MODAL ADDITIVE MODELS WITH DATA-DRIVEN STRUCTURE IDENTIFICATION

TIELIANG GONG AND CHEN Xu
Department of Mathematics and Statistics, University of Ottawa
Ottawa, ON, K1N 6N5, Canada

HONG CHEN∗
College of Science, Huazhong Agricultural University
Wuhan 430070, China

(Communicated by Ding-Xuan Zhou)

Abstract. Additive models, due to their high flexibility, have received a great deal of attention in high dimensional regression analysis. Many efforts have been made on capturing interactions between predictive variables within additive models. However, typical approaches are designed based on conditional mean assumptions, which may fail to reveal the structure when data is contaminated by heavy-tailed noise. In this paper, we propose a penalized modal regression method, Modal Additive Models (MAM), based on a conditional mode assumption for simultaneous function estimation and structure identification. MAM approximates the non-parametric function through forward neural networks, and maximizes modal risk with constraints on the function space and group structure. The proposed approach can be implemented by the half-quadratic (HQ) optimization technique, and its asymptotic estimation and selection consistency are established. It turns out that MAM can achieve satisfactory learning rate and identify the target group structure with high probability. The effectiveness of MAM is also supported by some simulated examples.

1. Introduction. The demand for high dimensional data analysis has increased significantly over past decades by the emergence of applications such as social networks, bioinformatics and computer vision. Variable selection, as an important topic in high dimensional data analysis, has received tremendous attention in statistics and machine learning communities. The main goal of variable selection is identifying the truly informative variables for prediction accuracy and interpretability. Generally, an ideal variable selection algorithm should be flexible, robust and with theoretical guarantees [16]. In particular, the algorithm should not impose model assumptions that are too restrictive so that it can be used in a wide range of applications. It should also be computationally efficient and compatible with high performance computational platforms. What’s more, theoretical guarantees should be provided to ensure the asymptotic consistency in identifying the target variables.

2020 Mathematics Subject Classification. Primary: 68T05; Secondary: 62J02.
Key words and phrases. Structure identification, modal regression, additive model, generalization error, kernel density estimation.

∗ Corresponding author: Hong Chen.
In recent years, a considerable of variable selection algorithms have been proposed rely on different model assumptions, the most popular one being the linear model assumption. Typical examples include non-negative garrote [1], Lasso [28], SCAD [10], adaptive Lasso [39] and truncated $L_1$ penalty [25]. The main idea is utilizing the least square loss function associated with a sparsity-inducing penalty to obtain a sparse representation of the predict function. Despite their widespread usage, the rigid parametric forms of linear models limit their ability to capture nonlinear covariate effects. Additive models, where each component is a univariate smooth function with respect to a single variable, offer a more flexible way to implement variable section compared with linear models, see [27, 18, 17, 33] for example. While much progress has been made on this problem, these methods are not able to exploit the interactions among the covariates that may exist as prior knowledge in many applications. To address this issue, some high-order additive models have been developed, allowing each function component contains more than one covariate, such as the ANOVA models [15], COSSO model [23], GroupSpAM [33, 3] and GASI model [2]. Albeit with more flexibility and interpretability compared to classic additive models, the number of candidate functional components to be estimated increases exponentially with the dimension, which makes these methods computationally infeasible in many applications. Variable screening is another popular line in high dimensional data analysis. It screens out uninformative variables by checking the marginal relationship between the response of each variable. The marginal relationship can be characterized by various criteria, including Pearson’s correlation [11] and the empirical function norm [9]. These methods are all computationally efficient and with the sure screening property, i.e. all the truly informative variables are retained after screening with probability one.

All the above mentioned variable selection methods are formulated in the framework of mean regression. Some recent studies [5, 29] show that mean regression models like ridge regression, Lasso [28], group Lasso [34] etc., fail to capture the data structure when training data is contaminated by outliers. Fortunately, this problem can be alleviated by modal regression models [6, 24, 32, 5, 12], which aims to estimate the conditional mode of output $Y$ for the given input $X = x$. Compared to mean regression, modal regression will be more efficient in the presence of outliers or heavy-tailed error distribution. It can reveal the structure of outputs and the trend of observations [5]. Inspired by this idea, we propose a penalized modal regression method based on conditional mode assumption, named modal additive models (MAM) for simultaneous function estimation and structure identification. MAM utilizes the forward neural network as a tool to approximate the target function due to its universal approximation ability and low computational cost. By incorporating function space complexity and data structure complexity into the model, MAM can achieve the identification of the possible interactions between predictor variables and function estimation. Different from the prior-based variable selection methods, MAM is essentially a data-driven method and it can be implemented without strong data priors. Since MAM contains two different types of regularization terms, the generalization analysis cannot be conducted by traditional techniques, as in [8, 36]. To overcome this difficulty, we introduce a double regularized stepping stone function to formulate the error decomposition of MAM and then establish its asymptotic estimation and selection consistency. Furthermore, we develop an efficient backward stepwise algorithm to identify the group structure of the given data. The main contributions of this paper are summarized as follows:
Following the structural risk minimization framework, we propose a new MAM for estimating the function and identifying informative variables. MAM is essentially a data-driven structure identification strategy and can be efficiently implemented by half-quadratic minimization method.

In order to conduct learning theory analysis, we develop a new way to formulate the error decomposition. Both the estimation consistency and selection consistency are established. The estimation consistency implies that MAM estimator approximates to the optimal with satisfactory convergence rate. The selection consistency shows that MAM can identify the truly informative group structure with probability tending to one.

A backward stepwise implementation algorithm for MAM is proposed. The empirical studies with simulated data and real data demonstrate the effectiveness of MAM.

The rest of the paper is organized as follows: Section 2 sets the notations and problem statements. Section 3 introduces the proposed modal additive model for structural identification. Section 4 establishes the estimation consistency and selection consistency for MAM. Section 5 presents optimization strategy for MAM. Section 6 reports the experimental results on simulated data and real data. Section 7 concludes this paper with discussion.

2. Preliminaries.

2.1. Modal regression. Let $\mathcal{X} \in \mathbb{R}^d$ and $\mathcal{Y} \in \mathbb{R}$ be the input and output spaces respectively. In the modal regression setting, training samples $z = \{x_i, y_i\}_{i=1}^n \subset \mathcal{X} \times \mathcal{Y}$ are generated independently by

$$Y = f^*(X) + \varepsilon, \quad (1)$$

where the mode of the conditional distribution of $\varepsilon$ at any $x \in \mathcal{X}$ is assumed to be zero, i.e. $\text{mode}(\varepsilon | X = x) = \arg \max_t p_{\varepsilon | X}(t | X = x) = 0$ for any $x \in \mathcal{X}$, where $p_{\varepsilon | X}$ be the conditional density of $\varepsilon$ on $X$. Then, the target function of modal regression can be represented by

$$f^*(x) = \text{mode}(Y | X = x) = \arg \max_t p_{Y | X}(t | X = x), \quad \forall x \in \mathcal{X}. \quad (2)$$

Throughout this paper, we assume that $\arg \max_t p_{Y | X}(t | X = x)$ is well defined for any $x \in \mathcal{X}$, which is equivalent to the existence and uniqueness of the global mode of the conditional density $p_{Y | X}$.

Denote $\rho$ on $\mathcal{X} \times \mathcal{Y}$ as the intrinsic distribution for data generated by (1) and denote $\rho_X$ as the corresponding marginal distribution on $\mathcal{X}$. For any measurable function $f : \mathcal{X} \to \mathbb{R}$, the modal regression performance can be characterized by

$$\mathcal{R}(f) = \int_{\mathcal{X}} p_{Y | X}(f(x) | X = x) \, d\rho_X(x). \quad (3)$$

It has been proved that $f^*$ is the maximizer of (3) over all measurable functions [12]. Since $\rho_X$ and $p_{Y | X}$ are usually unknown, we can not obtain the estimator directly through maximizing (3). A recent study (see [12], Theorem 5) shows $\mathcal{R}(f) = p_{\varepsilon_f}(0)$, where $p_{\varepsilon_f}$ is the density function of random variable $\varepsilon_f = Y - f(X)$. This implies that maximizing $\mathcal{R}(f)$ over some hypothesis spaces is equivalent to maximizing the density of $\varepsilon_f$ at 0, which can be estimated by non-parametric kernel density estimation.
Let $K_\sigma : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ be a kernel function, and $\phi(u-u') = K_\sigma(u, u')$ be a representing function which satisfies $\phi(u) = \phi(-u)$, $\phi(u) \leq \phi(0)$ for any $u \in \mathbb{R}$ and $\int_{\mathbb{R}} \phi(u) \, du = 1$. Typical examples include the Gaussian kernel, Epanechnikov kernel, quadratic kernel and Triangular kernel. With the help of $K_\sigma$, we can obtain the empirical estimation of $R_K$ kernel, quadratic kernel and Triangular kernel. For any function $f$ to the usual additive model when $G \in H_{\phi}(\mathbb{R})$, and $\phi$ representing function which satisfies $R_\phi$. Particularly, there holds $\phi(u) = \phi(-u)$, $\phi(u) \leq \phi(0)$ for any $u \in \mathbb{R}$ and $\int_{\mathbb{R}} \phi(u) \, du = 1$. Typical examples include the Gaussian kernel, Epanechnikov kernel, quadratic kernel and Triangular kernel. With the help of $K_\sigma$, we can obtain the empirical estimation of $R(f)$ by kernel density estimation, given by

$$R_\sigma^u(f) = \frac{1}{n\sigma} \sum_{i=1}^{n} K_\sigma(y_i - f(x_i), 0) = \frac{1}{n\sigma} \sum_{i=1}^{n} \phi\left(\frac{y_i - f(x_i)}{\sigma}\right).$$

For any $f : \mathcal{X} \to \mathbb{R}$, the expectation version of $R_\sigma^u(f)$ is

$$R^\sigma(f) = \frac{1}{\sigma} \int_{\mathcal{X} \times \mathcal{Y}} \phi\left(\frac{y-f(x)}{\sigma}\right) \, d\rho(x, y).$$

Particularly, there holds $R(f) - R^\sigma(f) \to 0$ as $\sigma \to 0$ [12].

### 2.2. Group additive structures

Let $G := \{u_j\}_{j=1}^p$ be a index partition of the input variables with $u_j \cap u_k = \emptyset$ if $j \neq k$ and $\cup_{j=1}^p u_j = \{1, 2, \ldots, d\}$. Denote $x_{u_j} = \{x_k : k \in u_j\}$ for $j = \{1, 2, \ldots, p\}$, then $\{x_1, x_2, \ldots, x_d\} = x_{u_1} \cup \cdots \cup x_{u_p}$. For any function $f(x)$, if there exists an index partition $G = \{u_1, u_2, \ldots, u_p\}$ and corresponding $f_{u_1}, \ldots, f_{u_p}$ such that

$$f(x) = f_{u_1}(x_{u_1}) + \cdots + f_{u_p}(x_{u_p}),$$

then we say that $f(x)$ admits the group structure $G$. In particular, (4) can reduce to the usual additive model when $G = \{(1), (2), \ldots, (d)\}$.

To ensure the uniqueness of group additive structure for an estimator $f(x)$, the following concepts from [2] is helpful to characterize the partial order between different group structures.

**Definition 2.1.** Let $G$ and $G'$ be two group additive structures. If for every group $u \in G$ there is a group $v \in G'$ such that $u \subset v$, then $G$ is called a sub group additive structure of $G'$, denoted by $G \leq G'$. Equivalently, $G'$ is a super group additive structure of $G$, denoted by $G' \geq G$.

For a single group $u$, denote the square integrable function space w.r.t. $u$ by $L^2_u(\mathcal{X}) := \{g \in L^2(\mathcal{X})| g(x) = f_u(x_u)\}$. Then the group additive model $f(x) = \sum_{j=1}^p f_{u_j}(x_{u_j})$ is an element of the direct sum function space $L^2_{G}(\mathcal{X}) := \oplus_{u \in G} L^2_u(\mathcal{X})$. In particular, $L^2_u = L^2_{G}$ if $u$ is the unique element in group additive structure and $|u| = d$ ($|u|$ is the cardinality of the group $u$).

**Definition 2.2.** Suppose that $f \in L^2_{G}$. For a group additive structure $G$, if there is a function $f_G \in L^2_{G}$ such that $f_G = f$, then $G$ is called an amiable group additive structure for $f$.

**Lemma 2.3.** [2] Suppose $G^A$ is the set of amiable group additive structure for $f$. If there is a unique minimal group structure additive structure $G^* \in G^A$ such that $G^* \leq G$ for all $G \in G^A$, then $G^*$ is the intrinsic group additive structure for $f$.

One can see from Lemma 2.3 that $G^*$ can achieve the greatest dimension reduction for modeling the relationship between input and output. Generally, the intrinsic group additive structure serves well in alleviating the curse of dimensionality while improving both the accuracy and interpretability of high-dimensional nonparametric regression [20, 23, 19].
3. Modal additive model for structural identification. This section aims at describing the proposed MAM. The primal interest of MAM is to perform function estimation and variable selection simultaneously. As mentioned in the previous section, the estimation function \( f \) admits the group structure \( G \). For a given \( G = \{u_j\}_{j=1}^p \), define a Sobolev space associated with \( G \) as

\[
\mathcal{F}_G = \left\{ f(x) = \sum_{j=1}^p \sum_{t=1}^h a_{tj} \psi(w_{tj} x_{u_j} + b_t) : \forall |\alpha| < s, \partial^\alpha f \in L_2^G \right\},
\]

where \( h \) is the number of hidden nodes, \( a = (a_1, a_2, \cdots, a_h) \), \( \psi(\cdot) \) is the activation function, \( w_{tj} \) is the weight vector associated with the \( j \)-th group and \( t \)-th hidden node, and \( b_t \) is the corresponding bias term. The identification of \( f \) is up to three parameters: \( a, w_{tj}, b_t \). In general, \( w_{tj}, b_t \) are randomly and independent chosen beforehand. Hence, we only focus on estimating the weight parameter \( a \).

There are two cases in specifying the form of \( f \). One is that the intrinsic group structure for \( f(x) \) is known as \( G^* = \{u_j\}_{j=1}^p \). Let \( \Omega(f) = \left\{ |a|_2^2 : f \in \mathcal{F}_G \right\} \), then the estimator can be directly solved by regularized modal regression

\[
f_{G^*} = \arg \max_{f \in \mathcal{F}_G} \left\{ \mathcal{R}_z(f) - \lambda \Omega(f) \right\}.
\]

Obviously, the solution of (5) depends on the group additive structure \( G^* \) and the regularization parameter \( \lambda \). However, \( G^* \) is usually unknown in practice. In such a case, the primary interest should be identifying \( G^* \) first, and then implementing model (5). The previous section has shown that \( G^* \) exists and unique. The possible selection for \( G^* \) is boiling down to two categories: \( G \in G^* \) and \( G \not\in G^* \).

If \( G \not\in G^* \), let \( G \) be the group additive structure used in (5), then the obtained estimator \( f_G \) will result in a bias since there is always a gap between \( G \) and \( G^* \). The bias can not be erased regardless of the size of training examples. Therefore, it is impossible to find a perfect model when using a non-amiable group structure.

However, the situation is completely different under the case that \( G \in G^* \). Since \( G \) is an amiable group additive structure, we have \( f_{G^*} = f_G \) almost surely. The bias term will be eliminated when the training size is increased. It should be noted that the goodness of fit is not enough to distinguish \( G \) and \( G^* \) when the number of samples is kept fixed, since the asymptotic rates of \( f_G \) and \( f_{G^*} \) are usually different. The essential difference between \( G \) and \( G^* \) lies in the structure complexities, i.e., \( G^* \) has the smallest group structure, that is \( G^* \leq G \) for any \( G \in G^* \). Once the structure complexity of a group additive structure is well-defined and denoted as \( \mathcal{C}(G) \), we can incorporate it into (5) as an additional penalty term and the standard regularized modal regression model becomes the MAM model

\[
\{f_{\hat{G}}, \hat{G}\} = \arg \max_{f \in \mathcal{F}_G, G} \left\{ \mathcal{R}_z(f) - \lambda \Omega(f) - \mu \mathcal{C}(G) \right\},
\]

where \( \mu > 0 \) is a regularization parameter that controls the structure complexity. As discussed above, the intrinsic group structure \( G^* \) can achieve best prediction performance while keeping the smallest structural complexity. Hence, it is expected that the probability of \( \hat{G} = G^* \) will tend to 1 as the training size increases.

Next we will focus on specifying the structure complexity \( \mathcal{C}(G) \). Covering number \([36, 8, 35]\) plays a central role in characterizing the complexity of hypothesis space. Hence we prefer using covering number to design a practical measurement for group additive structures. Recall that a Sobolev space \( \mathcal{F}^s \) can be embedded in the
continuous function space $\mathcal{C}(\mathcal{X})$, with the inclusion mapping $I_s : \mathcal{F}^s(\mathcal{X}) \to \mathcal{C}(\mathcal{X})$.

Let

$$
\mathcal{B}_R = \left\{ f \in \mathcal{F}^G, f = \sum_{j=1}^{p} \sum_{l=1}^{h} a_{ij} \psi(w_{lj}, \cdot), \sum_{l=1}^{h} |a_{ij}|^2 \leq R \right\}
$$

(7)

be the closed ball of radius $R$ centered at the origin in $\mathcal{F}^G$ and $\overline{\mathcal{B}_R}$ be the closure of $I_s(\mathcal{F}^s_{G,R})$ in $\mathcal{C}(\mathcal{X})$. We first recall the definition of covering number.

**Definition 3.1.** Let $(\mathcal{M}, d)$ be a pseudo-metric space and $S \subset \mathcal{M}$. For every $\varepsilon > 0$, the covering number $\mathcal{N}(S, \varepsilon, d)$ of $S$ with respect to $\varepsilon$ and $d$ is defined as the minimal number of balls of radius $\varepsilon$ whose union covers $S$, that is

$$
\mathcal{N}(S, \varepsilon, d) = \min \left\{ n \in \mathbb{N}: S \subset \bigcup_{j=1}^{n} B(o_j, \varepsilon) \text{ for some } \{o_j\}_{j=1}^{n} \subset \mathcal{M} \right\},
$$

where $B(o_j, \varepsilon) = \{o \in \mathcal{M}: d(o, o_j) \leq \varepsilon\}$ is a ball in $\mathcal{M}$.

By the rich knowledge on covering number of Sobolev spaces [37, 8], we know that

$$
\ln \mathcal{N}(\overline{\mathcal{B}_R}, \varepsilon) \leq C_s \left( \frac{R}{\varepsilon} \right)^{\frac{2d}{s}},
$$

(8)

The growth rate of $\mathcal{N}(\overline{\mathcal{B}_R}, \varepsilon)$ or its logarithm version can be regarded as a complexity measure. Note that when $\varepsilon \to 0$ with fixed $R$, $(R/\varepsilon)^{2d/s}$ would be the dominant factor in the upper bound. Here, we use $(R/\varepsilon)^{2d/s}$ to characterize the complexity of Sobolev space. For a fixed $s$, define $\mathcal{C}(\mathcal{F}^s_G) = (R/\varepsilon)^{2d/s} = \beta(\varepsilon)^{2d/s}$. In (6), $\mathcal{F}^s_G$ is a direct sum of Sobolev spaces. Let $G = \{u_1, u_2, \ldots, u_p\}$, $\mathcal{F}^s_G, \mathcal{F}^s_{u_1}, \ldots, \mathcal{F}^s_{u_p}$ be the Sobolev spaces corresponding to $G, u_1, \ldots, u_p$ respectively; and let $I_{s,G}, I_{s,u_1}, \ldots, I_{s,u_p}$ be the inclusion mappings of $\mathcal{F}^s_G, \mathcal{F}^s_{u_1}, \ldots, \mathcal{F}^s_{u_p}$ into $\mathcal{C}(\mathcal{X})$. We then have the following propositions.

**Proposition 1.** Let $G^* = \{u_j\}_{j=1}^{p}$ be a index partition with $u_j \cap u_k = \emptyset$ for $j \neq k$, $\bigcup_{j=1}^{p} u_j = \{1, 2, \ldots, d\}$. Let $\Omega \subset \mathbb{R}^d$ and $f \in L^2(\Omega)$ be a function belonging to the Sobolev space $\mathcal{F}^s_G$, i.e. for any $|\alpha| < s$, $\partial^\alpha f \in L^2(\Omega)$. Suppose that $\mathcal{F}^s$ endow with inner product

$$
\langle f, g \rangle_{s, \Omega} = \sum_{|\alpha| < s} \int_{\Omega} \partial^\alpha f(x) \partial^\alpha g(x) \, dx,
$$

(9)

and the associated norm $\|f\|_{s, \Omega} = \sqrt{\sum_{|\alpha| < s} \|\partial^\alpha f(x)\|_{L^2(\Omega)}^2}$, then

$$
\mathcal{F}^s_G = \mathcal{F}^s_{u_1} \oplus \mathcal{F}^s_{u_2} \oplus \cdots \oplus \mathcal{F}^s_{u_p}
$$

is still a Sobolev space.

**Proposition 2.** Let $G$ be a group additive structure and $\mathcal{F}_G$ be the induced direct sum Sobolev spaces defined in Proposition 1. Then

$$
\ln \mathcal{N}(I_{s,G}, \varepsilon, \|\cdot\|_s) \leq \sum_{j=1}^{p} \ln \mathcal{N}(I_{u_j}, \frac{\varepsilon}{|G|}, \|\cdot\|_s),
$$

where $|G|$ denotes the number of groups in $G$.

Proposition 2 indicates that $\ln \mathcal{N}(I_{s,G}, \varepsilon, \|\cdot\|_s) = O(\sum_{u \in G} \beta(\varepsilon)^{|u|})$, which gives a explicit form of the complexity measure $\mathcal{C}(G)$, i.e. $\mathcal{C}(G) = \sum_{j=1}^{p} \beta(\varepsilon)^{|u_j|}$.
4. **Theoretical assessment for MAM.** This section mainly concerns about the
generalization ability of MAM. The following assumptions are helpful for establishing
the generalization error bounds. We first state the basic conditions on the
kernel-induced representing function $\phi$.

**Assumption 1.** The representing function $\phi$ satisfies the following conditions: 1) 
$\forall u \in \mathbb{R}, \phi(u) \leq \phi(0) < \infty$, 2) $\phi$ is Lipschitz continuous with constant $L_{\phi}$, 3) 
$\int_{\mathbb{R}} \phi(u) \, du = 1$ and $\int_{\mathbb{R}} u^2 \phi(u) \, du < \infty$.

The three conditions in Assumption 1 are satisfied by most kernels used in kernel
density estimation, such as Gaussian kernel, Quadratic kernel, Epanechnikov kernel
etc. Since the MAM is closely related to $\mathbb{R}$
density estimation, such as Gaussian kernel, Quadratic kernel, Epanechnikov kernel

**Assumption 2.** The conditional density $p_{\epsilon|X}$ is second-order continuously differ-
nable and uniformly bounded.

A well-established approach for conducting error analysis of learning algorithms
is error decomposition [8, 30, 14, 4], which usually divides the generalization error
into sample error and approximation error. However, due to the structural con-
straint in the estimation model (6), we cannot formulate the error decomposition
in a traditional way. To overcome this difficulty, we introduce the stepping stone
function [31]

$$
\{f_\lambda, G\} = \arg \max_{f \in F^*_G} \{\mathcal{R}^\sigma(f) - \lambda \Omega(f) - \mu \mathcal{C}(G)\}.
$$

(10)

Then we obtain the following error decomposition of MAM.

**Proposition 3.** Let $f^*_G \in F^*_G$. Under Assumptions 1 and 2, we have

$$
\mathcal{R}(f^*_G) - \mathcal{R}(f_G) \leq S_1 + S_2 + \sigma^2 |p''_{\epsilon|X}| \infty \int_{\mathbb{R}} u^2 \phi(u) \, du + \lambda \Omega(f^*_G) + \mu \mathcal{C}(G^*),
$$

where

$$
S_1 = \mathcal{R}^\sigma(f^*_G) - \mathcal{R}^\sigma(f_G) - \{\mathcal{R}^\sigma(f^*_G) - \mathcal{R}^\sigma(f_G)\},
$$

$$
S_2 = \mathcal{R}^\sigma(f^*_G) - \mathcal{R}^\sigma(f_G) - \{\mathcal{R}^\sigma(f^*_G) - \mathcal{R}^\sigma(f_G)\}.
$$

**Proof.** From Theorem 10 in [12], we know that

$$
\mathcal{R}(f^*_G) - \mathcal{R}(f_G) \leq \mathcal{R}^\sigma(f^*_G) - \mathcal{R}^\sigma(f_G) + \sigma^2 |p''_{\epsilon|X}| \infty \int_{\mathbb{R}} u^2 \phi(u) \, du.
$$

(11)

According to the definition of $f_\lambda$ and $f^*_G$, we have

$$
\mathcal{R}^\sigma(f^*_G) - \mathcal{R}^\sigma(f_G)
$$

$$
= \mathcal{R}^\sigma(f^*_G) - \lambda \Omega(f^*_G) - \mu \mathcal{C}(G^*) - \mathcal{R}^\sigma(f_G) + \lambda \Omega(f_G) + \mu \mathcal{C}(G^*)
$$

$$
\leq \mathcal{R}^\sigma(f_\lambda) - \lambda \Omega(f_\lambda) - \mu \mathcal{C}(G_\lambda) - \mathcal{R}^\sigma(f_G) + \lambda \Omega(f_G) + \mu \mathcal{C}(G^*)
$$

$$
= \mathcal{R}^\sigma(f_\lambda) - \mathcal{R}^\sigma(f_G)
$$

$$
+ \{\mathcal{R}^\sigma(f_\lambda) - \lambda \Omega(f_\lambda) - \mu \mathcal{C}(G_\lambda) - \mathcal{R}^\sigma(f_G) - \lambda \Omega(f_G) - \mu \mathcal{C}(G^*))\}
$$

$$
+ \mathcal{R}^\sigma(f^*_G) - \mathcal{R}^\sigma(f_G) + \lambda \Omega(f^*_G) + \mu \mathcal{C}(G^*) - \lambda \Omega(f_G) - \mu \mathcal{C}(G^*)
$$
\[
S_1 + S_2 + \lambda \Omega(f_G^*) + \mu \mathcal{C}(G^*)
\]
where the last inequality holds because that the term \( R^*_n(f_\lambda) - \lambda \Omega(f_\lambda) - \mu \mathcal{C}(G_\lambda) - (R^*_n(f_\lambda) - \lambda \Omega(f_\lambda) - \mu \mathcal{C}(G_\lambda)) \) is less than zero. This completes the proof. \( \square \)

**Remark 1.** It should be emphasized that the conventional error analysis \([8, 7, 26, 30]\) is usually conducted under a fixed input space. However, in our setting, \( F_G^* \) will change over group structure \( G \), thus the previous error decomposition does not hold directly. We overcome this difficulty by considering all possibilities of \( G \) and introducing the stepping stone function \( f_\lambda \), which is essentially a data independent function. For any \( G \in \mathcal{G} \), we have

\[
R^*_n(f_\lambda) - \lambda \Omega(f_\lambda) - \mu \mathcal{C}(G_\lambda) - (R^*_n(f_\lambda) - \lambda \Omega(f_\lambda) - \mu \mathcal{C}(G_\lambda)) \leq 0.
\]

With the above settings, we now present theoretical results for MAM. The detail proofs of these results can be found in the appendix.

**Theorem 4.1.** Let \( G^* \) be the intrinsic group additive structure, \( G \in \mathcal{G}^A \) is a given amiable group structure. Under Assumptions 1 and 2, for any \( \varepsilon > 0 \), if \( f_G^* \in F_G^* \), \( \sigma = n^{-\frac{d+\varepsilon}{2\tau+2\varepsilon}}, h = [n^{\frac{1}{2\tau+2\varepsilon}}], \lambda = n^{-\frac{1}{2\tau+2\varepsilon}} \) and \( \mu = n^{-\frac{\varepsilon}{2\tau+2\varepsilon}} \), with confidence at least \( 1 - \delta \), there holds

\[
\mathbb{E}_{f_G^*}(R(f_{G^*}^*) - R(f_G^*)) \leq Cn^{-\frac{d}{2\tau+2\varepsilon}+\varepsilon} \log n,
\]

where \( C \) is a constant independent of \( \delta \) and \( n \).

**Remark 2.** Theorem 4.1 shows that the estimation consistency of MAM is guaranteed as \( n \to \infty \). The corresponding generalization error \( R(f_{G^*}^*) - R(f_G^*) \) is with polynomial decay. It follows from Theorem 4.1 that, with a properly chosen \( h, \lambda, \mu \) and \( \sigma \), the MAM estimator \( f_G^* \) tends to the target function \( f_{G^*}^* \), with approximation order \( \mathcal{O}(n^{-\frac{d}{2\tau+2\varepsilon}} \log n) \). Specifically, when \( s = d/2 \), the learning rate of MAM can achieve \( \mathcal{O}(n^{-\frac{1}{4}} \log n) \), which is faster than \( \mathcal{O}(n^{-\frac{1}{4}}) \) in \([29]\).

Theorem 4.1 states the statistical consistency of the MAM under the condition that \( G \in \mathcal{G}^A \). Next we consider the case that \( G \) is a non-amiable group additive structure, i.e. \( G \in \mathcal{G} \setminus \mathcal{G}^A \). The result is stated as follows.

**Theorem 4.2.** Let

\[
f_{\tilde{G}} = \arg \max_{f \in F_{\tilde{G}}} \{ R^*_n(f) - \lambda \Omega(f) \}.
\]

Then, (i) If \( \tilde{G} \in \mathcal{G} \setminus \mathcal{G}^A \), then there exists a constant \( \tilde{C} > 0 \) such that \( |R(f_{\tilde{G}}) - R(f_{G^*}^*)| \) converges to \( \tilde{C} \) in probability. (ii) There exist a constant \( \tilde{C} \) such that

\[
limit_{n \to \infty} \mathbb{P}\{|R(f_{\tilde{G}}) - R(f_{G^*}^*)| > \tilde{C} \text{ for all } \tilde{G} \in \mathcal{G} \setminus \mathcal{G}^A\} = 1.
\]

**Remark 3.** It follows from the first statement of Theorem 4.2 that \( R(f_{\tilde{G}}) \) fails to converge to \( R(f_{G^*}^*) \) in probability when \( G \) is a non-amiable group structure. There is a positive constant gap \( \tilde{C} \) between \( R(f_{\tilde{G}}) \) and \( R(f_{G^*}^*) \). Note that the number of non-amiable group structures is finite, the second statement shows that, for all the non-amiable group structure, the probability of the event \( |R(f_{\tilde{G}}) - R(f_{G^*}^*)| > \tilde{C} \) tends to 1.
Theorem 4.3. Suppose the conditions of Theorem 4.1 are satisfied, then the estimated group structure $\hat{G}$ is consistent for the intrinsic group additive structure $G^*$, i.e. $\mathbb{P}(\hat{G} = G^*) \to 1$ as $n \to \infty$.

Remark 4. Theorem 4.3 shows that the selected group structure by MAM can exactly recover the intrinsic group structure with probability tending to 1. This result extends the selection consistency of the mean regression [2] to the modal regression.

5. Optimization. We utilize the half-quadratic (HQ) theory [13] to optimize MAM. Consider a convex optimization problem $\min_x P(x)$. It has an equivalent half-quadratic reformulation

$$\min_{x,y} Q(x,y) + S(y),$$

where $Q(x,y)$ is quadratic for any $y \in \mathbb{R}$ and $S : \mathbb{R} \to \mathbb{R}$ satisfies $P(x) = \min_y Q(x,y) + S(y), \forall x \in \mathbb{R}$. The potential dual function $S$ can be specified via convex conjugacy given as follows: for a closed convex function $f$, there exist a convex $g$ such that $f(x) = \max_y (xy - g(y))$, where $g$ is the conjugate of $f$, i.e. $g = f^\ast$. The following lemma [29] is important for designing our algorithm.

Lemma 5.1. Given a closed convex function $f(x) = \max_y (xy - g(y))$, for any $x \in \mathbb{R}$

$$\text{arg max}_y (xy - g(y)) = f^\prime(x).$$

Recall that Proposition 2 allows us to express the complexity measure $C(G)$ as $C(G) = \sum_{j=1}^{p} \beta(\epsilon)^{|u_j|}$. Thus, the explicit form for MAM can be reformulate as

$$\{f_G, \hat{G}\} = \arg \max_{f \in F_G} \{R^\ast_x(f) - \lambda \Omega(f) - \mu \sum_{j=1}^{p} \beta |u_j|\}. \quad (12)$$

If the value of $\lambda$ and $\mu$ are pre-specified, the optimization problem (12) can be solved by alternative optimization. First, when the group structure $G$ is given, according to Lemma 5.1, $f_G$ can be obtained by solving the following problem

$$\mathcal{R}_{\lambda,G} = \max_{f \in F_G} \{R^\ast_x(f) - \lambda \Omega(f)\}. \quad (13)$$

The implementation details are summarized in Algorithm 1. Second, the optimal group structure can be specified through

$$\hat{G} = \arg \max_{G \in \mathcal{G}} \left\{ \mathcal{R}_{\lambda} - \mu \sum_{j=1}^{p} \beta |u_j| \right\}. \quad (14)$$

A backward stepwise selection strategy is illustrated in Algorithm 2.

The above two-step optimization procedure is designed to find the intrinsic group structure, i.e. $\hat{G} = G^*$. Note that a group structure has three possible categories: intrinsic, amiable and non-amiable structures. If $G$ is non-amiable, which will result in a biased estimate and hence $\mathcal{R}_{\lambda}$ is expected to be small. If $G$ is amiable, although $\mathcal{R}_{\lambda}$ is expected to be large, the complexity of $G$ is larger than that of $G^*$. Therefore, only $G^*$ can achieve a large $\mathcal{R}_{\lambda}$ while remaining small complexity penalty (by Lemma 2.3).
Algorithm 1: Half-quadratic Optimization for MAM

1: Require: Input data \((x_i, y_i)_{i=1}^n\), kernel-induced representing function \(\phi\), activating function \(\psi\), weight parameter \(w\) and bias term \(b\).
2: Ensure: \(a_z\);
3: Define function \(f\) such that \(f(x^2) = \phi(x)\);
4: Initialize \(\sigma, a\);
5: while not converge do
6: Update \(e_i\) by \(e_i = f' \left( \frac{y_i - f(x_i)}{\sigma} \right)^2\);
7: Update \(a\) by \(a = \arg\max_{a \in \mathbb{R}^h} \frac{1}{n\sigma} \sum_{i=1}^n \left( e_i \left( \frac{y_i - f(x_i)}{\sigma} \right)^2 - g(e_i) \right) - \lambda \|a\|_2^2\);
8: update \(\sigma\);
9: end while
10: Output: \(a_z = a\).

Algorithm 2: Backward Stepwise Selection for MAM

1: Start with the variable pool \(G = \{(1, 2, \ldots, d)\}\);
2: Solve (13) to obtain the maximum value \(\mathcal{R}_{\lambda,G}\);
3: for each variable \(j\) in \(G\) do
4: \(\hat{G}\) ← either divide \(j\) into subgroups or add to an existing group;
5: Solve (13) to obtain the maximum value \(\mathcal{R}_{\lambda,\hat{G}}\);
6: if \(\mathcal{R}_{\lambda,\hat{G}} > \mathcal{R}_{\lambda,G}\) then
7: Preserve \(\hat{G}\) as the new group structure;
8: end if
9: end for
10: Return \(\hat{G}\).

6. Experiments. In this section, we evaluate the performance of MAM through simulation studies. Table 1 gives the three models we are using and the corresponding intrinsic group structures. 1000 observations of \(X\) are generated independently from \(N(0,1)\) in \(M_1\), \(U(-1,1)\) in \(M_2\), and \(U(0,2)\) in \(M_3\). For each model, we consider two types of noise \(\epsilon\), i.e. standard Gaussian noise and Student-t noise with degree of freedom 3. In all experiments, we set \(k(x, x') = \exp(-\frac{||x-x'||^2}{\sigma^2})\) as the modal regression kernel and \(h = 20\). The regularization parameter \(\lambda\) is determined by cross-validation strategy. The grid value of \(\mu\) is equally partitioned in the interval \([1e-8, 64]\) on the log-scale and each \(\beta\) is an integer in \([1, 5]\).

Table 1: Selected models for simulation study and the corresponding intrinsic group structures

| ID | Model | Intrinsic group structure |
|----|-------|--------------------------|
| M1 | \(y = x_1 + x_2^2 + \frac{1}{1+x_1} + \sin(\pi x_4) + \log(x_5 + 5) + \sqrt{|x_6|} + \epsilon\) | \(\{1, (2), (3), (4), (5), (6)\}\) |
| M2 | \(y = \frac{\sin(x_2)}{x_2} + \cos((x_2 + x_3) \cdot \pi) + \arctan((x_4 + x_5 + x_6)^2) + \epsilon\) | \(\{1, (2, 3), (4, 5, 6)\}\) |
| M3 | \(y = \sin(x_1 + x_2) + 2\log(x_3 + 5) + x_4 + x_5 \cdot x_6 + \epsilon\) | \(\{(1, 2), (3), (4, 5, 6)\}\) |
In order to show that MAM can identify the intrinsic additive group structure, we apply algorithm 2 for each \((\mu, \beta)\) pair 50 times. If the true group structure for each model can be often identified for pairs \((\mu, \beta)\), then we suppose that MAM has the capacity to identify the intrinsic group structure. The average performance measures are summarized in Tables 3 and 4. Specifically, MF denotes the maximum frequency that the intrinsic additive structure are identified. Size represents the average group number selected, TP represents the number of truly informative groups selected, and U, O are the probability of under-fitting and over-fitting, respectively. It can be observed from Tables 3 and 4 that under the different noise conditions, MAM cannot identify the intrinsic group structure of the given models when \(\beta = 1\), regardless \(\mu\) changing from \(10^{-6}\) to \(10^{-1}\). When increasing \(\beta\) with tuning \(\mu\), the intrinsic group structure can be successfully identified. Specifically, MAM can identify the intrinsic group structure with probability 1 for \(M_1\) and \(M_3\). The probability for \(M_2\) is over 50%. One can also observe that when choosing large \(\mu\) and \(\beta\), MAM is inclined to divide the intrinsic group into more subgroups. This is due to the heavy penalty imposed on model (12) that guarantees the regularization effect.

The prediction ability of MAM is evaluated through comparing with GASI [2] on model \(M_1, M_2, M_3\). Each method is applied \(K = 50\) times repeatedly. The quality of the fit is measure by the mean absolute error (MAE): \(\text{MAE} = \frac{1}{K} \sum_{i=1}^{K} | \hat{Y}_i - Y_i |\), where \(\hat{Y}_i\) denotes the prediction values of each time. Table 2 reports the MAE comparison results between MAM and GASI. One can observe that MAM outperforms GASI on \(M_1\) and \(M_2\) while obtain higher MAE on \(M_3\). It might be because that MAM has a greater potential to model the highly non-linear structures as demonstrated in \(M_1\) and \(M_2\).

We further test the performance of MAM on Boston Housing data, which describes the median value of owner-occupied homes in the Boston Standard Metropolitan Statistical Area in 1970. It contains 506 observations and 13 predictor variables, which are CRIM, ZN, INDUS, CHAS, NOX, RM, AGE, DIS, RAD, TAX, PTRATIO, B, and LSTAT. Our goal is to identify probable group structure for the predictor variables. Algorithm 1 is utilized and the tuning parameter \(\mu\) and \(\beta\) are specified by 5-fold cross validation. The group additive structure identified by MAM with smallest cv-error is \(\{(1, 6), (2, 9), (3, 7), (4, 11), (5), (8, 12), (10, 13)\}\). Then the non-parametric functions for each group are estimated according to the identified group structure. Note that the variables in each group is no more than two, hence the estimation functions can be visualized.

Some patterns can be discovered in Fig. 1. The top-left depicts the function of the average number of rooms per dwelling and per capita crime rate by town. It can be seen that the price of houses is increasing with the the number of rooms, and decreasing with the crime rate. The houses with 6 – 8 rooms are preferred.
when the crime rate is low. The top-left figure shows that the house value is increasing with the proportion of Blacks. This phenomenon may arise from the market discrimination during that period. It can be also observed that the house value is increasing with the increase of the proportion of owner units built prior to 1940 and full value property tax rate. Unit age is generally related to structure quality. The higher proportion of the owner units, the higher value for these houses. Additionally, higher tax rate means better ancillary facility, hence higher house price.

Table 3: Average performance that intrinsic group structures are identified for ($\mu, \beta$) pair (Gaussian noise)

| Parameters | M1 | M2 | M3 |
|------------|----|----|----|
| $\mu$  | $\beta$ | MF | Size | TP | U | O | MF | Size | TP | U | O | MF | Size | TP | U | O |
| $1e^{-6}$ | 1  | 0  | 2  | 1  | 0  | 0 | 2 | 0.66 | 1  | 0  | 0 | 2  | 1  | 0  | 1  |
| $1e^{-5}$ | 1  | 0  | 2  | 1  | 0  | 0 | 2 | 0.84 | 1  | 0  | 0 | 2  | 1  | 0  | 1  |
| $1e^{-4}$ | 1  | 0  | 2  | 1  | 0  | 0 | 2 | 0.68 | 1  | 0  | 0 | 2  | 0.1 | 1  | 0  |
| $1e^{-3}$ | 1  | 0  | 2  | 1  | 0  | 0 | 2 | 0.46 | 0.46 | 0  | 0 | 2  | 1  | 1  | 0  |
| $1e^{-2}$ | 1  | 0  | 2  | 1  | 0  | 0 | 2 | 0.62 | 0.62 | 0  | 0 | 2  | 1  | 1  | 0  |
| $1e^{-1}$ | 1  | 0  | 2  | 1  | 0  | 0 | 2 | 0.78 | 0.78 | 0  | 0 | 2  | 1  | 0  | 0  |
| $1e^{-6}$ | 3  | 0  | 3  | 2  | 1  | 0 | 2 | 0.42 | 0.42 | 0  | 0 | 2  | 0.66 | 0.66 | 0  |
| $1e^{-5}$ | 3  | 0  | 2.84 | 1.78 | 0.94 | 0  | 2 | 0.54 | 0.54 | 0  | 0 | 2  | 0  | 1  | 0  |
| $1e^{-4}$ | 3  | 0  | 3.36 | 2.32 | 1  | 0  | 2 | 0.58 | 0.58 | 0  | 0 | 2.2 | 1.6 | 1  | 0  |
| $1e^{-3}$ | 3  | 0  | 4.9 | 3.9 | 1  | 0  | 2 | 0.78 | 0.78 | 0  | 50 | 4  | 4  | 0  | 0  |
| $1e^{-2}$ | 3  | 50 | 6  | 6  | 0  | 0  | 29 | 3.62 | 1.9 | 0  | 0.22 | 50 | 4 | 4 | 0  | 0  |
| $1e^{-1}$ | 3  | 50 | 6  | 6  | 0  | 0  | 5.38 | 1.62 | 0  | 1  | 0 | 6  | 2  | 0 | 1  |
| $1e^{-6}$ | 5  | 0  | 2.72 | 1.64 | 0.92 | 0  | 2 | 0.5 | 0.5 | 0  | 0.23 | 6 | 0.1 | 1  | 0  |
| $1e^{-5}$ | 5  | 0  | 3.4 | 1.6 | 0.8 | 0  | 2 | 0.58 | 0.58 | 0  | 0 | 3  | 2  | 1  | 0  |
| $1e^{-4}$ | 5  | 0  | 4.82 | 3.82 | 1  | 0  | 2 | 0.38 | 0.38 | 0  | 50 | 4  | 4 | 0  | 0  |
| $1e^{-3}$ | 5  | 27 | 5.54 | 5.08 | 0.46 | 0  | 28 | 3.44 | 1.76 | 0 | 0  | 50 | 4 | 4 | 0  | 0  |
| $1e^{-2}$ | 5  | 50 | 6  | 6  | 0  | 0  | 5 | 2 | 0 | 1  | 0  | 6  | 2 | 0 | 1  |
| $1e^{-1}$ | 5  | 50 | 6  | 6  | 0  | 0  | 6 | 1 | 0 | 1  | 0  | 6  | 2 | 0 | 1  |

7. Conclusion. This paper proposed a new variable selection method, named MAM, for identifying the group structure of given data. MAM is essentially a data-driven method and it can be implemented without strong priors of data. Theoretically, we propose a new error decomposition strategy to conduct generalization analysis. Both of the estimation consistency and the selection consistency for MAM are established. Our result shows that it can identify intrinsic group structure with high probability. Simulation studies demonstrate the effectiveness of the proposed method.

Appendix.

A. Basic lemmas. Now introduce the well-known Bernstein inequality and the Bernstein-type concentration equality for function-valued random variables for our theoretical analysis.
Table 4: Average performance that intrinsic group structures are identified for \((\mu, \beta)\) pair (Gamma noise)

| Parameters | M1 | M2 | M3 |
|------------|----|----|----|
| \(\mu\)    | \(\beta\) | MF Size | TP | U | O | MF Size | TP | U | O | MF Size | TP | U | O |
| 1e\(-6\)   | 1   | 0  | 2  | 1 | 1 | 0 | 0  | 2  | 0.6 | 0 | 0  | 2  | 0.6 | 0 |
| 1e\(-5\)   | 1   | 0  | 2  | 1 | 1 | 0 | 0  | 2  | 0.7 | 0 | 0  | 2  | 0.7 | 0 |
| 1e\(-3\)   | 1   | 0  | 2  | 1 | 1 | 0 | 0  | 2  | 0.92 | 0 | 0  | 2  | 0.92 | 0 |
| 1e\(-2\)   | 1   | 0  | 2  | 1 | 1 | 0 | 0  | 2  | 0.58 | 0 | 0  | 2  | 0.58 | 0 |
| 1e\(-1\)   | 1   | 0  | 2  | 1 | 1 | 0 | 0  | 2  | 0.76 | 0 | 0  | 2  | 0.76 | 0 |
| 1e\(-6\)   | 3   | 0  | 2  | 1 | 1 | 0 | 0  | 2  | 0.52 | 0 | 0  | 2  | 0.52 | 0 |
| 1e\(-5\)   | 3   | 0  | 2  | 1 | 1 | 0 | 0  | 2  | 0.52 | 0 | 0  | 2  | 0.52 | 0 |
| 1e\(-4\)   | 3   | 0  | 3  | 2 | 1 | 0 | 0  | 2  | 0.8 | 0 | 0  | 2  | 0.8 | 0 |
| 1e\(-3\)   | 3   | 0  | 4  | 3 | 1 | 0 | 0  | 5  | 2.26 | 0.92 | 0 | 0  | 2  | 0.92 | 0 |
| 1e\(-2\)   | 3   | 42 | 5.84 | 5.88 | 0.16 | 0 | 27 | 3.66 | 1.82 | 0 | 0  | 2  | 0.8 | 0 |
| 1e\(-1\)   | 3   | 50 | 6  | 6 | 0 | 0 | 0  | 6  | 1 | 0 | 0  | 6  | 2 | 0 |
| 1e\(-6\)   | 5   | 0  | 2.56 | 1.48 | 1 | 0 | 0  | 2  | 0.62 | 0 | 0  | 2  | 0.62 | 0 |
| 1e\(-5\)   | 5   | 0  | 3.52 | 2.5 | 1 | 0 | 0  | 2  | 0.66 | 0 | 0  | 2  | 0.66 | 0 |
| 1e\(-4\)   | 5   | 7  | 4.88 | 3.76 | 0.86 | 0 | 24 | 3.08 | 1.8 | 0 | 0  | 2  | 0.8 | 0 |
| 1e\(-3\)   | 5   | 8  | 4.94 | 3.84 | 0.84 | 0 | 27 | 3.4 | 1.6 | 0 | 0  | 2  | 0.8 | 0 |
| 1e\(-2\)   | 2   | 50 | 6  | 6 | 0 | 0 | 0  | 5  | 2 | 0 | 0  | 5  | 2 | 0 |
| 1e\(-1\)   | 5   | 50 | 6  | 6 | 0 | 0 | 0  | 6  | 1 | 0 | 0  | 6  | 2 | 0 |

Lemma 7.1. [38] Let \(\xi\) be a random variable on a probability space \(Z\) with mean \(E\xi\) and variance \(\nu^2\). If \(|\xi - E(\xi)| \leq M_\xi\) for almost \(z \in Z\), then with confidence at least \(1 - \delta\)

\[
E\xi - \frac{1}{n} \sum_{i=1}^{n} \xi(z_i) \leq \frac{2M_\xi \log(1/\delta)}{3n} + \sqrt{\frac{2\nu^2 \log(2/\delta)}{n}}.
\]

Lemma 7.2. [8] Let \(G\) be the set of measurable functions on \(Z = X \times Y\) such that, for some \(c \geq 0\), \(|g - E_g| \leq B\) almost everywhere and \(E_g^2 \leq c \cdot E_g\) for each \(g \in G\), where \(E_g = \int_Z g(x, y) \, d\rho\). Then, for every \(\varepsilon > 0\)

\[
P \left\{ \sup_{g \in G} \frac{E_g - \frac{1}{n} \sum_{i=1}^{n} g(z_i)}{\sqrt{E(g) + \varepsilon}} \geq \varepsilon \right\} \leq N(G, \varepsilon) \exp \left\{ - \frac{n\varepsilon^2}{2c + \frac{3B^2}{4}} \right\}.
\]

Lemma 7.3. [8] Suppose \(B_R\) is defined in (7), then

\[
\log N(B_R, \varepsilon) \leq h \log \left( \frac{4R}{\varepsilon} \right).
\]

Lemma 7.4. [21] Let \(B_1\) and \(B_2\) be Banach spaces and \(S, T\) are two operators that satisfy \(B_1 \rightarrow B_2\). Let \(\varepsilon_1, \varepsilon_2 > 0\), then we have

\[
N(S + T, \varepsilon_1 + \varepsilon_2) \leq N(S, \varepsilon_1) \cdot N(T, \varepsilon_2).
\]

B. Proof of proposition 1. The expression in (9) defines an inner product. It is sufficient to show that the space is complete for the associate norm and hence a Hilbert space. Consider a Cauchy sequence \(\{f_k\}_{k \in \mathbb{N}}\) for this norm, we have

\[
\lim_{i,j \to \infty} \sum_{|\alpha| < s} \| \partial^\alpha f_i - \partial^\alpha f_j \|_{L^2(\Omega)}^2 = 0,
\]

(15)
which implies that for each $\alpha$ satisfies $|\alpha| \leq s$, the sequence $\{\partial^\alpha f_i\}$ is a Cauchy sequence in $L^2(\Omega)$. Then, we know that there exists $g_\alpha$ such that $\partial^\alpha f_i \rightarrow g_\alpha$ in $L^2(\Omega)$ due to the completeness of $L^2(\Omega)$. Since the convergence in $L^2(\Omega)$ implies the convergence in the distributional sense, we have $f_i \rightarrow g_0$. Therefore, $\partial^\alpha f_i \rightarrow \partial^\alpha g_0$ in the distributional sense and $\partial^\alpha g_0 = g_\alpha \in L^2(\Omega)$, $g_0 \in F^s(\Omega)$. Moreover, we have

$$\|g_0 - f_i\|_{h,\Omega}^2 = \sum_{|\alpha| < s} \|g_\alpha - \partial^\alpha f_i\|_{L^2(\Omega)}^2 \rightarrow 0,$$

thus $f_i \rightarrow g_0$ for the $F^s(\Omega)$ norm. Considering that the direct sum of Hilbert spaces is still a Hilbert space, we finish the proof.

C. Proof of proposition 2. According to Lemma 7.4, we have

$$\mathcal{N}(I_{s,G}, \varepsilon, \|\cdot\|_s) \leq \prod_{u \in G} \mathcal{N}\left(I_{s,F^s_u}, \frac{\varepsilon}{|G|}, \|\cdot\|_s\right) = \prod_{u \in G} \mathcal{N}\left(I_{s,u}, \frac{\varepsilon}{|G|}, \|\cdot\|_s\right),$$

then the result follows by taking log on both sides.

D. Proofs of Theorems 4.1-4.3. The error term $S_1$ defined in Proposition 3 is bounded as below by the Bernstein inequality in Lemma 7.1.
Proposition 4. Under Assumption 1, for any $0 < \delta < 1$, there holds
\[
S_1 \leq \frac{8\phi(0)\log(2/\delta)}{3n\sigma} + 2\sigma^{-1}\phi(0)\sqrt{\frac{2\log(2/\delta)}{n}}
\]
with confidence at least $1 - \delta$.

Proof. For any $(x, y) \in X \times Y$ and $f \in \mathcal{F}_G$, define
\[
\xi(x, y) = \sigma^{-1}\phi\left(\frac{y - f_{\hat{G}}(x)}{\sigma}\right) - \sigma^{-1}\phi\left(\frac{y - f(x)}{\sigma}\right).
\]
Under Assumption 1, we have $|\xi| \leq 2\sigma^{-1}\phi(0)$ and almost surely
\[
|\xi - \mathbb{E}\xi| \leq 4\sigma^{-1}\phi(0).
\]
Moreover,
\[
\mathbb{E}\xi^2 \leq 4\sigma^{-2}\phi^2(0).
\]
Then according to Lemma 7.1, we obtain
\[
S_1 \leq \frac{8\phi(0)\log(1/\delta)}{3n\sigma} + 2\sigma^{-1}\phi(0)\sqrt{\frac{2\log(2/\delta)}{n}}.
\]
with confidence at least $1 - \delta$. \qed

We now turn to estimate $S_2$ in proposition 3.

Proposition 5. Under Assumption 1, for any $0 \leq \delta \leq 1$, the following inequality
\[
S_2 \leq \frac{1}{2}\{\mathcal{R}^\sigma(f_{\hat{G}}) - \mathcal{R}^\sigma(f_{\hat{G}})\} + C\lambda^{-1}n^{-1}\sigma^{-4}
\]
holds with confidence at least $1 - \delta$, where $C$ is a constant depending only on $L_\phi, \phi(0), s, d$.

Proof. Let
\[
\mathcal{G}_R := \left\{\sigma^{-1}\phi\left(\frac{y - f_{\hat{G}}(x)}{\sigma}\right) - \sigma^{-1}\phi\left(\frac{y - f(x)}{\sigma}\right) : f \in B_R\right\},
\]
then we know
\[
g(z) = \sigma^{-1}\phi\left(\frac{y - f_{\hat{G}}(x)}{\sigma}\right) - \sigma^{-1}\phi\left(\frac{y - f(x)}{\sigma}\right). \tag{16}
\]
Observe that
\[
\mathbb{E}g = \mathcal{R}^\sigma(f_{\hat{G}}) - \mathcal{R}^\sigma(f_{\hat{G}})
\]
and
\[
\frac{1}{n}\sum_{i=1}^ng(z_i) = \mathcal{R}^\sigma_x(f_{\hat{G}}) - \mathcal{R}^\sigma_x(f_{\hat{G}}),
\]
where $z_i = (x_i, y_i)$. It follows that $|g(z)| \leq \frac{2\phi(0)}{\sigma}$, $|g - \mathbb{E}g| \leq \frac{4\phi(0)}{\sigma}$ and $\mathbb{E}g^2 \leq \frac{2\phi(0)}{\sigma}\mathbb{E}g$. Now we apply Lemma 7.2 with $B = \frac{4\phi(0)}{\sigma}$ and $c = \frac{2\phi(0)}{\sigma}$ to the set of $\mathcal{G}_R$, yielding
\[
\sup_{f \in \mathcal{F}_G \cap B_R} \frac{\mathcal{R}^\sigma(f_{\hat{G}}) - \mathcal{R}^\sigma(f_{\hat{G}}) - (\mathcal{R}^\sigma_x(f_{\hat{G}}) - \mathcal{R}^\sigma_x(f_{\hat{G}}))}{\sqrt{\mathcal{R}^\sigma_x(f_{\hat{G}}) - \mathcal{R}^\sigma_x(f_{\hat{G}})}}
\]
\[
= \sup_{g \in \mathcal{G}_R} \frac{\mathbb{E}(g) - \frac{1}{n}\sum_{i=1}^ng(z_i)}{\sqrt{\mathbb{E}(g) + \varepsilon}} \leq \sqrt{\varepsilon},
\]
which holds with confidence at least \(1 - \mathcal{N}(\mathcal{G}_R, \varepsilon)\) exp \(\{-\frac{3n\sigma}{20\phi(0)}\}\). Observe that for any \(g_1, g_2 \in \mathcal{G}_R\), there exist \(f_1, f_2 \in \mathcal{B}_R\) such that

\[
|g_1 - g_2| \leq \left\| \sigma^{-1} \phi\left(\frac{y - f_1(x)}{\sigma}\right) - \sigma^{-1} \phi\left(\frac{y - f_2(x)}{\sigma}\right) \right\|_\infty \leq L_\phi \|f_1 - f_2\|_\infty,
\]

which means for any \(\varepsilon > 0\), an \(\varepsilon\sigma^2/L_\phi\)-covering of \(\mathcal{B}_R\) provides an \(\varepsilon\)-covering of \(\mathcal{G}_R\), i.e.

\[
\mathcal{N}(\mathcal{G}_R, \varepsilon) \leq \mathcal{N}(\mathcal{B}_R, \frac{\varepsilon^2 \sigma^2}{L_\phi}).
\]

Then

\[
1 - \mathcal{N}(\mathcal{G}_R, \varepsilon) \exp \left\{-\frac{3n\sigma\varepsilon}{20\phi(0)}\right\} \geq 1 - \mathcal{N}(\mathcal{B}_R, \frac{\varepsilon^2 \sigma^2}{L_\phi}) \exp \left\{-\frac{3n\sigma\varepsilon}{20\phi(0)}\right\}.
\]

It follows from Lemma 7.3 that

\[
\log N\left(\mathcal{B}_R, \frac{\varepsilon^2 \sigma^2}{L_\phi}\right) \leq h \log \left(\frac{4L_\phi R}{\sigma^2 \varepsilon}\right),
\]

which implies

\[
\mathbb{P}\left\{\sup_{f \in \mathcal{F}_{\mathcal{G}_R}} \frac{\mathcal{R}_\sigma(f_{\mathcal{G}_R}) - \mathcal{R}_\sigma(f_{\hat{\mathcal{G}}}) - (\mathcal{R}_\sigma(f_{\hat{\mathcal{G}}}) - \mathcal{R}_\sigma(f_{\mathcal{G}_R}))}{\sqrt{\mathcal{R}_\sigma(f_{\mathcal{G}_R}) - \mathcal{R}_\sigma(f_{\hat{\mathcal{G}}}) + \varepsilon}} \leq \sqrt{\varepsilon}\right\} \geq 1 - \exp \left\{h \log \left(\frac{4L_\phi R}{\sigma^2 \varepsilon}\right) - \frac{3n\sigma\varepsilon}{20\phi(0)}\right\}
\]

\[
\geq 1 - \exp \left\{\frac{4L_\phi R h}{\sigma^2 \varepsilon} - \frac{3n\sigma\varepsilon}{20\phi(0)}\right\},
\]

The last inequality holds due to the elementary inequality \(\log(x) \leq x - 1\) for \(x > 0\). Set \(\frac{4L_\phi R h}{\sigma^2 \varepsilon} - \frac{3n\sigma\varepsilon}{20\phi(0)} = \delta\), then

\[
\varepsilon^2 + \frac{20\phi(0)(h + \ln(\delta))}{3n\sigma} \varepsilon - \frac{80\phi(0)L_\phi h R}{3n\sigma^3} = 0.
\]

Solve the equation about \(\varepsilon\), we get

\[
\varepsilon^* \leq \frac{80\phi(0)L_\phi h R}{3n\sigma^3}.
\]

It follows from the definition of \(f_{\hat{\mathcal{G}}}\) that

\[
\mathcal{R}_\sigma(f_{\hat{\mathcal{G}}}) - \lambda \Omega(f_{\hat{\mathcal{G}}}) \geq \mathcal{R}_\sigma(0),
\]

hence

\[
\Omega(f_{\hat{\mathcal{G}}}) \leq \frac{2\phi(0)}{\lambda \sigma}.
\]

By the fact that \(\sqrt{\varepsilon^* \hat{a}} \leq \frac{1}{2} a + \varepsilon(a > 0)\), for all \(f_{\hat{\mathcal{G}}} \in \mathcal{B}_R\) with \(R = \frac{2\phi(0)}{\lambda \sigma}\), we conclude that

\[
\mathcal{S}_2 \leq \frac{1}{2} (\mathcal{R}_\sigma(f_{\hat{\mathcal{G}}}) - \mathcal{R}_\sigma(f_{\mathcal{G}_R})) + Ch\lambda^{-1}n^{-1}\sigma^{-4},
\]

(17)

with confidence at least \(1 - \delta\), where \(C = \frac{80L_\phi \phi(0)}{3}\). This finishes the proof. \(\square\)
Proof of Theorem 4.1. By combining Propositions 3, 4, and 5, we know that with confidence at least $1 - 2\delta$
\[
\mathcal{R}(f_{G,\cdot}) - \mathcal{R}(f_{\bar{G}}) \leq C \log(2/\delta) \left( n^{-\frac{1}{2}} \sigma^{-1} + h \lambda^{-1} \sigma^{-4} + \sigma^2 + \mu \lambda^{-\frac{3}{2}} \sigma^{-\frac{7}{2}} \right),
\]
where $C$ is a constant depending only on $\phi(0), L_\phi$. Set $\sigma = n^{-\frac{\delta}{4+\epsilon}}$, $h = \lfloor n^{\frac{\delta}{4+\epsilon}} \rfloor$, $\lambda = n^{-\frac{\delta}{4+\epsilon}}$ and $\mu = n^{-\frac{2\delta}{4+\epsilon}}$, we obtain
\[
\mathcal{R}(f_{G,\cdot}) - \mathcal{R}(f_{\bar{G}}) \leq C \log(2/\delta) n^{-\frac{\delta}{4+\epsilon} + \epsilon}.
\]
Note that
\[
\mathbb{E}_{\rho^n}(\mathcal{R}(f_{G,\cdot}^*) - \mathcal{R}(f_{\bar{G}})) = \int_{0}^{\infty} \mathbb{P}(\mathcal{R}(f_{G,\cdot}^*) - \mathcal{R}(f_{\bar{G}}) > \varepsilon) \, d\varepsilon,
\]
direct computation yields to the upper bound. This completes the proof.

Proof of Theorem 4.2. Observe that
\[
|\mathcal{R}(f_{G}) - \mathcal{R}(f_{G,\cdot})| \leq |\mathcal{R}(f_{G}) - \mathcal{R}^\sigma(f_{G})| + |\mathcal{R}^\sigma(f_{G}) - \mathcal{R}^\sigma(f_{G,\cdot})| + |\mathcal{R}^\sigma(f_{G,\cdot}) - \mathcal{R}(f_{G,\cdot})|.
\]
It has been proved in [12, 29] that
\[
|\mathcal{R}^\sigma(f) - \mathcal{R}(f)| \leq \frac{\sigma^2}{2} \left( \|p''_{\rho}\|_\infty \int_{\mathbb{R}} u^2 \phi(u) \, du \right),
\]
(19)
Hence the first term and the last term of (18) can be bounded by $c_1 \sigma^2$, where $c_1 = \|p''_{\rho}\|_\infty \int_{\mathbb{R}} u^2 \phi(u) \, du$.

Let $\{P_j\}_{j=1}^{d} \subseteq \{1, \ldots, p\}$ be a set of commuting projections on $\mathcal{F}$ that satisfy: $P_j(f) = f$ if $f$ does not depend on $x_j$, and $P_j(f)$ does not depend on $x_j$. Following the function decomposition techniques developed by [22], we can express $f_{G}$ and $f_{G,\cdot}$ as
\[
f_{G} = \sum_{u \subseteq \{1, \ldots, p\}} \hat{f}_{G,u} \quad \text{with} \quad \hat{f}_{G,u} := \sum_{v \subseteq u} (-1)^{|u| - |v|} P_{\{1, \ldots, p\} \setminus v} (f_{G}),
\]
\[
f_{G,\cdot} = \sum_{u \subseteq \{1, \ldots, p\}} f_{G,\cdot,u} \quad \text{with} \quad f_{G,\cdot,u} := \sum_{v \subseteq u} (-1)^{|u| - |v|} P_{\{1, \ldots, p\} \setminus v} (f_{G,\cdot}),
\]
Note that $\bar{G}$ is non-amiable group structure, there is at least one subset of $u \subseteq \{1, \ldots, p\}$ such that $f_{G,\cdot,u} \neq \hat{f}_{G,u}$. Denote $u^c = \{1, \ldots, p\} \setminus u$,
\[
c_2 = L_\phi \max_{u \subseteq \{1, \ldots, p\}} \int_{X \times Y} |\hat{f}_{G,u} - f_{G,\cdot,u}| \, d\rho(x, y),
\]
\[
c_3 = L_\phi \max_{u \subseteq \{1, \ldots, p\}} \int_{X \times Y} |\hat{f}_{G,u}(x_{u^c}) - f_{G,\cdot,u}(x_{u^c})| \, d\rho(x_{u^c}, y),
\]
then
\[
|\mathcal{R}^\sigma(f_{G}) - \mathcal{R}^\sigma(f_{G,\cdot})| \leq \sigma^{-1} L_\phi \int_{X \times Y} |f_{G}(x) - f_{G,\cdot}(x)| \, d\rho(x, y)
\]
\[
= \sigma^{-1} L_\phi \int_{X_u \times Y} |\hat{f}_{G,u}(x_u) - f_{G,\cdot,u}(x_u)| \, d\rho(x_u, y)
\]
\[
+ \sigma^{-1} L_\phi \int_{X_{u^c} \times Y} |\hat{f}_{G,u^c}(x_{u^c}) - f_{G,\cdot,u^c}(x_{u^c})| \, d\rho(x_{u^c}, y)
\]
\[
\leq (c_2 + c_3) \sigma^{-1}
\]
(20)
where the first inequality is due to the Lipschitz property of $\phi$, and the equality is due to orthogonality possessed by a direct sum Sobolev space.

Combining the estimation (18), (19), (20), we have

$$|R(f_G) - R(f_{G^*})| \leq \epsilon + (c_2 + c_3)\sigma^{-1},$$

where $\epsilon = c_1^2 \sigma^2$. Theorem 4.1 tells us that $\sigma^2 = n^{-\frac{d}{2(d+2)}}$ can guarantee the fast convergence rate, hence $\epsilon \sim O(n^{-\frac{d}{2(d+2)}})$. The second term becomes $(c_2 + c_3)n^{-\frac{d}{d+2}}$, which means no matter how much samples are given, $|R(f_G) - R(f_{G^*})|$ can not converge to 0, the conclusion follows.

**Proof of Theorem 4.3.** Theorem 4.2 tells us that the amiable group structure would be identified with high probability when choosing $\epsilon < C$. Denote $\epsilon_1 = |R(f_G) - R(f_{G^*})|$ and $\epsilon_2 = \mu C(G) - \mu C(G^*)$. Note that $G$ is not the true group structure, hence $C(G) \geq C(G^*)$. For an amiable group structure $G$, we know from Theorem 4.1 that $\epsilon_1$ converges to 0 in probability. Therefore, the true group structure will be identified with high confidence through solving the problem (6).

**Acknowledgments.** This work was supported in part by the NSERC grant RGPIN-2016-05024 and in part by National Natural Science Foundation of China (NSFC) under Grant 11690014, 11671161, 11971373, 61806162. The content is solely the responsibility of the authors and does not necessarily represent the official views of the aforementioned funding agencies.

**REFERENCES**

[1] L. Breiman and J. Friedman, Estimating optimal transformations for multiple regression and correlation, *Journal of the American Statistical Association*, **80** (1985), 580–598.

[2] P. Chao and M. Zhu, Group additive structure identification for kernel non-parametric regression, *Advances in Neural Information Processing Systems*, (2017).

[3] H. Chen, X. Wang, C. Deng and H. Huang, Group sparse additive machine, *Advances in Neural Information Processing Systems*, (2017).

[4] H. Chen and Y. L. Wang, Kernel-based sparse regression with the correntropy-induced loss, *Appl. Comput. Harmon. Anal.*, **44** (2018), 144–164.

[5] Y.-C. Chen, R. Genovese, R. Tibshirani and L. Wasserman, Nonparametric modal regression, *Annals of Statistics*, **44** (2016), 489–514.

[6] G. Collomb, W. Härde and S. Hassanis, A note on prediction via estimation of the conditional mode function, *Journal of Statistical Planning and Inference*, **15** (1986), 227–236.

[7] F. Cucker and S. Smale, Best choices for regularization parameters in learning theory: On the bias-variance problem, *Foundations of Computational Mathematics*, **2** (2002), 413–428.

[8] F. Cucker and S. Smale, On the mathematical foundations of learning, *Bulletin of the American Mathematical Society*, **39** (2002), 1–49.

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

[10] J. Q. Fan and R. Z. Li, Variable selection via non-concave penalized likelihood and its oracle properties, *Journal of the American Statistical Association*, **96** (2001), 1348–1360.

[11] J. Q. Fan and J. C. Lv, Sure independence screening for ultrahigh dimensional feature space, *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **70** (2008), 849–911.

[12] Y. Feng, J. Fan and Y. Suykens, A statistical learning approach to modal regression, *Journal of Machine Learning Research*, **21** (2020), 1–35.

[13] D. Geman and C. Yang, Nonlinear image recovery with half-quadratic regularization, *IEEE Transactions on Image Processing*, **4** (1995), 932–946.

[14] T. L. Gong, Z. B. Xu and H. Chen, Generalization analysis of Fredholm kernel regularized classifiers, *Neural Computation*, **29** (2017), 1879–1901.

[15] C. Gu, *Smoothing Spline ANOVA Models*, Second edition, Springer Series in Statistics, 297. Springer, New York, 2013.
[16] X. He, J. Wang and S. Lv, Scalable kernel-based variable selection with sparsistency, preprint, arXiv:1802.09246.
[17] J. Huang, J. Horowitz and F. R. Wei, Variable selection in nonparametric additive models, *Annals of Statistics*, 38 (2010), 2282–2313.
[18] J. Huang and L. J. Yang, Identification of non-linear additive autoregressive models, *Journal of the Royal Statistical Society, Series B*, 66 (2004), 463–477.
[19] J. Huang, S. G. Ma and C.-H. Zhang, Adaptive lasso for sparse high-dimensional regression models, *Statistica Sinica.*, 18 (2008), 1603–1618.
[20] K. Kandasamy and Y. Yu, Additive approximations in high-dimensional non-parametric regression via the salsa, *International Conference on Machine Learning*, (2016).
[21] T. Kühn, Covering numbers of Gaussian reproducing kernel Hilbert spaces, *Journal of Complexity*, 27 (2011), 489–499.
[22] F. Kuo, G. Sloan, G. Wasiakowski and H. Woźniakowski, On decompositions of multivariate functions, *Mathematics of Computation, Mathematics of Computation*, 79 (2010), 953–966.
[23] Y. Lin and H. Zhang, Component selection and smoothing in multi-variate nonparametric regression, *Annals of Statistics*, 34 (2006), 2272–2297.
[24] T. Sager and R. Thisted, Maximum likelihood estimation of isotonic modal regression, *Annals of Statistics*, 10 (1982), 690–707.
[25] X. T. Shen, W. Pan and Y. Z. Zhu, Likelihood-based selection and sharp parameter estimation, *Journal of the American Statistical Association*, 107 (2012), 222–232.
[26] L. Shi, Y.-L. Feng and D.-X. Zhou, Concentration estimates for learning with ℓ1-regularizer and data dependent hypothesis space, *Applied and Computational Harmonic Analysis*, 31 (2011), 286 – 302.
[27] T. Shively, R. Kohn and S. Wood, Variable selection and function estimation in additive non-parametric regression using a data-based prior, *Journal of the American Statistical Association*, 94 (1999), 777–794.
[28] R. Tibshirani, Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society, Series B*, 58 (1996), 267–288.
[29] X. Wang, H. Chen, W. Cai, D. Shen and H. Huang, Regularized modal regression with applications in cognitive impairment prediction, *Advances in Neural Information Processing Systems*, (2017).
[30] Q. Wu, Y. M. Ying and D.-X. Zhou, Multi-kernel regularized classifiers, *Journal of Complexity*, 23 (2007), 108–134.
[31] Q. Wu and D.-X. Zhou, Learning with sample dependent hypothesis spaces, *Computers and Mathematics with Applications*, 56 (2008), 2896–2907.
[32] W. Yao, R. Lindsay and R. Li, Local modal regression, *Journal of Nonparametric Statistics*, 24 (2012), 647–663.
[33] J. Yin, X. Chen and E. Xing, Group sparse additive models, *International Conference on Machine Learning*, (2012).
[34] M. Yuan and Y. Lin, Model selection and estimation in regression with grouped variables, *Journal of the Royal Statistical Society, Series B*, 68 (2006), 49–67.
[35] T. Zhang, Covering number bounds of certain regularized linear function classes, *Journal of Machine Learning Research*, 2 (2002), 527–550.
[36] D.-X. Zhou, The covering number in learning theory, *Journal of Complexity*, 18 (2002), 739–767.
[37] D.-X. Zhou, Capacity of reproducing kernel space in learning theory, *IEEE Transactions on Information Theory*, 49 (2003), 1743–1752.
[38] D.-X. Zhou and K. Jetter, Approximation with polynomial kernels and SVM classifiers, *Advances in Computational Mathematics*, 25 (2006), 323–344.
[39] H. Zou, The adaptive lasso and its oracle properties, *Journal of the American Statistical Association*, 101 (2006), 1418–1429.

Received January 2020; 1st revision March 2020, final revision May 2020.

E-mail address: adidasgtl@gmail.com
E-mail address: chenh@mail.hzau.edu.cn
E-mail address: cx3@uottawa.ca