldsymbol{\delta} \in \mathbb{R}^{p_n m_n} : \|\boldsymbol{\delta}\|_2 \neq 0,\ \|\boldsymbol{\delta}\|_2 \leq N_n \text{ and } \text{card}_B(\boldsymbol{\delta}) = o(s_n) \right\}. \tag{10}$$

Assumption 3 (On the exponential family distribution). The function $b(\theta)$ is three times differentiable with $c_1 \leq b''(\theta) \leq c_1^{-1}$ and $|b'''(\theta)| \leq c_1^{-1}$ in its domain for some constant $c_1 > 0$. For unbounded and non-Gaussian distributed Y_i, there exists a diverging sequence $M_n = o(\sqrt{n})$ such that

$$\sup_{\boldsymbol{\beta} \in B_1} \max_{1 \leq i \leq n} \left| b' \left(\left| \Phi_i^T \boldsymbol{\beta} \right| \right) \right| \leq M_n. \tag{11}$$

Additionally the error term $\epsilon_i = y_i - \mu_{y_i}$s follow the uniform sub-Gaussian distribution, that is, there exist constants $c_2 > 0$ such that uniformly for all $i = 1, \ldots, n$, we have

$$P(|\epsilon_i| \geq t) \leq 2 \exp(-c_2 t^2) \quad \text{for any } \ t > 0. \tag{12}$$

Assumption 4 (On nonzero function coefficients). There exist a sequence $c_{f,n}$ that may tend to zero as $n \to \infty$ such that for all $j \in T$, the true nonzero functions f_j satisfy

$$\min_{j \in T} \|f_j\|_2 \geq c_{f,n}.$$

We note that Assumption 1 is a standard assumption in high-dimensional models, where the design matrix needs to be bounded from above. Assumption 2 is a well-known condition in high-dimension setup on the empirical Gram matrix (Bickel et al., 2009). It is different from the regular eigenvalue condition, since when $n < p$, the $p \times p$ Gram matrix has rank less than p, thus it must have zero eigenvalues. Therefore, it is not realistic to bound the eigenvalues away from zero for all $\boldsymbol{\nu} \in \mathbb{R}^{p_n m_n}$, but we need to restrict to some space C. In our setup, C is the restricted subblock eigenvalue condition on subblocks of the Gram matrix studied by Belloni and Chernozhukov (2013). Though the lower bound and upper bound are imposed on the fixed design matrix, we gave a derivation in supplementary materials that this condition holds when $\boldsymbol{X}$ is drawn from a continuously differentiable density function which is bounded away from 0 and infinity on the domain of $\boldsymbol{X}$. This result is similar to the results in Huang et al. (2010).

Assumption 3 is a standard assumption to generalized models. Equations (11) and (12) together control the tail behavior of the responses, and as mentioned by Fan and Tang (2013), ensure a general and broad applicability of the method. Analogous assumptions to (11) can also be seen in Fan and Song (2010) and Bühlmann and van de Geer (2011). Specifically, for example,

for GAM with logit link function, we have $b(\theta) = \log(1 + \exp(\theta))$. It is easy to verify that both its second and third derivatives have their absolute values all bounded from above by 1. For Equation (11), observe that the first derivative is the mean of Bernoulli distribution, and thus it is also bounded. The error term is also bounded by 1, therefore, taking $c_2 = \log(2)$ will make Equation (12) satisfy all logistic regression cases. Moreover, bounded second moment in logistic regression ensure that there exists ϵ such that the probability p_i of each observation satisfies $\epsilon < p < 1 - \epsilon$.

Assumption 4 appears often in variable selection methodologies, because intuitively a nonzero function or covariate has to contribute enough to the response in order to be considered nonzero.

Remark 1. In Assumption 2, $\boldsymbol{\delta} = \boldsymbol{\beta} - \boldsymbol{\beta}_0$ is the difference vector between a $\boldsymbol{\beta}$ and the true coefficients $\boldsymbol{\beta}_0$, thus we can view C as a restricted neighborhood of $\boldsymbol{\beta}_0$, that is,

$$\mathcal{N}^{\mathrm{RE}}_{\boldsymbol{\beta}_0} = \left\{ \boldsymbol{\beta} : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2 \le N_n,\ m_n \times \mathrm{card}_B(\boldsymbol{\delta}) \le n^* = o(n) \right\}.$$

If $\boldsymbol{\beta} \in \mathcal{N}^{\mathrm{RE}}_{\boldsymbol{\beta}_0}$, then by Assumption 2 we have

$$\frac{(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T \Phi^T \Phi (\boldsymbol{\beta} - \boldsymbol{\beta}_0)}{n\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2^2} \ge \gamma_0 \gamma_2^{2s_n} m_n^{-1}.$$

This, together with the bounded variance assumption in Assumption 3, ensures the restricted strong convexity of the target function, that is, for a $\boldsymbol{\beta}^* \in \mathcal{N}^{\mathrm{RE}}_{\boldsymbol{\beta}_0}$, we have

$$\frac{(\boldsymbol{\beta}^* - \boldsymbol{\beta}_0)^T \Phi^T \boldsymbol{\Sigma}(\boldsymbol{\beta}) \Phi (\boldsymbol{\beta}^* - \boldsymbol{\beta}_0)}{n\|\boldsymbol{\beta}^* - \boldsymbol{\beta}_0\|_2^2} \ge \gamma_0 c_1 \gamma_2^{2s_n} m_n^{-1}, \quad \forall\ \boldsymbol{\beta} \in \mathcal{N}^{\mathrm{RE}}_{\boldsymbol{\beta}_0}. \tag{13}$$

3 | METHODOLOGY AND THEORETICAL PROPERTIES

We propose a two-step procedure for selecting high-dimensional additive models with generalized link that has improved convergence rates compared with a single stage selection.

3.1 | First step: Model screening

The objective of this step is to recover the true support T of the additive components. Let $\hat{T}$ be a random support given by a model selection procedure and $|\hat{T}|$ be the number of variables selected. A good model selection procedure should satisfy the common screening consistency conditions

$$T \subset \hat{T},\ |\hat{T}| = O(s_n),\ \text{w.p. converging to 1.} \tag{14}$$

There have been many variable selection penalization (Fan & Lv, 2011; Fan & Peng, 2004; Fan & Song, 2010; Van de Geer, 2008) in GLMs and (Huang et al., 2010) in linear additive models where this condition holds. Specifically, Fan and Song (2010) satisfies the requirements in (14) in GLMs and Huang et al. (2010) also satisfies (14) in additive models. In this article, we show that under mild conditions, by maximizing the log-likelihood with group lasso-like penalization,

we can select a model that satisfies (14). We also provide a rate of convergence of this first step selection.

Define the objective function to be

$$L(\boldsymbol{\beta};\lambda_{n1}) = -\frac{1}{n}\sum_{i=1}^{n}\left[y_i\left(\boldsymbol{\beta}^T\Phi_i\right) - b\left(\boldsymbol{\beta}^T\Phi_i\right)\right] + \lambda_{n1}\sum_{j=1}^{p_n}\|\boldsymbol{\beta}_j\|_2. \tag{15}$$

Let $\hat{\boldsymbol{\beta}}$ be the optimizer for (15), that is,

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^{p_n m_n}}{\arg\min}\, L(\boldsymbol{\beta};\lambda_{n1}).$$

Let $\hat{T} = \text{supp}_B(\hat{\boldsymbol{\beta}})$.

The objective function is the negative log-likelihood plus the group lasso penalization term, and the parameters are estimated as the minimizers of the objective function. Here the negative log likelihood function is averaged among the n observations to ensure that it is under the same scale as the penalization function.

With this group lasso type penalized log-likelihood, the selected model has the following properties.

Theorem 1. *Consider the model $\hat{T}$ obtained by minimizing (15). Under Assumptions 1–4, for some constant C and any diverging sequence $\gamma_n > 0$, choose the regularization parameter*

$$\lambda_{n1}^{b} = C\sqrt{m_n}\sqrt{\frac{\gamma_n + \log(p_n m_n)}{n}}$$

for bounded response (i.e., $|y_i| < c$), and the regularization parameter

$$\lambda_{n1}^{\text{ub}} = \sqrt{m_n}\gamma_n\sqrt{\frac{\log(p_n m_n)}{n}}$$

for unbounded sub-Gaussian response, as the sample size increases,

(i) *With probability tending to 1,*

$$|\hat{T}| = O(s_n).$$

(ii) *With probability tending to 1,*

$$\sum_{j=1}^{p_n}\left\|\boldsymbol{\beta}_j^0 - \hat{\boldsymbol{\beta}}_j\right\|_2^2 = O_P\left(s_n\gamma_2^{-2s_n}\frac{m_n^2\log(p_n m_n)}{n}\right) + O(\lambda_{n1}^{b\,2}m_n^2 s_n\gamma_2^{-2s_n}) + O(s_n^2 m_n^{1-2d}\gamma_2^{-2s_n})$$

for the bounded response and

$$\sum_{j=1}^{p_n}\left\|\boldsymbol{\beta}_j^0 - \hat{\boldsymbol{\beta}}_j\right\|_2^2 = O_P\left(s_n\gamma_2^{-2s_n}\gamma_n\frac{m_n^2\log(p_n m_n)}{n}\right) + O(\lambda_{n1}^{\text{ub}\,2}m_n^2 s_n\gamma_2^{-2s_n}) + O(s_n^2 m_n^{1-2d}\gamma_2^{-2s_n})$$

for any diverging sequence γ_n and unbounded sub-Gaussian response.

(iii) *If* $s_n\gamma_2^{-2s_n}m_n^2\log(p_nm_n)/n << c_{f,n}$ ($s_n\gamma_2^{-2s_n}m_n^2\gamma_n\log(p_nm_n)/n << c_{f,n}$ *in the unbounded case),* $\gamma_2^{-2s_n}m_n^2\lambda_{n1}^2 s_n/m_n << c_{f,n}$ *and* $s_n^2 m_n^{1-2d}\gamma_2^{-2s_n} << c_{f,n}$*, with probability tending to 1, all nonzero coefficients are selected.*

The proof of this theorem is given in supplementary materials.

Remark 2. In practice, for a specific data set, to avoid estimability issues, the constants C is selected to be large enough such that the number of parameters to be estimated, that is, the number of selected nonzero functions $|\hat{T}|$ multiplied by the number of basis function m_n should be less than or equal to n. Moreover, considering the multicollinearity in the design matrix, the constants are chosen such that $m_n \times |\hat{T}| << n$.

Remark 3. The additional term γ_n in the convergence rate is due to unboundedness nature of the response variable rather than due to nonlinear link function.

Remark 4. For the special case, linear (Gaussian) additive model, our results coincide with Huang et al. (2010). The difference is that we study a fixed design with assumptions on the eigenvalues of the design matrix and they studied a random design with assumption on the distribution of the design matrix. We have put further assumption on the eigenvalue due to the divergence of s_n, the number of nonzero variables. In the special case that s_n is fixed, our assumptions coincides with the assumptions in Huang et al. (2010). Another difference is that we include a diverging term γ_n that establishes the rate of convergence with probability converging to one.

There are three terms in the convergence rate: the first term comes from the regression itself, the second term comes from shrinkage, and the third term comes from the spline approximation error.

Remark 5. Let $\hat{f}_{nj}(x) = \sum_{k=1}^{m_n}\hat{\beta}_{jk}\phi_k(x)$. We can also state the results of the first selection step in terms of functions, which is a direct consequence of theorem 1. First, we have (i) $|\hat{T}| = O(s_n)$ with probability tending to 1, and (ii) if $s_nm_n\gamma_2^{-2s_n}\log(p_nm_n)/n << c_{f,n}$ ($s_nm_n\gamma_2^{-2s_n}\gamma_n\log(p_nm_n)/n << c_{f,n}$ in the unbounded case), $\lambda_{n1}^2 s_nm_n\gamma_2^{-2s_n} << c_{f,n}$ and $s_n^2m_n^{-2d}\gamma_2^{-2s_n} << c_{f,n}$, with probability tending to 1, all nonzero coefficients are selected.

Moreover, by the properties of spline in De Boor (2001), see, for example, Stone (1986) and Huang et al. (2010), there exist positive constants c_1 and c_2 such that

$$c_1m_n^{-1}\|\hat{\boldsymbol{\beta}}_{nj}-\boldsymbol{\beta}_{nj}\|_2^2 \le \|\hat{f}_{nj}-f_{nj}\|_2^2 \le c_2m_n^{-1}\|\hat{\boldsymbol{\beta}}_{nj}-\boldsymbol{\beta}_{nj}\|_2^2, \tag{16}$$

we have

$$\sum_{j=1}^{p_n}\left\|f_j-\hat{f}_{nj}\right\|_2^2 = O_P\left(s_n\gamma_2^{-2s_n}\frac{m_n\log(p_nm_n)}{n}\right) + O(\lambda_{n1}^{b\,2}m_ns_n\gamma_2^{-2s_n}) + O(s_n^2m_n^{-2d}\gamma_2^{-2s_n})$$

for the bounded response case and

$$\sum_{j=1}^{p_n}\left\|f_j-\hat{f}_{nj}\right\|_2^2 = O_P\left(s_n\gamma_2^{-2s_n}\gamma_n\frac{m_n\log(p_nm_n)}{n}\right) + O(\lambda_{n1}^{\mathrm{ub}\,2}m_ns_n\gamma_2^{-2s_n}) + O(s_n^2m_n^{-2d}\gamma_2^{-2s_n})$$

for the unbounded case, for any diverging sequence γ_n.

Remark 6. The theorem and its remark together tell us that under Assumptions 1–4, by choosing proper γ_n, the functions selected by minimizing the first target function satisfy

$$T \subset \hat{T} \quad \text{and} \quad |\hat{T}| = O(s_n)$$

with probability converging to 1, that is, we obtained screening consistency.

3.2 | Second step: Postselection

After we have a "good" initial estimator, we use the adaptive group lasso to recover the true model (Huang et al., 2010) and we are able to achieve selection consistency in probability under some mild assumptions. The adaptive group lasso idea is similar to adaptive lasso (Zou, 2006) which enjoys better theoretical properties than simple lasso. Chatterjee and Lahiri (2013) and Das et al. (2017) studied rate of convergence and other asymptotic properties of the adaptive lasso estimator. Define the objective function to be

$$L_a(\boldsymbol{\beta}; \lambda_{n2}) = -\frac{1}{n}\sum_{i=1}^{n} \left[y_i \left(\boldsymbol{\beta}^T \Phi_i\right) - b\left(\boldsymbol{\beta}^T \Phi_i\right)\right] + \lambda_{n2} \sum_{j=1}^{p_n} w_{nj} \|\boldsymbol{\beta}_j\|_2, \tag{17}$$

where the weights depend on the screening stage group lasso estimator

$$w_{nj} = + \begin{cases} \|\hat{\boldsymbol{\beta}}_j\|_2^{-1}, & \text{if } \|\hat{\boldsymbol{\beta}}_j\|_2 > 0 \\ \infty, & \text{if } \|\hat{\boldsymbol{\beta}}_j\|_2 = 0 \end{cases}. \tag{18}$$

Let $\hat{\boldsymbol{\beta}}_{\text{AGL}}$ be the optimizer for (17), that is,

$$\hat{\boldsymbol{\beta}}_{\text{AGL}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^{m_n p}}{\arg\min}\, L_a(\boldsymbol{\beta}; \lambda_{n2}).$$

For the choice of weights, the first stage estimators need not to be necessarily the solution of group lasso, but could be more general estimators that satisfy following assumptions.

Assumption 5. The initial estimator $\hat{\boldsymbol{\beta}}$ is r_n consistent at zero, that is,

$$r_n \max_{j \in T^c} \|\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^0\|_2 = O_P(1), \tag{19}$$

and there exists a constant c_3 such that

$$\mathbb{P}\left(\min_{j \in T} \|\hat{\boldsymbol{\beta}}_j\|_2 \geq c_3 b_{n1}\right) \to 1, \tag{20}$$

where $b_{n1} = \min_{j \in T} \|\boldsymbol{\beta}_j^0\|_2$.

Assumption 6. Let $s_n^* = p_n - s_n$ be the number of zero components. The tuning parameter λ_{n2} satisfies

$$\frac{\sqrt{\log(s_n^* m_n)}}{n^{1/2} \lambda_{n2} r_n} + \frac{s_n}{\lambda_{n2} r_n m_n^{d+1/2}} + \frac{\lambda_{n2} r_n}{\gamma_n \sqrt{s_n/n}} = o(1) \tag{21}$$

for any diverging sequence γ_n.

Assumption 5 gives the restrictions on the initial estimator. We do not require our initial estimator to be the group lasso estimator. Any initial estimator satisfying Assumption 5 will be able to make the adaptive group lasso estimator consistently selects and estimates the true nonzero components. However, the rate of convergence of the adaptive group lasso estimator depends on the rate of convergence of the initial estimator, which is assumed to be r_n in Assumption 5. Moreover, the initial estimator must not have a 0 estimation for the nonzero components, otherwise it will mislead the results in the proceeding step. Assumption 6 put restrictions on the tuning parameter λ_{n2} in the adaptive group lasso step. The first two terms gives the lower bound for λ_{n2} and the third term gives the upper bound. Only with "appropriate" choice of λ_{n2} we can have the selection consistency and estimation consistency.

It is worth noting that if we take the group lasso estimator as our initial estimator, Assumptions 5 and 6 are automatically satisfied. Specifically, a trivial choice of r_n would be

$$r_n = O_P^{-1}\left(\sqrt{s_n}\gamma_2^{-s_n}\frac{m_n\sqrt{\log(p_n m_n)}}{\sqrt{n}}\right) + O^{-1}(\lambda_{n1}^b m_n\sqrt{s_n}\gamma_2^{-s_n}) + O^{-1}(s_n m_n^{0.5-d}\gamma_2^{-s_n})$$

for the bounded response and

$$r_n = O_P^{-1}\left(\sqrt{s_n}\gamma_2^{-s_n}\sqrt{\gamma_n}\frac{m_n\sqrt{\log(p_n m_n)}}{\sqrt{n}}\right) + O^{-1}(\lambda_{n1}^{\text{ub}} m_n\sqrt{s_n}\gamma_2^{-s_n}) + O^{-1}(s_n m_n^{0.5-d}\gamma_2^{-s_n})$$

for the unbounded case and any diverging sequence γ_n, since we observe that for $j \in T^c$, $\hat{\boldsymbol{\beta}}_j$ is either estimated as zero, or has a rate of convergence to $\boldsymbol{\beta}_j$ bounded by the rate of convergence in theorem (1). For Equation (20), we observe that the rate of convergence of the group lasso estimator is higher-order infinitesimal of the minimal signal strength of nonzero coefficients, thus taking $c_3 = 0.5$ is sufficient. In Assumption 6, with our trivial choice of r_n, we are able to find a range of tuning parameters that satisfy Equation (21). Therefore, it is reasonable to take the group lasso estimator as an initial estimator for the adaptive group lasso.

Let the notation $\hat{\boldsymbol{\beta}}_n \overset{0}{=} \boldsymbol{\beta}^0$ denote that the sign of each $\hat{\boldsymbol{\beta}}_j$ and $\boldsymbol{\beta}_j^0$ are either both zero or both nonzero. Then we have the following asymptotic properties for the adaptive group lasso estimator.

Theorem 2. *Assume Assumptions 1–6 hold, consider the estimator $\hat{\boldsymbol{\beta}}_{\text{AGL}}$ by minimizing (17), we have*

(i) If $c_{f,n} >> \sqrt{s_n/n}$, the adaptive group lasso consistently selects the true active predictors with probability converging to 1, that is,

$$\mathbb{P}\left(\hat{\boldsymbol{\beta}}_{\text{AGL}} \overset{0}{=} \boldsymbol{\beta}^0\right) \to 1. \tag{22}$$

(ii) The rate of convergence of the adaptive group lasso estimator is given by

$$\sum_{j\in T}\|\hat{\boldsymbol{\beta}}_{\text{AGL}\,j} - \boldsymbol{\beta}_j^0\|_2^2 = O_p\left(s_n\gamma_2^{-2s_n}m_n^2\frac{\log(s_n m_n)}{n}\right) + O(s_n^2\gamma_2^{-2s_n}m_n^{1-2d}) + O(\lambda_{n2}^2 m_n^2 s_n\gamma_2^{-2s_n})$$

for the bounded response case and

$$\sum_{j\in T}\|\hat{\boldsymbol{\beta}}_{\text{AGL}\,j} - \boldsymbol{\beta}_j^0\|_2^2 = O_p\left(\gamma_n s_n\gamma_2^{-2s_n}m_n^2\frac{\log(s_n m_n)}{n}\right) + O(s_n^2\gamma_2^{-2s_n}m_n^{1-2d}) + O(\lambda_{n2}^2 m_n^2 s_n\gamma_2^{-2s_n})$$

for the unbounded response case, where γ_n is any diverging sequence.

The proof of this theorem is given in the supplementary materials. It is interesting to compare the adaptive group lasso results with Wang and Tian (2019), who studied the asymptotic properties of the adaptive group lasso for GLMs. It is worth noting that we considered a more general case by allowing the group size to diverge with n, and the eigenvalue to be bounded by sequences that depending on n on a broader domain. In the special case that corresponds to their assumptions, our results (Theorem 2) coincide with their results.

Similar to the group lasso estimator, we also derive the results for the nonparametric function estimations, stated in the following remark.

Remark 7. Let $\hat{f}_{\text{AGL } j}(x) = \Phi_j(x)\hat{\boldsymbol{\beta}}_{\text{AGL } j}$. We can also state the results of the second selection step in terms of functions, which is a direct consequence of theorem 2. First, we have the true nonzero subset is recovered with probability tending to 1. Moreover, by the same properties of spline as in Remark 5, we have

$$\sum_{j\in T}\|\hat{f}_{\text{AGL } j} - f_j\|_2^2 = O_p\left(s_n\gamma_2^{-2s_n}m_n\frac{\log(s_nm_n)}{n}\right) + O(s_n^2\gamma_2^{-2s_n}m_n^{-2d}) + O(\lambda_{n2}^2 m_n s_n \gamma_2^{-2s_n})$$

for the bounded response case and

$$\sum_{j\in T}\|\hat{f}_{\text{AGL } j} - f_j\|_2^2 = O_p\left(\gamma_n s_n\gamma_2^{-2s_n}m_n\frac{\log(s_nm_n)}{n}\right) + O(s_n^2\gamma_2^{-2s_n}m_n^{-2d}) + O(\lambda_{n2}^2 m_n s_n \gamma_2^{-2s_n})$$

for the unbounded case, for any diverging sequence γ_n.

The convergence rate for the group lasso estimator is

$$\sum_{j=1}^{p_n}\left\|f_j - \hat{f}_{nj}\right\|_2^2 = O_P\left(s_n\gamma_2^{-2s_n}\frac{m_n\log(p_nm_n)}{n}\right) + O(\lambda_{n1}^{b\ 2}m_n s_n\gamma_2^{-2s_n}) + O(s_n^2 m_n^{-2d}\gamma_2^{-2s_n}),$$

while for the adaptive group lasso estimator it is

$$\sum_{j\in T}\|\hat{f}_{\text{AGL } j} - f_j\|_2^2 = O_p\left(s_n\gamma_2^{-2s_n}m_n\frac{\log(s_nm_n)}{n}\right) + O(\lambda_{n2}^2 m_n s_n\gamma_2^{-2s_n}) + O(s_n^2\gamma_2^{-2s_n}m_n^{-2d}).$$

The regression term differs by the size of candidate set. The price we pay by not knowing the true set is $\log(pm_n)$ in the group lasso step, and becomes $\log(s_nm_n)$ in the adaptive group lasso step, since the initial estimator has recovered a super set of the true set with cardinality $O(s_n)$. The penalty term's difference appears on the tuning parameter, where λ_{n2} is of a smaller order than λ_{n1} with a multiplier of r_n^{-1}. According to our choice of λ_{n2}, it has a trivial upper bound which is of order $O(\lambda_{n1}^2)$. Therefore, the tuning parameter part in the penalty convergence rate term becomes quadratic. The approximation error term is not affected by the adaptive group lasso step.

The adaptive group lasso is important in two reasons: first, with probability tending to 1, this is able to select the true nonzero components accurately, which is not always the case in group lasso; second, the rate of convergence of the adaptive group lasso estimator is faster than the rate of convergence of the group lasso estimator. The difference in the leading terms are in the order of r_n^{-1}. This makes the adaptive group lasso estimator to achieve a better error with the same sample size, or the same error with a smaller sample size.

The theorem and remark in this section ensure that under mild assumptions, we are able to recover the true model with probability tending to 1 and achieve a rate of convergence better than the initial estimator. Particularly, if the restrictions of n, p_n, m_n, and s_n in the previous section satisfy, the group lasso estimator is actually a good initial estimator. Therefore, this two-step procedure actually is a complete procedure that gives us a way to do this model selection and estimation on any high-dimensional generalized additive model. However, the procedure is not practically complete without proper selection of the tuning parameter λ. Therefore, we propose a theoretically validated tuning parameter selection in the next section.

4 | TUNING PARAMETER SELECTION

One important issue in penalized methods is choosing a proper tuning parameter. It is known that the selection results are sensitive to the choice of tuning parameters. The theoretical results only provide the order of the tuning parameter, which is not very useful in practice. The reason is that the order of a sequence describes the limiting properties when n goes to infinity. In reality, our n is a fixed number, so we must have a practical instruction on selecting the tuning parameter.

Despite its importance, there is no much development for tuning parameter selection in the high-dimensional literature. The conventional tuning parameter selection criteria tend to select too many predictors, thus is hard to reach selection consistency. Another reason, especially in group lasso problems, is that the solution path of group lasso is piecewise nonlinear, which makes the testing procedures even harder. Here, we propose the GIC (Fan & Tang, 2013; Zhang et al., 2010) that supports consistent model selection.

Let $\hat{\boldsymbol{\beta}}^{\lambda}$ be the adaptive group lasso solution with tuning parameter λ. The GIC is defined as

$$\mathrm{GIC}(\lambda) = \frac{1}{n}\{D(\hat{\mu}_{\lambda}; \boldsymbol{Y}) + a_n|\hat{T}_{\lambda}|\}, \tag{23}$$

where $D(\hat{\mu}_{\lambda}; \boldsymbol{Y}) = 2\{l(\boldsymbol{Y}; \boldsymbol{Y}) - l(\hat{\mu}_{\lambda}; \boldsymbol{Y})\}$. Here the $l(\boldsymbol{\mu}; \boldsymbol{Y})$ is the log-likelihood function in Equation (3) expressed as a function of the expectation $\boldsymbol{\mu}$ and $\boldsymbol{Y}$. $l(\boldsymbol{Y}; \boldsymbol{Y})$ represents the saturated model with $\boldsymbol{\mu} = \boldsymbol{Y}$, and $\hat{\mu}_{\lambda} = b'(\sum_{i=1}^{p_n} \hat{f}_j^{\lambda}(x_{ij})) = b'(\phi\hat{\boldsymbol{\beta}}^{\lambda})$ is our estimated expectation when the tuning parameter is λ. The hyperparameter a_n is to penalize the size of the model. Using GIC, under proper choice of a_n, we are able to select all active predictors consistently.

The importance of the following consistency theorem is that the result in the previous section guarantees that with probability converging to 1, there exists a λ_{n0} that will be able to identify the true model. Therefore, a good choice of a_n will be able to identify the true model with probability converging to 1. For a support $A \subset \{1, \ldots, p\}$ such that $|A| \leq q_n$, where $q_n \geq s_n$ and $q_n = o(n)$, let

$$I(\boldsymbol{\beta}(A)) = E\left[\log(f^*/g_A)\right] = \sum_{i=1}^{n} \left[b'(\Phi_i\boldsymbol{\beta}^0)\Phi_i^T(\boldsymbol{\beta}^0 - \boldsymbol{\beta}(A)) - b(\Phi_i^T\boldsymbol{\beta}^0) + b(\Phi_i^T\boldsymbol{\beta}(A))\right] \tag{24}$$

be the Kullback-Leibler (KL) divergence between the true model and the selected model, where f^* is the density of the true model, and g_A is the density of the model with population parameter $\boldsymbol{\beta}(A)$. Let $\boldsymbol{\beta}^*(A)$ be the model with the smallest KL divergence over all models with support A, and let

$$\delta_n = \inf_{\substack{A \not\supset T \\ |A| \leq q_n}} \frac{1}{n} I(\boldsymbol{\beta}^*(A)).$$

Here we note that if $T \subset A$, the minimizer is automatically $\boldsymbol{\beta}^0$ and thus the KL divergence is zero. For an underfitted models $T \not\subset A$, δ_n describes how easily one can distinguish the models from the true model by measuring the minimum distance from the true model to the "best estimated models." Later in the theorems we will need to assume lower bounds on δ_n so that we will be able to reach our consistency results. The following theorem proves that GIC works under mild conditions.

Theorem 3. *Under Assumptions 1–6, suppose that* $\delta_n q^{-1} R_n^{-1} \to \infty$, $n\delta_n s_n^{-1} a_n^{-1} \to \infty$ *and* $a_n \psi^{-1} \to \infty$, *where* R_n *and* ψ_n *are defined in Lemmas B.3 and B.4, we have, as* $n \to \infty$,

$$\mathbb{P}\{\inf_{\lambda \in \Omega_- \cup \Omega_+} \mathrm{GIC}_{a_n}(\lambda) > \mathrm{GIC}_{a_n}(\lambda_{n0})\} \to 1, \tag{25}$$

where

$$\Omega_- = \{\lambda \in [\lambda_{\min}, \lambda_{\max}] : T_\lambda \not\supset T\},$$

$$\Omega_+ = \{\lambda \in [\lambda_{\min}, \lambda_{\max}] : T_\lambda \supset T \text{ and } T_\lambda \neq T\},$$

where T_λ *is the set of predictors selected by tuning parameter* λ. $\lambda_{\min}$ *can be chosen as the smallest* λ *such that the selected model has size* q *that satisfies the theorem assumption, and* $\lambda_{\max}$ *simply corresponds to a model with no variables.*

The proof of this theorem is given in supplementary materials. In practice, a choice of a_n is proposed to be $m_n \log(\log(n)) \log(p_n)$. We have

Corollary 1. *Under Assumptions 1–6, with choice of* $a_n = m_n \log(\log(n)) \log(p_n)$, *we have*

$$\mathbb{P}\{\inf_{\lambda \in \Omega_- \cup \Omega_+} \mathrm{GIC}_{a_n}(\lambda) > \mathrm{GIC}_{a_n}(\lambda_{n0})\} \to 1.$$

In our two-step procedure, there are two tuning parameters to be selected: λ_{n1} in the group lasso step and λ_{n2} in the adaptive group lasso step. The choice of λ_{n2} is of more importance, since λ_{n1} only serve as the parameter in screening. As long as we have a screening step that satisfies (14), we are ready for the adaptive group lasso step. To be simple, we propose to use GIC for selecting both λ_{n1} and λ_{n2}. As a result of the previous theorem, we are able to reach selection consistency. It is worth noting that the range of candidate tuning parameters is not explicitly specified by the assumptions. A practical guidance is to choose $\lambda_{\min}$ to be the minimum value such that the model is still estimable, and to choose $\lambda_{\max}$ such that we obtain a null model.

5 | NUMERICAL PROPERTIES

In this section we conduct various empirical exercises to illustrate our theoretically guided method in practice. To optimize the group lasso problems, we apply the algorithm named groupwise-majorization-descent (GMD) by Yang and Zou (2015), which approximates the convex log-likelihood part with second-order Taylor expansion and solves it with a quadratic function's closed-form solution, wrapped in a block coordinate descent algorithm. We made the algorithm in GAM available as a python class, which is accessible at https://github.com/KaixuYang/penalisedGAM.

As smoothness is a concern in practical GAM computations, we bring the P-spline (Eilers & Marx, 1996) penalty into the model while implementing the model numerically. The P-spline penalty controls the differences between coefficients of consecutive basis functions, and thus yields smoother spline functions.

Specifically, let $l(\boldsymbol{\beta}; \boldsymbol{X}, \boldsymbol{y})$ be the objective function in Section 3, either the group lasso objective function or the adaptive group lasso objective function. The objective function with smoothness penalty is defined as

$$l_s(\boldsymbol{\beta}; \boldsymbol{X}, \boldsymbol{y}) = l(\boldsymbol{\beta}; \boldsymbol{X}, \boldsymbol{y}) + \lambda_s \sum_{j=1}^{p} \boldsymbol{\beta}_j^T \boldsymbol{D} \boldsymbol{\beta}_j, \tag{26}$$

where

$$\boldsymbol{D} = \begin{bmatrix} 1 & -1 & 0 & . \\ -1 & 2 & -1 & . \\ 0 & -1 & 2 & . \\ . & . & . & . \end{bmatrix}$$

It is worth noting that the penalty term in (26) is different from the smooth-sparsity penalty terms in some existing literature (Amato et al., 2016; Meier et al., 2009). The difference mainly lies on whether square root is taken on the quadratic smoothness penalty. In our case, as we only need a smoothness penalty to work along with the sparsity penalty, the effect of the square root can be obtained with a proper choice of the independent smoothness penalty λ_s. The examples in this section shows the effectiveness of this penalty combination in terms of accurate variable selection. On the other hand, smoothness penalty without the square root brings more efficiency in computation. A slightly modified soft-thresholding function is used to handle the combination of the group lasso penalty and the smoothness penalty.

5.1 | Simulated examples

Here we undertake extensive simulation study to see the performance of our proposed two-step selection and estimation approach. We investigate the performance of both uncorrelated and correlated covariates and we consider different sample sizes and varying number of predictors in each case.

In this section, we consider three different types of generalized models: the logistic regression (Bernoulli distribution), the Poisson regression (Poisson distribution), and the Gamma regression (Gamma distribution). Through the whole subsection, we choose $l = 4$ which implies a cubic B-spline. We choose $m_n = 9$ for most cases unless stated otherwise. The choice of l and m_n implies that there are $m_n - l = 5$ inner knots, which are evenly placed over the empirical percentiles of the training data. In this subsection, we compare the performance of the two-step approach with the Lasso (Tibshirani, 1996), the GAMBoost (Tutz & Binder, 2006) and the GAMSEL (Chouldechova & Hastie, 2015). We implement our two-step approach with our own package mentioned above. The Lasso is implemented with the scikit-learn package in python. The GAMBoost and GAMSEL methods are implemented using their packages in R. In the group lasso step, we choose the tuning parameter corresponding to n_g variables, where n_g is the largest number such that $n_g \times m_n <= n$.

This choice prevents estimation issues when we have too many parameters. The GIC procedure is applied in the adaptive group lasso step to select tuning parameters. In the GIC procedure, the tuning parameter selection criterion is defined as

$$\text{GIC}(\lambda) = \frac{1}{n}\{D(\hat{\mu}_\lambda; \boldsymbol{Y}) + a_n|\hat{T}|\}. \tag{27}$$

From our results in the previous section, we choose $a_n = (\log\log n)(\log p)m_n$.

5.1.1 | Logistic regression

First, we consider the logistic regression

$$y_i \sim \text{Bernoulli}(\theta_i), \quad i = 1, \ldots, n, \tag{28}$$

where $\theta_i = \text{logit}^{-1}[\alpha + \sum_{j=1}^{p} f_j(x_{ij})]$ and x_{ij} is the (i,j)th element of the design matrix X.

Example 1. We first consider the logistic additive model on an independent design matrix case, where each predictor in X is independent of other predictors. Each element of the design matrix is generated from a Unif$(-1, 1)$ distribution. We consider three different cases with all n, p, and s increasing, which coincides with our theory in Section 3. Specifically, the three cases are: $n = 100$, $p = 200$, and $s = 3$; $n = 200$, $p = 500$, and $s = 4$; $n = 300$, $p = 3000$, and $s = 5$. A testing sample of size 1000 is generated independently to measure the performance. For all three cases, we have nonzero functions $f_1(x) = 5\sin(3x)$, $f_2(x) = -4x^4 + 9.33x^3 + 5x^2 - 8.33x$, and $f_3(x) = x(1 - x^2)\exp(3x) - 4$. These three general terms include a periodic term, a polynomial term, and an exponential term. The last two cases have one more function of $f_4(x) = 4x$, a linear term. Finally, the last case has an addition $f_5(x) = 4\sin(-5\log(\sqrt{x+3})$, a complicated composite function. Without loss of generality, the first s functions are set to be nonzero. The constants in the functions are to ensure similar signal strength and smoothness. The other functions $f_{s+1}(x) = \cdots = f_p(x) = 0$.

Our results focus on NV, the average number of variables being selected; TPR, the true positive rate (what percent of the truly nonzero variables are selected); FPR, the false positive rate (what percent of the zero variables are selected); and PE, the prediction error. In the logistic regression problem, our metric to measure the prediction error will be the misclassification rate, which is also the measurement in Chouldechova and Hastie (2015). The simulation results are averaged over 100 repetitions.

The simulation results are summarized in Table 1. Compared with the classical method Lasso and the existing GAM methods GAMSEL and GAMBoost, the two-step approach performs the best in terms of both variable selection and estimation in the high-dimensional setup. The two-step approach performs significantly better in prediction errors. In variable selection, the two-step approach selects the closest number of variables to the ground truth, while keeping the TPR high and FPR low. The existing GAM algorithms have similar TPR but includes too many false positives. The existing GAM algorithms were not intended for very high-dimensional data, and thus fails to handle the variable selection and prediction at the same time. As mentioned in Fan and Li (2001), the tuning parameter in the Lasso for consistent variable selection is not the same as the tuning parameter for best prediction. We can see this may also be true for the group lasso case, since the estimated nonzero coefficients in the group lasso step are overpenalized.

TABLE 1 Simulation results for the two-step approach compared with the Lasso, GAMSEL, and GAMBoost in the three cases of Example 1

	$n = 100, p = 200, s = 3$				$n = 200, p = 500, s = 4$				$n = 300, p = 3000, s = 5$			
	NV	**TPR**	**FPR**	**PE**	**NV**	**TPR**	**FPR**	**PE**	**NV**	**TPR**	**FPR**	**PE**
Two-step	3.56	0.920	0.004	0.148	4.82	0.989	0.002	0.128	4.92	0.968	0.000	0.122
	(1.19)	(0.146)	(0.005)	(0.027)	(1.02)	(0.057)	(0.002)	(0.018)	(0.535)	(0.086)	(0.000)	(0.018)
Lasso	30.0	0.920	0.138	0.249	64.7	0.978	0.122	0.229	85.2	0.816	0.027	0.211
	(17.9)	(0.144)	(0.090)	(0.041)	(19.2)	(0.452)	(0.039)	(0.024)	(68.3)	(0.243)	(0.022)	(0.024)
GAMSEL	10.1	0.820	0.039	0.241	14.0	0.943	0.021	0.214	33.9	0.986	0.010	0.208
	(11.1)	(0.209)	(0.055)	(0.035)	(12.6)	(0.112)	(0.025)	(0.023)	(27.9)	(0.065)	(0.009)	(0.016)
GAMBoost	44.7	0.738	0.213	0.231	85.4	1.00	0.164	0.196	138	0.996	0.044	0.186
	(4.84)	(0.055)	(0.025)	(0.027)	(6.88)	(0.000)	(0.014)	(0.018)	(9.64)	(0.028)	(0.003)	(0.015)

Note: Results are averaged over 100 repetitions. Enclosed in parentheses are the corresponding standard errors.
Abbreviations: FPR, the false positive rate; NV, average number of the variables being selected; PE, prediction error (here is the misclassification rate); TPR, the true positive rate.

This also proves that an adaptive group lasso step is important, in terms of both variable selection and prediction.

In practice, the predictors are sometimes correlated to each other. It is interesting to see how well the procedure performs in correlated predictor cases. Therefore, we also perform the same comparison on correlated predictors.

Example 2. In this example, we study the case where the design matrix contains correlated predictors. We generate the data in the following way. First we generate each element of $X_{n\times p}$ independently from Unif$(-1, 1)$. Then we generate u from Unif$(-1, 1)$, independently from $X_{n\times p}$. Then all columns of X are transformed using $X_j = (X_j + tu)/\sqrt{1 + t^2}$. This procedure controls the correlation among predictors through t such that $\text{corr}(x_{ik}, x_{ij}) = t^2/(1 + t^2)$. Here the simulation is run on $n = 100$, $p = 200$, and $s = 3$. All other setups are kept same as Example 1. In our example, we choose $t = \sqrt{3/7}$, where the correlation is 0.3 and $t = \sqrt{7/3}$, where the correlation is 0.7.

The results are summarized in Table 2. In the correlated cases, all four methods are influenced, more or less. In terms of variable selection, the two-step approach still has the closest number of selected variables. The methods behave differently in terms of TPR and FPR. GAMBoost tends to have greater numbers in both TPR and FPR, while GAMSEL tends to have both lower numbers. The two-step approach balances between those two methods, while maintaining the smallest FPR among all methods. In terms of the prediction error, the two-step approach significantly beats the other methods. The results show good performance of the two-step approach, and again emphasize that the adaptive group lasso step is necessary for better selection and estimation.

This underselection for correlated predictors has been an issue for the lasso and adaptive lasso methods. For nonparametric additive models, Huang et al. (2010) found the same issue when dealing with correlated predictors. Also the NIS proposed by Fan et al. (2011) did not perform well in correlated predictors compared with uncorrelated case. Our two-step approach is not affected too much with the correlation, in terms of both variable selection and prediction.

It also happens in the real world that the signal strength is low. Therefore, it is interesting to consider a case where we have lower signal strength than in Example 1.

TABLE 2 Simulation results for the two-step approach compared with the Lasso, GAMSEL, and GAMBoost in Example 2 with correlation 0.3 and 0.7 for $n = 100$, $p = 200$, and $s = 3$

	Cor = 0.3				Cor = 0.7			
	NV	TPR	FPR	PE	NV	TPR	FPR	PE
Two-step	2.82 (0.994)	0.753 (0.229)	0.003 (0.004)	0.171 (0.033)	2.05 (0.829)	0.557 (0.170)	0.002 (0.003)	0.174 (0.022)
Lasso	37.0 (38.2)	0.690 (0.259)	0.176 (0.194)	0.312 (0.069)	21.9 (37.9)	0.327 (0.291)	0.103 (0.193)	0.288 (0.047)
GAMSEL	15.4 (16.0)	0.573 (0.285)	0.069 (0.079)	0.342 (0.065)	12.5 (9.15)	0.397 (0.271)	0.057 (0.044)	0.264 (0.033)
GAMBoost	44.2 (5.21)	0.977 (0.085)	0.209 (0.026)	0.268 (0.033)	33.7 (4.52)	0.860 (0.178)	0.158 (0.014)	0.203 (0.026)

Note: Results are averaged over 100 repetitions. Enclosed in parentheses are the corresponding standard errors.
Abbreviations: NV, average number of the variables being selected; FPR, the false positive rate; PE, prediction error (here is the misclassification rate); TPR, the true positive rate.

TABLE 3 Simulation results for the two-step approach compared with the Lasso, GAMSEL, and GAMBoost in Example 3, with $n = 100$, $p = 200$, $s = 3$, and reduced signal strength

	NV	TPR	FPR	PE
Two-step	3.91 (2.05)	0.703 (0.240)	0.009 (0.009)	0.218 (0.033)
Lasso	30.0 (30.5)	0.770 (0.304)	0.142 (0.154)	0.258 (0.036)
GAMSEL	15.3 (18.0)	0.510 (0.266)	0.070 (0.090)	0.377 (0.054)
GAMBoost	50.3 (5.11)	0.980 (0.079)	0.240 (0.026)	0.308 (0.028)

Note: Results are averaged over 100 repetitions. Enclosed in parentheses are the corresponding standard errors.
Abbreviations: FPR, the false positive rate; NV, average number of the variables being selected; PE, prediction error (here is the misclassification rate); TPR, the true positive rate.

Example 3. In the next example, we reduce the signal strength of Example 1 by a factor of 2, while all other assumptions are kept the same. The results are shown in Table 3. From the table we see that minimal signal strength is an important factor to the performance of variable selection in the generalized models. The performance is impacted by the signal strength for all models. The two-step approach still have the closest number of nonzero variables to the ground truth. Though the true positive rate is lower than that of the lasso or the GAMBoost, the latter two methods have too many false positives. The Lasso or GAMBoost selects too many variables and should not be considered as good variable selection methods. Moreover, the prediction error of the two-step approach remains the best among all four methods.

5.1.2 | Other link functions

In this subsection, we study the performance of the two-step approach numerically on the Poisson regression and Gamma regression. In the Poisson regression, we have

$$y_i \sim \text{Poisson}(\theta_i), \quad i = 1, \dots, n, \tag{29}$$

TABLE 4 Simulation results for the two-step approach compared with the Lasso, GAMSEL, and GAMBoost in Example 4 for Poisson regression and Gamma regression with $n = 100$, $p = 200$, and $s = 3$

	Poisson regression				Gamma regression			
	NV	TPR	FPR	PE	NV	TPR	FPR	PE
Two-step	4.30 (1.51)	0.930 (0.172)	0.008 (0.009)	2.34 (0.703)	3.57 (0.98)	0.997 (0.033)	0.003 (0.005)	14.4 (19.5)
Lasso	13.4 (9.79)	0.867 (0.189)	0.054 (0.050)	3.51 (0.403)	12.5 (7.72)	0.887 (0.196)	0.048 (0.039)	42.3 (11.5)
GAMBoost	82.1 (4.27)	1.00 (0.000)	0.401 (0.022)	15.4 (2.12)	NA	NA	NA	NA

Note: Results are averaged over 100 repetitions. Enclosed in parentheses are the corresponding standard errors. The GAMBoost method does not support Gamma regression with noncanonical link function, while the canonical link falls outside of range, therefore it does not support Gamma regression.
Abbreviations: FPR, the false positive rate; NV, average number of the variables being selected; PE, prediction error (here is the misclassification rate); TPR, the true positive rate.

where $\theta_i = \exp[\alpha + \sum_{j=1}^{p} f_j(x_{ij})]$ and x_{ij} is the (i,j)th element of the design matrix X. In the Gamma regression, we have

$$y_i \sim \text{Gamma}(\theta_i, \phi), \quad i = 1, \ldots, n, \tag{30}$$

where $\theta_i = \exp[\alpha + \sum_{j=1}^{p} f_j(x_{ij})]$ and x_{ij} is the (i,j)th element of the design matrix X. The dispersion parameter ϕ is assumed to be known. Without loss of generality, we take $\phi = 1$.

Example 4. In this example, we keep the same setup as in Example 1 to generate the design matrix, and use the Poisson distribution/Gamma distribution above to generate response variables. All other parameters are kept the same as in Example 1, but the signal strength is set to 1/4 of the original signal strength, and we set $n = 100$, $p = 200$, and $s = 3$. We compare the two-step approach with GLMs and the GAMBoost. Note that the GAMSEL only supports Gaussian and binomial link, thus is not used as a comparison here. The GAMBoost only supports generalized models with canonical link. The canonical link for Gamma regression suffers from the risk that the mean might fall outside of its range, thus the canonical link is not useful in practice. Therefore, we only use GAMBoost in Poisson regression as a comparison. Our algorithm works for both Gamma regression and Poisson regression, and to the best of our knowledge, is the only publicly available algorithm that supports both in the high-dimensional settings. The GLMs are run with the scikit-learn package in python.

The results are provided in Table 4. We see the two-step approach works significantly better than the linear model, and than the GAMBoost in the Poisson regression case, except for the true positive rate. The GAMBoost has a perfect true positive rate, which is slightly better than that of our two-step approach. However, the same issue as before is that it selected too many variables and make the false positive rate much higher than tolerable. Moreover, the prediction performance on the two-step approach is also in the first place in both the cases.

5.2 | Real data examples

In this section, we provide three real data examples to illustrate our procedure. In the first example, we consider the case $n > p$ in the classification setup, in the second example, we consider

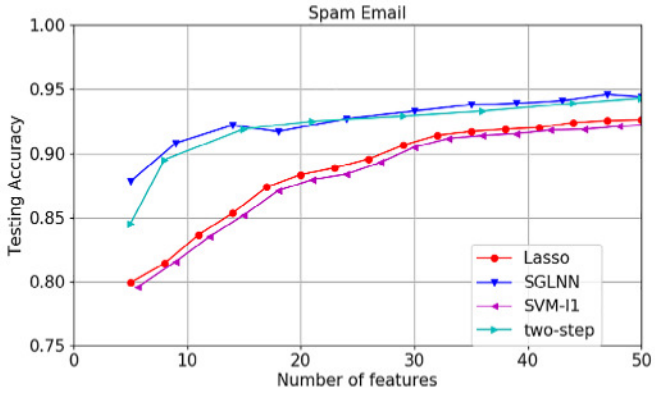

FIGURE 1 The classification accuracy against the number of nonzero variables measured on a testing set for Example 5 over 50 repetitions. The two-step approach, the logistic regression with Lasso, the l_1 norm penalized SVM, and the sparse group lasso neural network are included for comparison [Colour figure can be viewed at wileyonlinelibrary.com]

the high-dimensional setup $n < p$ in the classification setup, and in the third example, we consider a Gamma regression model.

Example 5. In this example, we use the data set in Example 5 of Friedman et al. (2001), the spam data as an example of the case $n > p$. The data set is available at https://web.stanford.edu/~hastie/ElemStatLearn/data.html. This data set has been studied in many different contexts with the objective being to predict whether an email is a spam or not based on a few features of the emails. There are $n = 4601$ observations, among which 1813 (39.4%) are spams. There are $p = 57$ predictors, including 48 continuous real [0, 100] attributes of the relative frequency of 48 "spam" words out of the total number of words in the email, six continuous real [0, 100] attributes of the relative frequency of six "spam" characters out of the total number of characters in the email, one continuous real attribute of average length of uninterrupted sequences of capital letters, one continuous integer attribute of length of longest uninterrupted sequence of capital letters, and one continuous integer attribute of total number of capital letters in the email. The data were first log transformed, since most of the predictors have long-tailed distribution, as mentioned in Friedman et al. (2001). They were then centered and standardized.

The data were split into a training data set with 3067 observations and a testing data set with 1534 observations. We choose order $l = 4$ which implies a cubic B-spline. We choose $m_n = 15$, which implies there are 11 inner knots, evenly placed over the empirical percentiles of the data. We compare the result with the logistic regression with Lasso penalty, the support vector machine (SVM) with Lasso penalty, and the sparse group lasso neural network (SGLNN, Feng & Simon, 2017, see also Yang & Maiti, 2020). The Lasso and SVM are implemented with the scikit-learn module in python, and the SGLNN is implemented with the algorithm in the paper in python. By changing the tuning parameter or stopping criterion, we get estimations with different sparsity levels. All results are averaged over 50 repetitions. The classification error with different level of sparsity is shown in Figure 1. The two-step approach and the neural network perform better than the linear models, which indicates a nonlinear relationship. The two-step approach has maximum accuracy 0.944, while that for the neural network is 0.946. The neural network performs a little better than the two-step approach due to its ability to model the interactions among predictors, but this difference is not significant. However, neural network has no interpretation

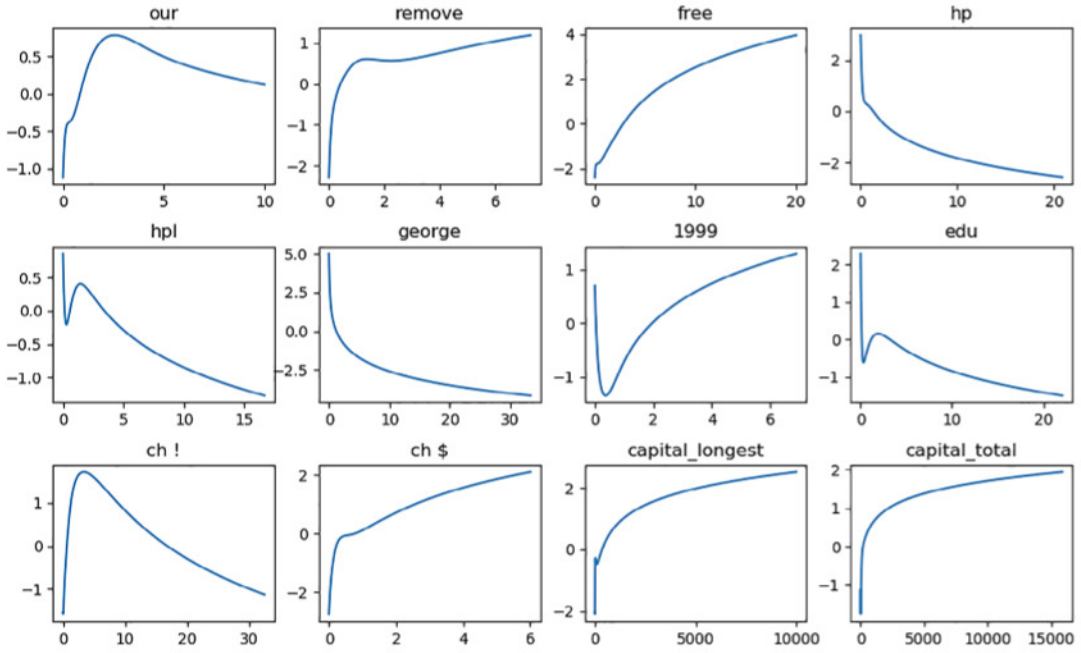

FIGURE 2 The estimated functions for the most frequently selected functions for Example 5 [Colour figure can be viewed at wileyonlinelibrary.com]

and takes longer to train. All four methods have performance increase as more predictors are included, which indicates that all predictors contributes to some effect to the prediction. However, we are able to reach more than 0.9 accuracy with only 15 predictors included. With the GIC criterion, the two-step approach selects 14.6 ± 1.52 predictors, with an average accuracy of 0.914 ± 0.015. The most frequently selected functions are shown in Figure 2, which also shows that these functions are truly nonlinear. The plots are of the original functions, that is, before the logarithm transformation. The estimated functions are close to the results in Friedman et al. (2001), chapter 9, with slight scale difference due to different penalization. The results show that the additive model by the adaptive group lasso is more suitable for this data than linear models.

Example 6. For high-dimensional classification example, we use the prostate cancer gene expression data described in http://featureselection.asu.edu/datasets.php. The data set has a binary response. 102 observations were studied on 5966 predictor variables, which indicates that the data set is really a high-dimensional data set. The responses have values 1 (50 sample points) and 2 (52 sample points), where 1 indicates normal and 2 indicates tumor. All predictors are continuous predictors, with positive values.

To see the performance of our procedure, we ran 100 replications. In each replication, we randomly choose 76 of the observations as training data set and the rest 26 observations as testing data set. We choose order $l = 4$ which implies a cubic B-spline. We choose $m_n = 9$, which implies there are five inner knots, evenly placed over the empirical percentiles of the data. Similar to the last example, we compare the result with the logistic regression with Lasso penalty, the SVM with Lasso penalty, and SGLNN. The classification error with different level of sparsity is shown in Figure 3. From the figure we see that compared with linear methods such as the logistic regression or SVM, the nonparametric approaches converges faster. The two-step approach reaches a testing

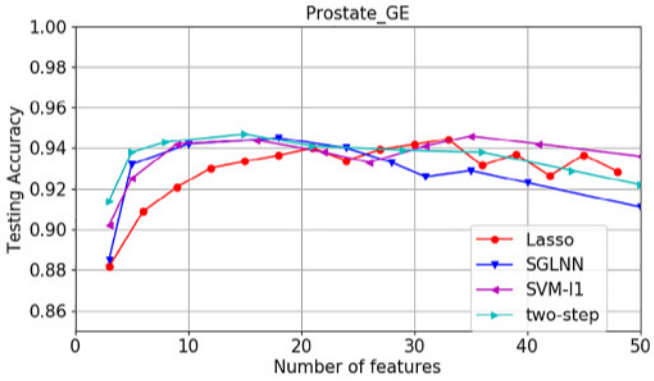

FIGURE 3 The classification accuracy against the number of nonzero variables measured on a testing set for Example 6 over 500 repetitions. The two-step approach, the logistic regression with Lasso, the l_1 norm penalized SVM, and the sparse group lasso neural network are included for comparison [Colour figure can be viewed at wileyonlinelibrary.com]

accuracy of 0.945 when around 15 variables are included in the model, while the linear methods need over 30 variables to reach competitive results. Compared with neural network, the two-step approach is easier to implement with stabilized performances. A drawback of the nonparametric methods is to easily overfit for small sample, and that is the reason the performance drops as too many variables entered the into the model. With the GIC criterion, the two-step approach selects 3.25 ± 1.67 predictors, with an average accuracy of 0.914 ± 0.016. To show the nonlinear relationship, Figure 4 shows the estimated functions for the six most frequently selected variables.

Example 7. In this example, we investigate the performance of the two-step approach on Gamma regression. The data set is from National Oceanic and Atmospheric Administration (NOAA). We use the storm data, which includes the occurrence of storms in the United States with the time, location, property damage, a narrative description, and so on. Here we only take the data in Michigan from 2010 to 2018 and keep the narrative description as our predictor variable and the property damage as our response variable. The description is in text, therefore we applied wording embedding algorithm Word2vec (Mikolov et al., 2013) to transform each description into a numeric representation vector of length $p = 701$, similar word embedding preprocessing can be found in Lee et al. (2020). The response variable property damage has a long tail distribution, thus we use a Gamma regression here. After removing outliers, the data set contains 3085 observations. In order to study the high-dimensional case, we randomly sample 10% of the observations as our training data ($n = 309$) and the rest are used for validation. Moreover, the response is normalized with the location and scale parameters of gamma distribution.

To see the performance of our procedure, we ran 50 replications. We choose order $l = 4$ which implies a cubic B-spline. We choose $m_n = 9$, which implies there are five inner knots, evenly placed over the empirical percentiles of the data. Since there is limited libraries available for variable selection under high-dimensional gamma model, we compare the two-step approach with the linear regression with Lasso on a logarithm transformation on the response variable. The prediction error with different level of sparsity is shown in Figure 5. With the GIC criterion, the two-step approach selects 34.45 ± 3.52 predictors, with an average mean squared error (MSE) of 0.004334 ± 0.000115. However, from the plot we see that the linear model was not able to reach this accuracy through the whole solution path, with the best accuracy of 0.004337 at around 80

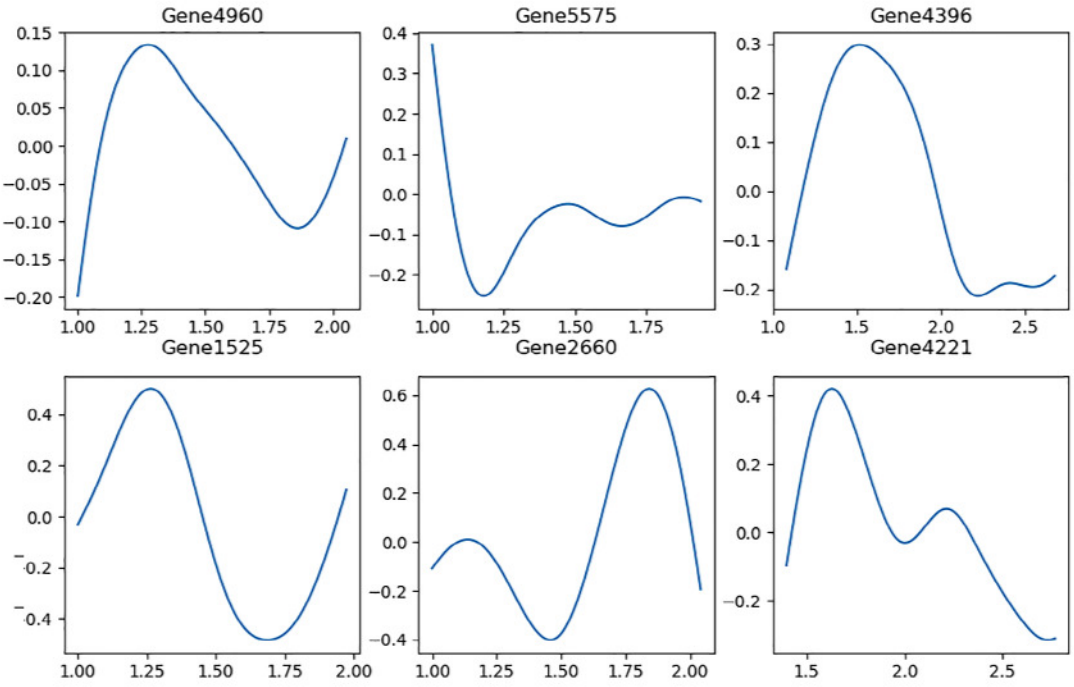

FIGURE 4 The estimated functions for the most frequently selected functions ordered by descending in frequency for Example 6 [Colour figure can be viewed at wileyonlinelibrary.com]

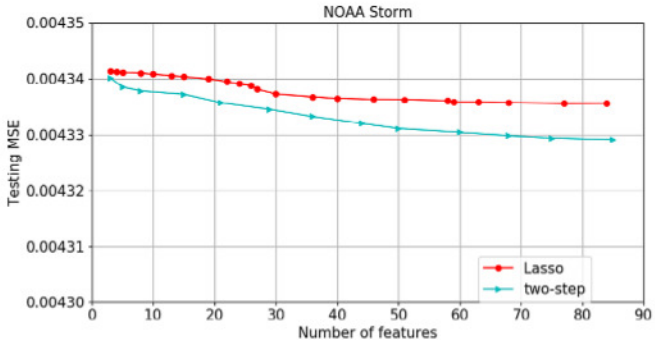

FIGURE 5 The testing MSE against the number of nonzero variables measured on a testing set for Example 7 over 50 repetitions. The two-step approach and logarithm transformation with the Lasso are included for comparison [Colour figure can be viewed at wileyonlinelibrary.com]

nonzero variables. This example also shows the superior of the nonparametric model over linear models.

6 | DISCUSSION

In this article, we considered ultrahigh-dimensional ($\log p_n = O(n^\rho)$) generalized additive model with a diverging number of nonzero functions ($s_n \to 0$ *as* $n \to \infty$). After using basis expansion

on the nonparametric functions, we used two-step procedures—group lasso and adaptive group lasso to select the true model. We have proved the screening consistency of the group lasso estimator and the selection consistency of the adaptive group lasso estimator. The rates of convergence of both estimators were also derived, which proved that the adaptive group lasso does have an improvement on the estimator. The whole article provides a solid foundation for the existing methods. Finally we proved that under this nonparametric setup, the GIC is a good way to select the tuning parameter that consistently selects the true model.

In this article, we used a fixed design on the data matrix $\boldsymbol{X}$. A random design on $\boldsymbol{X}$ could be considered, that is, X has a continuous distribution function $f_X(X)$ on its interval $[a, b]$; however, extra assumptions such as the boundedness of the density function are needed to reach the same result. Also we proved the selection consistency of the GIC procedure on the adaptive group lasso estimator, conditioning that the initial estimator satisfies (14), which is possessed by the group lasso procedure with probability tending to 1. However, the properties of the group lasso initial estimator could be further studied beyond the screening consistency, for example, could it have selection consistency under appropriate conditions? This is a challenging problem, since there does not have to exist a tuning parameter that gives selection consistency in the group lasso procedure, but this is an interesting problem that deserves further investigation.

Moreover, the heteroscedastic error case is also attractive in high-dimensional GAM. The square root Lasso (Belloni et al., 2011) has been proved to overcome this issue; however, it has not been extended to the nonparametric setup. It could be interesting to apply square root Lasso on the GAM to incorporate this case. This is a demanding topic that deserves further investigation as well.

ORCID

Kaixu Yang https://orcid.org/0000-0002-8971-0257
Tapabrata Maiti https://orcid.org/0000-0002-9362-4984

REFERENCES

Amato, U., Antoniadis, A., & De Feis, I. (2016). Additive model selection. *Statistical Methods & Applications*, *25*(4), 519–564.

Bakin, S. (1999). *Adaptive regression and model selection in data mining problems* (Ph.D. thesis). School of Mathematical Sciences, Australian National University.

Belloni, A., & Chernozhukov, V. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli*, *19*(2), 521–547.

Belloni, A., Chernozhukov, V., & Wang, L. (2011). Square-root lasso: Pivotal recovery of sparse signals via conic programming. *Biometrika*, *98*(4), 791–806.

Bickel, P. J., Ritov, Y., & Tsybakov, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, *37*, 1705–1732.

Bühlmann, P., & van de Geer, S. (2011). *Statistics for high-dimensional data: Methods, theory and applications* (1st ed.). Springer Publishing Company, Incorporated.

Chatterjee, A., & Lahiri, S. (2013). Rates of convergence of the adaptive lasso estimators to the oracle distribution and higher order refinements by the bootstrap. *The Annals of Statistics*, *41*(3), 1232–1259.

Chen, J., & Chen, Z. (2008). Extended bayesian information criteria for model selection with large model spaces. *Biometrika*, *95*(3), 759–771.

Chouldechova, A., & Hastie, T. (2015). Generalized additive model selection. *arXiv preprint arXiv:1506.03850*.

Das, D., Gregory, K., and Lahiri, S. (2017). Perturbation bootstrap in adaptive lasso. *arXiv preprint arXiv:1703.03165*.

De Boor, C. (2001). *A practical guide to splines* (revised ed.). Springer.

Eilers, P. H., & Marx, B. D. (1996). Flexible smoothing with b-splines and penalties. *Statistical Science*, *11*, 89–102.

Fan, J., Feng, Y., & Song, R. (2011). Nonparametric independence screening in sparse ultra-high-dimensional additive models. *Journal of the American Statistical Association*, *106*(494), 544–557.

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, *96*(456), 1348–1360.

Fan, J., & Lv, J. (2011). Nonconcave penalized likelihood with NP-dimensionality. *IEEE Transactions on Information Theory*, *57*(8), 5467–5484.

Fan, J., & Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics*, *32*(3), 928–961.

Fan, J., & Song, R. (2010). Sure independence screening in generalized linear models with np-dimensionality. *The Annals of Statistics*, *38*(6), 3567–3604.

Fan, Q., & Zhong, W. (2018). Nonparametric additive instrumental variable estimator: A group shrinkage estimation perspective. *Journal of Business & Economic Statistics*, *36*(3), 388–399.

Fan, Y., & Tang, C. Y. (2013). Tuning parameter selection in high dimensional penalized likelihood. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *75*(3), 531–552.

Feng, J., & Simon, N. (2017). Sparse-input neural networks for high-dimensional nonparametric regression and classification. *arXiv preprint arXiv:1711.07592*.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning Springer series in statistics* (Vol. *1*). Springer.

Hastie, T., & Tibshirani, R. (1986). [Generalized additive models]: Rejoinder. *Statistical Science*, *1*(3), 314–318.

Huang, J., Horowitz, J. L., & Wei, F. (2010). Variable selection in nonparametric additive models. *Annals of Statistics*, *38*(4), 2282.

Lee, G. Y., Manski, S., & Maiti, T. (2020). Actuarial applications of word embedding models. *ASTIN Bulletin: The Journal of the IAA*, *50*(1), 1–24.

Liu, R., Yang, L., & Härdle, W. K. (2013). Oracally efficient two-step estimation of generalized additive model. *Journal of the American Statistical Association*, *108*(502), 619–631.

Marra, G., & Wood, S. N. (2011). Practical variable selection for generalized additive models. *Computational Statistics & Data Analysis*, *55*(7), 2372–2387.

Meier, L., Van de Geer, S., & Bühlmann, P. (2009). High-dimensional additive modeling. *The Annals of Statistics*, *37*(6B), 3779–3821.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Nandy, S., Lim, C. Y., & Maiti, T. (2017). Additive model building for spatial regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *79*(3), 779–800.

Schumaker, L. (1981). *Spline functions: Basic theory*. John Wiley & Sons.

Stone, C. J. (1986). The dimensionality reduction principle for generalized additive models. *The Annals of Statistics*, *14*(2), 590–606.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288.

Tutz, G., & Binder, H. (2006). Generalized additive modeling with implicit variable selection by likelihood-based boosting. *Biometrics*, *62*(4), 961–971.

Van de Geer, S. A. (2008). High-dimensional generalized linear models and the lasso. *The Annals of Statistics*, *36*(2), 614–645.

Wang, H., Li, B., & Leng, C. (2009). Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *71*(3), 671–683.

Wang, M., & Tian, G.-L. (2019). Adaptive group lasso for high-dimensional generalized linear models. *Statistical Papers*, *60*(5), 1469–1486.

Yang, K., & Maiti, T. (2020). Statistical aspects of high-dimensional sparse artificial neural network models. *Machine Learning and Knowledge Extraction*, *2*(1), 1–19.

Yang, Y., & Zou, H. (2015). A fast unified algorithm for solving group-lasso penalize learning problems. *Statistics and Computing*, *25*(6), 1129–1141.

Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *68*(1), 49–67.

Zhang, N. R., & Siegmund, D. O. (2007). A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics*, *63*(1), 22–32.

Zhang, Y., Li, R., & Tsai, C.-L. (2010). Regularization parameter selections via generalized information criterion. *Journal of the American Statistical Association*, *105*(489), 312–323.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, *101*(476), 1418–1429.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Yang, K., & Maiti, T. (2021). Ultrahigh-dimensional generalized additive model: Unified theory and methods. *Scandinavian Journal of Statistics*, 1–26. https://doi.org/10.1111/sjos.12548