entropy

MDPI

Article

# Error Bound of Mode-Based Additive Models

**Hao Deng [1], Jianghong Chen [2], Biqin Song [1,*] and Zhibin Pan [1,*]**

[1] College of Science, Huazhong Agricultural University, Wuhan 430070, China; dengh@mail.hzau.edu.cn
[2] College of Electrical and New Energy, China Three Gorges University, Yichang 443002, China; chenjh97@126.com
* Correspondence: biqin.song@mail.hzau.edu.cn (B.S.); pzbhallow@mail.hzau.edu.cn (Z.P.)

**Abstract:** Due to their flexibility and interpretability, additive models are powerful tools for high-dimensional mean regression and variable selection. However, the least-squares loss-based mean regression models suffer from sensitivity to non-Gaussian noises, and there is also a need to improve the model's robustness. This paper considers the estimation and variable selection via modal regression in reproducing kernel Hilbert spaces (RKHSs). Based on the mode-induced metric and two-fold Lasso-type regularizer, we proposed a sparse modal regression algorithm and gave the excess generalization error. The experimental results demonstrated the effectiveness of the proposed model.

**Keywords:** modal regression; additive models; reproducing kernel Hilbert spaces; error bound

check for updates

**Citation:** Deng, H.; Chen, J.; Song, B.; Pan, Z. Error Bound of Mode-Based Additive Models. *Entropy* **2021**, *23*, 651. https://doi.org/10.3390/e23060651

Academic Editor: Ercan Kuruoglu

Received: 22 March 2021
Accepted: 19 May 2021
Published: 22 May 2021



## 1. Introduction

Regression estimation and variable selection are two important tasks for high-dimensional data mining [1]. Sparse additive models [2,3], aiming to deal with the above tasks simultaneously, have been extensively investigated in the mean regression setting. As a class of models between linear and nonparametric regression, these methods inherit the flexibility from nonparametric regression and the interpretability from linear regression. Typical methods include COSSO [4] and SpAM [2] and its variants, such as Group SpAM [3], SAM [5], Group SAM [6], SALSA [7], MAM [8], SSAM [9], and ramp-SAM [10]. From the lens of nonparametric regression, the additive structure on the hypothesis space is crucial to overcome the curse of dimensionality [7,11,12].

Usually, the aforementioned models are limited to the estimation of the conditional mean under the mean-squared error (MSE) criterion. However, for the complex non-Gaussian noises (e.g., the skewed noise, the heavy-tailed noise), it is difficult to extract the intrinsic trends from the mean-based approaches, resulting in degraded performance. Beyond the traditional mean regression, it is interesting to formulate a new regression framework under the (conditional) mode-based criterion. With the help of the recent works in [13–19], this paper aimed to propose a new robust sparse additive model, rooted in modal regression associated with the RKHS.

As an alternative approach to mean regression, modal regression has been investigated on statistical behavior [14,15,17] and real-world applications [20,21]. Yao [14] proposed a modal linear regression algorithm and characterized its theoretical properties under the global mode assumption. As a natural extension of Lasso [22], Wang et al. [15] considered the regularized modal regression and established its analysis on the generalization bound and variable selection consistency. Feng et al. [17] studied modal regression by a learning theory approach and illustrated its relation with MCC [23,24]. Different from the above global approaches, some local modal regression algorithms were formulated in [16,25] with convergence guarantees. Recent literature [26] gave a general overview of modal regression, and a more comprehensive list of references can be found there.

The proposed robust additive models are formulated under the Tikhonov regularization scheme, which involves three building blocks, including the mode-based metric, the RKHS-based hypothesis space, and two Lasso-type penalties. Since the linear function space, polynomial function space, and Sobolev/Besov space are special cases of the RKHS, the kernel-based function space is more flexible than the traditional spline-based spaces or other dictionary-based hypotheses [2,5,27–29]. The mode-induced regression metric is robust to the non-Gaussian noise according to the theoretical and empirical evaluations [14,15,17]. The regularized penalty addresses the sparsity and smoothness of the estimator, which has shown promising performance for mean regression [2,29–31]. Therefore, different from mean-based kernel regression and additive models, the mode-based approach enjoys robustness and interpretability simultaneously due to its metric criterion and trade-off penalty. The estimator of our approach can be obtained by integrating the half-quadratic (HQ) optimization [32] and the second-order cone programming (SOCP) [33].

The rest of this article is organized as follows. After introducing the robust additive model in Section 2, we state its generalization error bound in Section 3. Finally, Section 5 ends this paper with a brief conclusion.

## 2. Methodology

### 2.1. Modal Regression

In this section, we recall the basic background on modal regression [19,34]. Let $\mathcal{X}$ be a compact subset of $\mathbb{R}^p$ associated with the input covariate vector and $\mathcal{Y} \in \mathbb{R}$ be the response variable set. In this paper, we considered the following nonparametric model:

$$Y = f^*(X) + \epsilon, \tag{1}$$

where $X = (X_1, \dots, X_p)^T \in \mathcal{X}$, $Y \in \mathcal{Y}$, and $\epsilon$ is a random noise. For feasibility, we denote by $\rho$ the underlying joint distribution of $(X, Y)$ generated by (1).

Being different from the traditional mean regression under the noise condition $E(\epsilon|X = x) = 0$ (e.g., Gaussian noise), we just require that the mode of the conditional distribution of $\epsilon$ equal zero at each $x \in \mathcal{X}$. That is:

$$\forall x \in \mathcal{X}, mode(\epsilon|X = x) = \arg\max_{t\in\mathbb{R}} P_{\epsilon|X}(t|X = x) = 0, \tag{2}$$

where $P_{\epsilon|X}$ is the conditional density of $\epsilon$ given $X$. Notice that the zero condition is not specified to the homogeneity or symmetry distribution of noise $\epsilon$, and some non-Gaussian noises (e.g., the skewed noise, the heavy-tailed noise) are not excluded.

From (1), we further deduce that:

$$f^*(u) \quad := \quad \sum_{j=1}^{p} f_j^*(u_j) = mode(Y|X = u) = \arg\max_t P_{Y|X}(t|X = u),$$

where $u = (u_1, \dots, u_p)^T \in \mathcal{X}$ and $P_{Y|X}$ denotes the density of $Y$ conditional on $X$. Then, the purpose of modal regression is to find the target function $f^*$ according to the empirical data $\mathbf{z} = \{z_i\}_{i=1}^n = \{(x_i, y_i)\}_{i=1}^n$ drawn independently from $\rho$.

For modal regression, the performance of a predictor $f : \mathcal{X} \to \mathbb{R}$ is measured by the mode-based metric:

$$\mathcal{R}(f) = \int_{\mathcal{X}} P_{Y|X}(f(x)|X = x)d\rho_{\mathcal{X}}(x), \tag{3}$$

where $\rho_{\mathcal{X}}$ is the marginal distribution of $\rho$ with respect to input space $\mathcal{X}$.

Although the target function $f^*$ is the maximizer of $\mathcal{R}(f)$ over all measurable functions, it cannot be estimated directly via maximizing (3) due to the unknown $P_{Y|X}$ and $\rho_{\mathcal{X}}$. Fortunately, some indirect density-estimation-based strategies were proposed in [14,15,17]. As shown in Theorem 5 of [17], $\mathcal{R}(f)$ equals the density function of random variable $E_f = Y - f(X)$ at zero, e.g.,

$$\mathcal{R}(f) = P_{E_f}(0).$$

Therefore, we can find an approximation of $f^*$ by maximizing the empirical version of $P_{E_f}(0)$ with the help of kernel density estimation (KDE).

Let $K_\sigma : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ be a kernel with bandwidth $\sigma$, and its representing function $\phi : \mathbb{R} \to [0, \infty)$ satisfies $\phi(\frac{u-u'}{\sigma}) = K_\sigma(u, u'), \forall u, u' \in \mathbb{R}$. Typical kernels used in KDE include the Gaussian kernel, the Epanechnikov kernel, the logistic kernel, and the sigmoid kernel. The KDE-based estimator of $P_{E_f}(0)$ is defined as:

$$\hat{P}_{E_f}(0) = \frac{1}{n\sigma} \sum_{i=1}^{n} K_\sigma(y_i - f(x_i), 0) = \frac{1}{n\sigma} \sum_{i=1}^{n} \phi(\frac{y_i - f(x_i)}{\sigma}) := \hat{\mathcal{R}}^\sigma(f).$$

Learning models for modal regression are usually formulated by Tikhonov regularization schemes associated with the empirical metric $\hat{\mathcal{R}}^\sigma(f)$; see, e.g., [15,35].

Naturally, the data-free modal regression metric, *w.r.t.* $\hat{\mathcal{R}}^\sigma(f)$, can be defined as:

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

In theory, the learning performance of estimator $f : \mathcal{X} \to \mathbb{R}$ can be evaluated in terms of $\mathcal{R}(f) - \mathcal{R}(f^*)$, which can be further bounded via $\mathcal{R}^\sigma(f) - \mathcal{R}^\sigma(f^*)$ (see Theorem 10 in [17]).

**Remark 1.** *As illustrated in [17], when taking $K_\sigma$ as a Gaussian kernel, the modal regression for maximizing $\mathcal{R}^\sigma(f)$ is consistent with learning under the maximum correntropy criterion (MCC). By employing different kernels, we can provide rich evaluated metrics for better robust estimation.*

### *2.2. Mode-Based Sparse Additive Models*

The additive model is formulated as follows,

$$Y = \sum_{j=1}^{p} f_j^*(X_j) + \epsilon, \tag{4}$$

where $X_j \in X$, $(j = 1, 2, \cdots, p)$, $Y \in \mathcal{Y}$, and $f_j^*$ are unknown component functions. By employing nonlinear hypothesis function spaces with an additive structure, the additive model provides better flexibility for regression estimation and variable selection [19]. In [28], the theoretical properties of the sparse additive model with the quantile loss function were discussed. We introduce some basic notation and assumptions in a similar way.

Suppose that $Ef_j^*(X_j) = 0$ and $\|f_j^*\|_{K_j} \leq 1$ for each $f_j^*$ in (4) with $j \in \mathcal{S}$. Here, $f_j^* : \mathcal{X}_j \to \mathbb{R}$ is an unknown univariate function in a reproducing kernel Hilbert space (RKHS) $\mathcal{H}_j := \mathcal{H}_{K_j}$ associated with kernel $K_j$ and norm $\|\cdot\|_{K_j}$ [30,31], and $\mathcal{S} \subseteq \{1, \ldots, p\}$ is an intrinsic subset with cardinality $|\mathcal{S}| < p$. This means each observation $(x_j, y_j)$ is generated according to:

$$y_i = \sum_{j \in \mathcal{S}} f_j^*(x_{ij}) + \epsilon_i, i = 1, \ldots, n,$$

where $x_i = (x_{i1}, \ldots, x_{ip})^T \in \mathbb{R}^p$, $f_j^* \in \mathcal{H}_j$ and $\epsilon$ satisfies the condition (2).

For any given $j \in \{1, \ldots, p\}$, denote $\mathcal{B}_r(\mathcal{H}_j) = \{g \in \mathcal{H}_j : \|g\|_{K_j} \leq r\}$. The hypothesis space considered here is defined by:

$$\mathcal{F} = \{f = \sum_{j=1}^{p} f_j : f_j \in \mathcal{B}_r(\mathcal{H}_j), i = 1, \ldots, p\}, \tag{5}$$

which is a subset of the RKHS $\mathcal{H} = \{f = \sum_{j=1}^{p} f_j : f_j \in \mathcal{H}_j\}$ with the norm:

$$\|f\|_K^2 = \inf\{\sum_{j=1}^{p} \|f_j\|_{K_j}^2 : f = \sum_{j=1}^{p} f_j\}.$$

For each $\mathcal{X}_j$ and the corresponding marginal distribution $\rho_{\mathcal{X}_j}$, we denote $\|f_j\|_2^2 := \int_{\mathcal{X}_j} |f_j(u)|^2 d\rho_{\mathcal{X}_j}(u)$. Given inputs $\{x_i\}_{i=1}^n$, define the empirical norm of each $f_j$ as:

$$\|f_j\|_n^2 := \frac{1}{n}\sum_{i=1}^{n} f_j^2(x_{ij}), \forall f_j \in \mathcal{H}_j, j \in \{1,\ldots,p\}.$$

With the help of the mode-based metric (3) and the hypothesis space (5), we formulated the mode-based sparse additive model as:

$$\hat{f} = \arg\max_{f\in\mathcal{F}}\{\hat{\mathcal{R}}^{\sigma}(f) - \lambda_1 \sum_{j=1}^{p}\|f_j\|_n - \lambda_2 \sum_{j=1}^{p}\|f_j\|_{K_j}\}, \tag{6}$$

where $(\lambda_1, \lambda_2)$ is a pair of positive regularization. The first regularization term is sparsity-promoting [11,36], and the second one guarantees smoothness in the solution.

By the representer theorem of kernel methods (e.g., [37]), the solution of (6) admits the following form:

$$\hat{f}(u) = \sum_{i=1}^{n}\sum_{j=1}^{p} \hat{\alpha}_{ij} K(u_j, x_{ij}), u = (u_1,\ldots,u_p)^T$$

with a collection of coefficients $\{\hat{\alpha}_j = (\alpha_{1j},\ldots,\alpha_{nj})^T \in \mathbb{R}^n : j = 1,\ldots,p\}$.

The optimal coefficients with respect to (6) are the solution to the following non-convex optimization:

$$\max_{\alpha_j\in\mathbb{R}^n, \alpha_j^T K_j \alpha_j \le 1}\{\frac{1}{n}\sum_{i=1}^{n}\phi(\frac{y_i - \sum_{j=1}^{p} K_{ji}^T\alpha_j}{\sigma}) - \frac{\lambda_1}{\sqrt{n}}\sum_{j=1}^{p}\|K_j\alpha_j\|_2 - \lambda_2\sum_{j=1}^{p}\sqrt{\alpha_j^T K_j \alpha_j}\}$$

where $K_{ji} = (K_j(x_{1j}, x_{ij}),\ldots,K_j(x_{nj}, x_{ij}))^T \in \mathbb{R}^n$ and $K_j = (K_j(x_{ij}, x_{lj}))_{i,l}^n = (K_{j1},\ldots,K_{jn}) \in \mathbb{R}^{n\times n}$.

**Remark 2.** *There are various combinations of sparsity and smoothness regularization for additive models [2,3,29–31]. The regularization in this paper adopting a two-fold group Lasso scheme, which was employed in [28], but in quantile regression settings, is also different from the coefficient-based regularized modal regression in [19].*

**Remark 3.** *From the lens of computation, the proposed algorithm (6) can be transformed into a regularized least-squares regression problem by HQ optimization [32]. Then, the transformed problem can be tackled with the SOCP [33] easily.*

## 3. Error Analysis

This section states the upper bounds of the excess quantity $\mathcal{R}(f^*) - \mathcal{R}(\hat{f})$. For the ease of presentation, we only considered the special setting where $\mathcal{H}_j \equiv \mathcal{H}_{j'}, \forall j, j' \in \{1,\ldots,p\}$, and we denote $\oplus_{j=1}^{p}\mathcal{H}_j$ as $\mathcal{H}_K$ with $\sup K(x,x) \le 1$.

Recall that the Mercer kernel $K : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ admits the following spectral expansion [38]:

$$K(x,x') = \sum_{\ell\ge 1} b_\ell \psi_\ell(x)\psi_\ell(x'), x, x' \in \mathcal{X},$$

where $\{(b_\ell, \psi_\ell)\}_{\ell \geq 1}$ are the pairs of eigenvalue-eigenfunctions of integral operator $\mathcal{T}f$ : $\int K(\cdot, x)f(x)d\rho_{\mathcal{X}}(x)$ with $b_1 \geq b_2 \geq \ldots \geq 0$.

To evaluate the complexity of $\mathcal{H}_K$ in terms of the decay rate of eigenvalues $\{b_\ell\}_{\ell \geq 1}$ [27,28], we refer to Assumption 1 in [28] as the basis of our analysis.

**Assumption 1.** *There exist $s \in (0,1)$ and constant $c_1 > 0$ such that $b_\ell \leq c_1 \ell^{-\frac{1}{s}}, \forall \ell \geq 1$.*

As illustrated in [27,28], the requirement $s < 1$ is a weak condition since $\sum_\ell b_\ell = EK(x,x) \leq 1$. In particular, it holds $b_\ell \asymp \ell^{-2h}$ for the Sobolev space $\mathcal{H}_K = W_2^h(h > \frac{1}{2})$ with the Lebesgue measure on $[0,1]$.

To describe the hypothesis in RKHS, we refer to Assumption 2 in [28].

**Assumption 2.** *For some $s \in (0,1)$ given in Assumption 1, there exists a positive constant $c_2$ such that $\|f\|_\infty \leq c_2 \|f\|_2^{1-s} \|f\|_K^s, \forall f \in \mathcal{H}_K$.*

**Remark 4.** *To understand the statistical performance of the proposed estimator without any "correlatedness" conditions on covariates, Rademacher complexity [39] was used to measure functional complexity in [28]. We drew on the experience of [28].*

In general, Assumption 2 is stronger than Assumption 1 and is satisfied when the RKHS is continuously embeddable in a Sobolev space. For the uniformly bounded $\{\psi_\ell\}_{\ell \geq 1}$, this sub-norm condition is consistent with Assumption 1.

For any given independent input variables $\{x_i\}_{i=1}^n \subset \mathcal{X}$, define the Rademacher complexity:

$$\mathcal{R}_n(f) := \frac{1}{n}\sum_{i=1}^{n} \sigma_i f(x_i), \forall f \in \mathcal{H}_K,$$

where $\{\sigma_i\}_{i=1}^n$ is an *i.i.d.* sequence of Rademacher variables that take $\{\pm 1\}$ with probability $1/2$. As shown in [40], it holds:

$$E\mathcal{R}_n\{f \in \mathcal{H}_K\{\|f\|_K = 1, \|f\|_2 \leq t\}\} \asymp \frac{1}{\sqrt{n}}[\sum_{\ell}^{\infty} \min\{t^2, b_\ell\}]^{\frac{1}{2}}.$$

Moreover, from Assumption 1, define:

$$\begin{aligned} \gamma_n &:= \inf\{\gamma \geq \sqrt{\frac{A\log\tilde{p}}{n}}, E[\sup_{\|f\|_K=1, \|f\|_2 \leq t} |\mathcal{R}_n(f)|] \leq \gamma t + \gamma^2, \forall t \in (0,1)\} \\ &\asymp \max\{\sqrt{\frac{A\log\tilde{p}}{n}}, (\frac{1}{n})^{\frac{1}{2(1+\alpha)}}\}. \end{aligned}$$

The main idea of our error analysis is to first state a theory result for a defined event and then investigate the behavior of $\hat{f}$ in (6) conditional on that event.

Define $\eta(t) := \max\{1, \sqrt{t}, t/\sqrt{n}\}$ for any $t > 0$ and $\xi_n := \xi_n(\lambda) = \max\{\lambda^{-\frac{\alpha}{2}} n^{-\frac{1}{2}}, \lambda^{-\frac{1}{2}} n^{-\frac{1}{1+\alpha}}, \sqrt{\frac{\log p}{n}}\}$, and consider the event:

$$\theta(t) = \{|\frac{1}{n}\sum_{i=1}^{n} \epsilon_i f(x_i)| \leq c_\alpha \eta(t) \xi_n (\|f\|_2 + \lambda^{\frac{1}{2}} \|f\|_K), \forall f \in \mathcal{H}_K\},$$

where $\{\epsilon_i\}_{i=1}^n$ are zero-mean *i.i.d.* random variables with $|\epsilon_i| \leq L$ and $c_\alpha$ is a constant depending on $\alpha$ and $L$.

**Remark 5.** *To analyze the behavior of the regularized estimator conditioned on the event, several basic facts of the empirical processes were introduced in [28]. Our work can be boiled down to this framework. We introduced the relevant lemmas in [28] as a stepping stone.*

**Lemma 1.** *Let Assumptions 1 and 2 be true. If* $\frac{\log p}{\sqrt{n}} \le 1$, *it holds:*

$$P(\theta(t)) \ge 1 - \exp\{-t\}, \forall \lambda > 0, t \ge 1.$$

The following lemma (see also Theorem 4 in [41]) demonstrates the relationship between the empirical norm $\|\cdot\|_n$ and $\|\cdot\|_2$ for functions in $\mathcal{H}_K$.

**Lemma 2.** *For* $A \ge 1$ *and any given* $\tilde{p} \ge p$ *with* $\log \tilde{p} \ge 2 \log\log n$, *there exists a constant* $c$ *such that:*

$$\|f\|_2 \le c(\|f\|_n + \gamma_n \|f\|_K)$$

*and:*

$$\|f\|_n \le c(\|f\|_2 + \gamma_n \|f\|_K)$$

*with confidence at least* $1 - \tilde{p}^{-A}$, *where* $\gamma_n \asymp \max(\sqrt{\frac{A \log \tilde{p}}{n}}, (\frac{1}{n})^{\frac{1}{2(1+\alpha)}})$.

**Lemma 3.** *Let* $\{z_i\}_{i=1}^n \subset \mathcal{Z}$ *be independent random variables, and let* $\Gamma$ *be a class of real-valued functions on* $\mathcal{Z}$ *satisfying:*

$$\|\gamma\| \le \eta_n, \forall \gamma \in \Gamma, \quad \textit{and} \quad \frac{1}{n}\sum_{i=1}^n var(\gamma(z_i)) \le \iota_n^2,$$

*for some positive constants* $\eta_n$ *and* $\iota_n$. *Define* $\zeta := \sup_{\gamma \in \Gamma} |\frac{1}{n}\sum_{i=1}^n \gamma(z_i) - E\gamma(z)|$. *Then,*

$$P\{\zeta \ge E\zeta + t\sqrt{2(\iota_n^2 + 2\eta_n E z)} + \frac{2\eta_n t^2}{3} \le \exp\{-nt^2\}$$

For any given $\Delta_-$ and $\Delta_+$, define:

$$\mathcal{F}(\Delta_-, \Delta_+) \quad = \quad \{f = \sum_{j=1}^p f_j \in \mathcal{H}_K : \gamma_n \sum_{j=1}^p \|f_j - f_j^*\|_2 \le \Delta_-, \gamma_n^2 \sum_{j=1}^p \|f_j - f_j^*\|_K \le \Delta_+\},$$

**Lemma 4.** *Let Assumptions 1 and 2 be true for each* $\mathcal{H}_j$. *For any given* $A \ge 2$, *with confidence at least* $1 - \tilde{p}^{-A}$, *it holds:*

$$\mathcal{R}^\sigma(f^*) - \mathcal{R}^\sigma(f) - (\hat{\mathcal{R}}^\sigma(f^*) - \hat{\mathcal{R}}^\sigma(f)) \le c_* \eta(t_0)(\Delta_- + \Delta_+) + \exp\{-\tilde{p}\},$$

*for any* $f \in \mathcal{F}(\Delta_-, \Delta_+)$ *with* $max\{\Delta_-, \Delta_+\} \le e^{\tilde{p}}$, *where* $t_0 = 2\log(\frac{2\sqrt{3}}{\log 2}) + A \log \tilde{p} + 2 \log \tilde{p}$, $\lambda = n^{-\frac{1}{1+\alpha}}$, *and* $c_*$ *is a positive constant.*

**Proof.** Denote $\Gamma = \{\gamma(z) : \gamma(z) = \frac{1}{\sigma}\phi(\frac{y - f^*(x)}{\sigma}) - \frac{1}{\sigma}\phi(\frac{y - f(x)}{\sigma}), f \in \mathcal{F}(\Delta_-, \Delta_+)\}$. It is easy to verify that:

$$E\gamma(z) - \frac{1}{n}\sum_{i=1}^n \gamma(z_i) = \mathcal{R}(f^*) - \mathcal{R}(f) - (\hat{\mathcal{R}}(f^*) - \hat{\mathcal{R}}(f)), \gamma \in \Gamma.$$

Let $\zeta := \sup_{\gamma \in \Gamma} |\frac{1}{n}\sum_{i=1}^n \gamma(z_i) - E\gamma(z)|$. From Lemma 3, we have:

$$\zeta \le E\zeta + \sqrt{\frac{2t(\iota_n^2 + 2\eta_n E\zeta)}{n}} + \frac{2\eta_n t}{3n}, \tag{7}$$

with probability at least $1 - \exp\{-t\}$, where constants $\sup_{\gamma \in \Gamma} \|\gamma\|_\infty = \eta_n$ and $\sup_{\gamma \in \Gamma} \sqrt{\frac{1}{n}\sum_{i=1}^n var(\gamma(z_i))} = \iota_n$. Observing that:

$$\sqrt{\frac{2t(\iota_n^2 + 2\eta_n E\zeta)}{n}} \le \sqrt{\frac{2t\iota_n^2}{n}} + 2\sqrt{\frac{\eta_n E\zeta}{n}} \le \sqrt{\frac{2t}{n}}\iota_n + E\zeta + \frac{\eta_n}{n}, \tag{8}$$

we can take:

$$\iota_n^2 \le 2E(\gamma(z))^2 = 2E(\frac{1}{\sigma}\phi(\frac{y-f^*(x)}{\sigma}) - \frac{1}{\sigma}\phi(\frac{y-f(x)}{\sigma}))^2 \le \frac{2\|\phi'\|_\infty^2}{\sigma^4}\|f-f^*\|_2^2 \le \frac{2\|\phi'\|_\infty^2}{\sigma^4}\frac{\Delta_-^2}{\gamma^2}, \tag{9}$$

and:

$$\eta_n = \sup_{\gamma\in\Gamma}\|\gamma\|_\infty \le \frac{\|\phi'\|_\infty}{\sigma^2}\|f^*-f\|_\infty \le \frac{\|\phi'\|_\infty}{\sigma^2}\|f^*-f\|_K \le \frac{\|\phi'\|_\infty}{\sigma^2}\frac{\Delta_+}{\gamma_n^2}. \tag{10}$$

Combining (7)–(10), we obtain with confidence at least $1-\exp\{-t\}$

$$\zeta \le 2E\zeta + \frac{2\|\phi'\|_\infty}{\gamma_n\sigma^2}\sqrt{\frac{t}{n}} + \frac{\kappa\|\phi'\|_\infty\Delta_+}{\sigma^2\gamma_n^2}\frac{1+t}{n}.$$

By a symmetrization technique in [42], we have:

$$E\zeta \le 2E\mathcal{R}_n(\Gamma) \le \frac{2\|\phi'\|_\infty}{\sigma^2}E\mathcal{R}_n(\mathcal{F}-f^*).$$

Applying Lemma 3 for $\mathcal{R}_n(\mathcal{F}-f^*)$, we obtain that:

$$E[\mathcal{R}_n(\mathcal{F}-f^*)] \le \mathcal{R}_n(\mathcal{F}-f^*) + 4\frac{\Delta_-}{\gamma_n}\sqrt{\frac{2t}{n}} + \frac{\Delta_+}{\gamma_n^2}\frac{1+t}{n},$$

with probability at least $1-2\exp\{-t\}$. Moreover, with probability at least $1-2\exp\{-t\}$, it holds:

$$\begin{aligned}\zeta &\le \frac{8\|\phi'\|_\infty}{\sigma^2}\mathcal{R}_n(\mathcal{F}-f^*) + \frac{6\|\phi'\|_\infty\Delta_-}{\gamma_n\sigma^2}\sqrt{\frac{t}{n}} + \frac{5\|\phi'\|_\infty\Delta_+}{\gamma_n^2\sigma^2}\frac{1+t}{n}\\ &\le \frac{8\|\phi'\|_\infty}{\sigma^2}\sum_{j=1}^{p}\mathcal{R}_n(\mathcal{H}_j-f_j^*) + \frac{6\|\phi'\|_\infty\Delta_-}{\gamma_n\sigma^2}\sqrt{\frac{t}{n}} + \frac{5\|\phi'\|_\infty\Delta_+}{\gamma_n^2\sigma^2}\frac{1+t}{n}.\end{aligned}$$

For the event $\theta(t)$, Lemma 1 demonstrates that:

$$|\mathcal{R}_n(f)| \le c_\alpha\eta(t)\xi_n(\|f\|_2 + \lambda^{\frac{1}{2}}\|f\|_K), \forall f\in\mathcal{H}_K, \forall\lambda>0,$$

with confidence $1-\exp\{-t\}$. Then,

$$\zeta \le \frac{8\|\phi'\|_\infty c_\alpha\eta(t)\xi_n}{\sigma^2}\sup_{f\in\mathcal{F}}\{\sum_{j=1}^{p}\|f-f_j^*\|_2 + \lambda^{\frac{1}{2}}\sum_{j=1}^{p}\|f_j-f_j^*\|_K\} + \frac{6\|\phi'\|_\infty\Delta_-}{\gamma_n\sigma^2}\sqrt{\frac{t}{n}} + \frac{5\|\phi'\|_\infty\Delta_+}{\gamma_n^2\sigma^2}\frac{1+t}{n}.$$

Taking $\lambda = n^{-\frac{1}{1+\alpha}}$, we can verify that $c\gamma_n \ge \xi_n$ and $\xi_n\lambda^{\frac{1}{2}} \ge c\gamma_n^2$. Then,

$$\zeta \le \frac{8c_\alpha\eta(t)\|\phi'\|_\infty}{\sigma^2}(\Delta_+ + \Delta_-) + \frac{6\Delta_-\|\phi'\|_\infty}{\sigma^2}\sqrt{\frac{t}{A\log\tilde{p}}} + \frac{5\Delta_+\|\phi'\|_\infty t}{\sigma^2 A\log\tilde{p}},$$

for some event $\theta(\Delta_-,\Delta_+)$.

For $t = 2\log(2\sqrt{3}/\log 2) + A\log\tilde{p} + 2\log\tilde{p}$, we deduce that $e^{-\tilde{p}} \le \Delta_- \le e^{\tilde{p}}$ and $e^{-\tilde{p}} \le \Delta_+ \le e^{\tilde{p}}$ considering $(2\tilde{p}+1)^2$ different discrete pairs $\Delta_-^j = \Delta_+^j := 2^{-j}$, $j = -\tilde{p}, \ldots, \tilde{p}$, and we deduce that:

$$P(\bigcap_{k,j}\theta(\Delta_-^j, \Delta_+^j)) \le 1 - 3(\frac{2}{\log 2}^2 \tilde{p}^2 \exp\{-2\log(\frac{2\sqrt{3}}{\log 2} - A\log\tilde{p} - 2\log\tilde{p}\} \le 1 - \tilde{p}^{-A}.$$

When $\Delta_- \le e^{-\tilde{p}}$ or $\Delta_+ \le e^{-\tilde{p}}$, it is trivial to obtain the desired result. □

The proof of Lemma 4 is derived from the proof of Proposition 1 in [28] for the quantile regression. We state our main result on the error bound.

**Theorem 1.** *Let the regularization parameters of $\hat{f}$ defined in (6) be $\lambda_1 = \sqrt{\xi}\gamma_n$ and $\lambda_2 = \xi\gamma_n^2$, where $\xi = \max\{2c\eta(t_0), 4\}$ with $\eta(t) = \max\{1, \sqrt{t}, t/\sqrt{n}\}$, $t_0 = 2\log(2\sqrt{3}/\log 2) + A\log\tilde{p} + 2\log\tilde{p}$, and $\gamma_n \asymp \max(\sqrt{\frac{A\log\tilde{p}}{n}}, (\frac{1}{n})^{\frac{1}{2(1+\alpha)}})$. Under the conditions of Assumptions 1 and 2, for any $\tilde{p} \ge p$ such that $\log p \le \sqrt{n}$ and $\log\tilde{p} \ge 2\log\log n$, then for some constant $A \ge 2$, such that with probability at least $1 - 2\tilde{d}^{-A}$:*

$$\begin{aligned}
\mathcal{R}(f^*) - \mathcal{R}(\hat{f}) &\le cs\|\phi'\|_\infty \eta(t_0)(\eta(t_0))^{\frac{1}{4}}\sqrt{\gamma_n} \le c(\eta(t_0))^{\frac{5}{4}} \max\{(\frac{A\log\tilde{p}}{c})^{\frac{1}{4}}, (\frac{1}{n})^{\frac{1}{4(1+\alpha)}}\} \\
&\le c\max\{\sqrt{A\log\tilde{p}}, \frac{A\log\tilde{p}}{\sqrt{n}}\}^{\frac{5}{4}} \cdot \max\{(\frac{A\log\tilde{p}}{n})^{\frac{1}{4}}, (\frac{1}{n})^{\frac{1}{4+4\alpha}}\} \\
&\le c\max\{\frac{(A\log\tilde{p})^{\frac{7}{8}}}{n^{\frac{1}{4}}}, \frac{(A\log\tilde{p})^{\frac{1}{2}}}{n^{\frac{1}{4+4\alpha}}}, \frac{(A\log\tilde{p})^{\frac{3}{2}}}{n^{\frac{3}{4}}}, \frac{(A\log\tilde{p})^{\frac{5}{4}}}{n^{\frac{3+2\alpha}{4+4\alpha}}}\}.
\end{aligned}$$

**Proof.** By the definition of $\hat{f}$ in (6), we know that:

$$\hat{\mathcal{R}}^\sigma(\hat{f}) - \lambda_1 \textstyle\sum_{j=1}^p \|\hat{f}_j\|_n - \lambda_2 \sum_{j=1}^p \|\hat{f}_j\|_{K_j} \ge \hat{\mathcal{R}}^\sigma(f^*) - \lambda_1 \sum_{j=1}^p \|f_j^*\|_n - \lambda_2 \sum_{j=1}^p \|f_j^*\|_{K_j}.$$

This implies that:

$$\begin{aligned}
&\hat{\mathcal{R}}^\sigma(\hat{f}) - \mathcal{R}^\sigma(f^*) - \lambda_1 \sum_{j=1}^p \|\hat{f}_j\|_n - \lambda_2 \sum_{j=1}^p \|\hat{f}_j\|_{K_j} \\
\ge\ & [\mathcal{R}^\sigma(\hat{f}) - \mathcal{R}^\sigma(f^*)] - [\hat{\mathcal{R}}^\sigma(\hat{f}) - \hat{\mathcal{R}}^\sigma(f^*)] - \lambda_1 \sum_{j=1}^p \|f_j^*\|_n - \lambda_2 \sum_{j=1}^p \|f_j^*\|_{K_j}.
\end{aligned}$$

Moreover,

$$\begin{aligned}
&\mathcal{R}^\sigma(f^*) - \mathcal{R}^\sigma(\hat{f}) \le \mathcal{R}^\sigma(f^*) - \mathcal{R}^\sigma(\hat{f}) + \lambda_1 \sum_{j\notin\mathcal{S}} \|\hat{f}_j\|_n + \lambda_2 \sum_{j\notin\mathcal{S}} \|\hat{f}_j\|_K \\
\le\ & [\mathcal{R}^\sigma(f^*) - \mathcal{R}^\sigma(\hat{f})] - [\hat{\mathcal{R}}^\sigma(f^*) - \hat{\mathcal{R}}^\sigma(\hat{f})] + \lambda_1 \sum_{j\in\mathcal{S}} (\|f_j^*\|_n - \|\hat{f}_j\|_n) + \lambda_2 \sum_{j\in\mathcal{S}} (\|f_j^*\|_K - \|\hat{f}_j\|_K) \\
\le\ & [\mathcal{R}^\sigma(f^*) - \mathcal{R}^\sigma(\hat{f})] - [\hat{\mathcal{R}}^\sigma(f^*) - \hat{\mathcal{R}}^\sigma(\hat{f})] + \lambda_1 \sum_{j\in\mathcal{S}} \|\hat{f}_j - f_j^*\|_n + \lambda_2 \sum_{j\in\mathcal{S}} \|\hat{f}_j - f_j^*\|_K. \qquad (11)
\end{aligned}$$

Taking $\lambda_1 = \sqrt{\xi}\gamma_n$, $\lambda_2 = \xi\gamma_n^2$ with $\gamma_n = \max\{\sqrt{\frac{A\log\tilde{p}}{n}}, (\frac{1}{n})^{\frac{1}{2+2\alpha}}\}$, $\alpha \in (0,1)$, we deduce that:

$$\gamma_n \sum_{j=1}^p \|\hat{f}_j - f_j^*\|_2 \le 2p(\frac{1}{n})^{\frac{1}{2+2\alpha}} \le 2\tilde{p}(\frac{1}{4}) \le e^{\tilde{p}}, \forall n \ge 1, \tilde{p} \ge p,$$

and:

$$\gamma_n^2 \sum_{j=1}^p \|f_j - f_j^*\|_{K_j} \le \gamma_n\gamma_n \sum_{j=1}^p \|\hat{f} - f^*\|_{K_j} \le e^{-\tilde{p}}.$$

Therefore, we verify that $\hat{f} \in \mathcal{F}(\Delta_-, \Delta_+)$ with $\Delta_- \leq e^{\tilde{p}}$ and $\Delta_+ \leq e^{\tilde{p}}$. With the choices $\lambda_2 = \lambda_1^2 = \xi\gamma_n^2$, it holds:

$$\lambda_1\|\hat{f}_j - f_j^*\|_n + \lambda_2\|\hat{f}_j - f_j^*\|_K \leq 2(\lambda_1 + \lambda_2) = 4\sqrt{\xi}\gamma_n, j \in \mathcal{S}.$$

due to the fact $\|f_j\|_n \leq \|f_j\|_K \leq 1$, for any $f_j \in \mathcal{H}_{K_j}$.

According to Lemma 4 and (11), we obtain:

$$\begin{aligned} & \mathcal{R}^\sigma(f^*) - \mathcal{R}^\sigma(\hat{f}) \\ \leq\; & \frac{c\eta t_0\|\phi'\|_\infty}{\sigma^2}(\gamma_n \sum_{j=1}^{p}\|\hat{f}_j - f_j^*\|_2 + \gamma_n^2 \sum_{j=1}^{p}\|\hat{f}_j - f_j^*\|_K) + \lambda_1 \sum_{j\in\mathcal{S}}\|\hat{f}_j - f_j^*\|_n + \lambda_2 \sum_{j\in\mathcal{S}}\|\hat{f}_j - f_j^*\|_K + e^{-\tilde{p}} \\ \leq\; & \frac{c\eta(t_0)\|\phi'\|_\infty}{\sigma^2}\sqrt{\xi}\gamma_n + e^{-\tilde{p}}, \end{aligned}$$

with probability at least $1 - 2\tilde{p}^{-A}$.

Notice that $\log\tilde{p} \geq 2\log\log n$ implies that $e^{-\tilde{p}} \leq n^{-2} \leq \gamma_n$. Then:

$$\mathcal{R}^\sigma(f^*) - \mathcal{R}(\hat{f}) \leq \frac{c\eta(t_0)\|\phi'\|_\infty}{\sigma^2}\sqrt{\xi}\gamma_n.$$

Combining this with Theorem 9 in [17] and setting $\sigma = (\|\phi'\|_\infty\eta(t_0)\sqrt{\xi}\gamma_n)^{\frac{1}{4}}$, we obtain the desired result. □

The proof of Theorem 1 is inspired by that of Theorem 1 in [28]; see [28] for more details. According to Theorem 1, we can conclude that the mode-based SpAM can achieve the learning rate with polynomial decay $\mathcal{O}(n^{-\frac{1}{4+4\alpha}})$ since $\epsilon \in [0,1]$ and $A, \tilde{p}$ are positive constants.

## 4. Experimental Evaluation

To demonstrate the efficiency of our method, in this section, we evaluated our model on some synthetic datasets. The data in $\mathbb{R}^p$ with dimension $p = 5$ and $p = 10$ were generated randomly according to the uniform distribution on the interval $[0,1]$. Then, we computed the MSE of our estimator $\hat{f}$. Figures 1–3 depict the MSE of $\hat{f}$ when the parameter pair $(\lambda_1, \lambda_2) = (0,1), (1,0)$ and $(1,1)$, respectively, while the number of samples $n$ varies from 50/60 to 80/90. This paper used Yalmip [43] modeling in the MATLAB environment and called *fmincon* to solve the problem. From the figures, we can notice that the MSEs tended to decrease with the increase of the number of samples $n$ under three kinds of parameter settings, which verified that our method was effective in the regression of high-dimensional data.

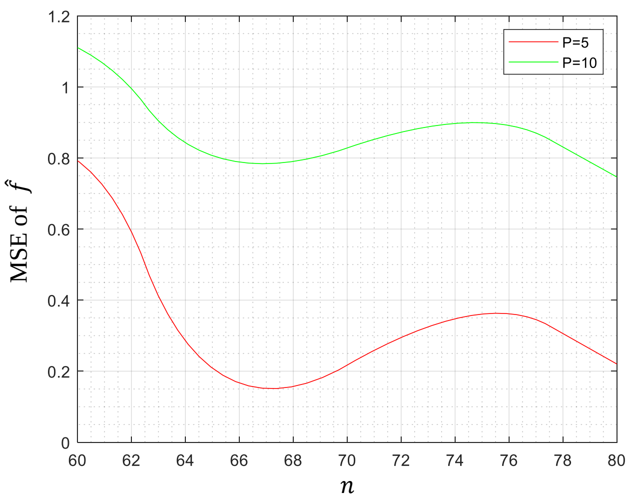


**Figure 1.** MSE of $\hat{f}$ when $(\lambda_1, \lambda_2) = (0,1)$.

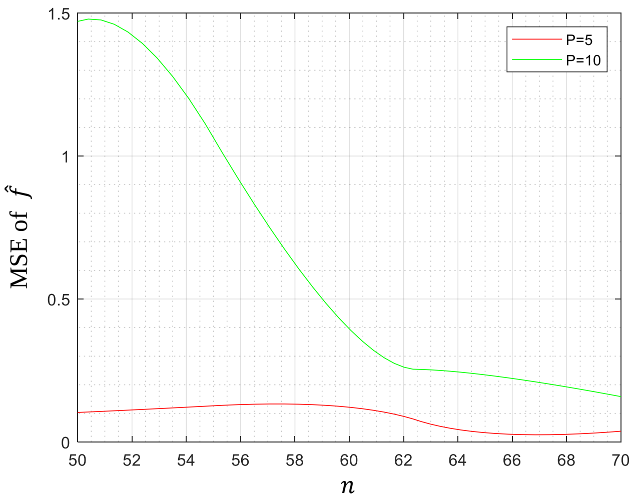


**Figure 2.** MSE of $\hat{f}$ when $(\lambda_1, \lambda_2) = (1, 0)$.

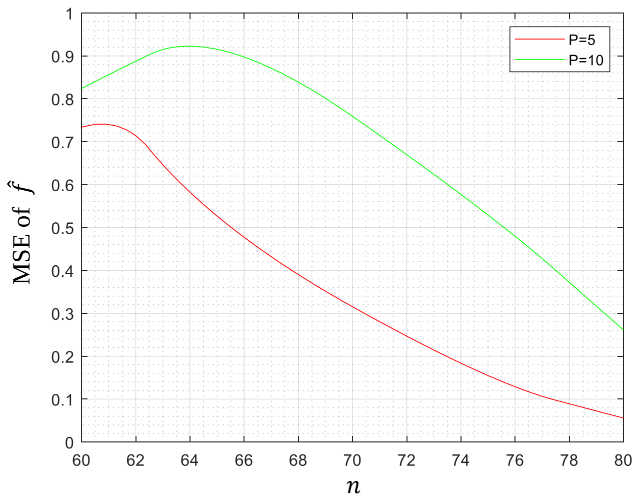


**Figure 3.** MSE of $\hat{f}$ when $(\lambda_1, \lambda_2) = (1, 1)$.

## 5. Conclusions

In this work, we proposed a mode-based sparse additive model and established its generalization error bound. The theoretical results extended the previous mean-based analysis to the mode-based approach. We demonstrated that the mode-based SpAM can achieve the learning rate with polynomial decay $\mathcal{O}(n^{-\frac{1}{4+4\alpha}})$, which is comparable to the previous result in [15] with $\mathcal{O}(n^{-\frac{1}{7}})$. In the future, it will be important to further explore the variable selection consistency of the proposed model.

**Author Contributions:** Conceptualization, H.D., B.S., J.C. and Z.P.; methodology, H.D. and Z.P.; validation, B.S. and Z.P.; formal analysis, H.D., B.S. and Z.P.; investigation, H.D. and J.C.; resources, Z.P.; data curation, H.D. and J.C.; writing—original draft preparation, H.D. and J.C.; writing—review and editing, H.D. and J.C.; visualization, H.D. and J.C.; supervision, B.S. and Z.P.; project administration, B.S. and Z.P.; funding acquisition, H.D., B.S. and Z.P. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported in part by the Fundamental Research Funds for the Central Universities of China (Grant Nos. 2662019FW003 and 2662020LXQD001) and the National Natural Science Foundation of China (Grant No. 12001217).

**Data Availability Statement:** The synthetic data generation method of the simulation experiment has been introduced in the experimental part.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References


1. Xia, Y.; Hou, Y.; Lv, S. Learning Rates for Partially Linear Support Vector Machine in High Dimensions. *Anal. Appl.* **2021**, *19*, 167–182. [CrossRef]
2. Ravikumar, P.; Liu, H.; Lafferty, J.; Wasserman, L. SpAM: Sparse Additive Models. *J. R. Stat. Soc. Ser. B* **2009**, *71*, 1009–1030. [CrossRef]
3. Yin, J.; Chen, X.; Xing, E.P. Group Sparse Additive Models. In Proceedings of the International Conference on Machine Learning (ICML), Edinburgh, UK, 26 June–1 July 2012; pp. 1643–1650.
4. Lin, Y.; Zhang, H.H. Component Selection and Smoothing in Multivariate Nonparametric Regression. *Ann. Stat.* **2006**, *34*, 2272–2297. [CrossRef]
5. Zhao, T.; Liu, H. Sparse additive machine. In Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS), La Palma, Spain, 21–23 April 2012; Volume 22, pp. 1435–1443.
6. Chen, H.; Wang, X.; Deng, C.; Huang, H. Group Sparse Additive Machine. In Proceedings of the Advances in Neural Information Processing Systems (NIPS), Long Beach, CA, USA, 4–9 December 2017; Volume 30, pp. 197–207.
7. Kandasamy, K.; Yu, Y. Additive Approximations in High Dimensional Nonparametric Regression via the SALSA. In Proceedings of the International Conference on Machine Learning (ICML), New York, NY, USA, 19–24 June 2016; Volume 48, pp. 69–78.
8. Wang, Y.; Chen, H.; Zheng, F.; Xu, C.; Gong, T.; Chen, Y. Multi-task Additive Models for Robust Estimation and Automatic Structure Discovery. In Proceedings of the Advances in Neural Information Processing Systems (NIPS), Online, 6–12 December 2020; Volume 33, pp. 11744–11755.
9. Chen, H.; Liu, G.; Huang, H. Sparse Shrunk Additive Models. In Proceedings of the International Conference on Machine Learning (ICML), Vienna, Austria, 12–18 July 2020; Volume 119, pp. 6194–6204.
10. Chen, H.; Guo, C.; Xiong, H.; Wang, Y. Sparse Additive Machine with Ramp Loss. *Anal. Appl.* **2021**, *19*, 509–528. [CrossRef]
11. Meier, L.; Geer, S.V.D.; Buhlmann, P. High-dimensional Additive Modeling. *Ann. Stat.* **2009**, *37*, 3779–3821. [CrossRef]
12. Raskutti, G.; Wainwright, M.J.; Yu, B. Minimax-optimal Rates for Sparse Additive Models over Kernel Classes via Convex Programming. *J. Mach. Learn. Res.* **2012**, *13*, 389–427.
13. Kemp, G.C.R.; Silva, J.M.C.S. Regression towards the mode. *J. Econom.* **2012**, *170*, 92–101. [CrossRef]
14. Yao, W.; Li, L. A New Regression model: Modal Linear Regression. *Scand. J. Stat.* **2014**, *41*, 656–671. [CrossRef]
15. Wang, X.; Chen, H.; Cai, W.; Shen, D.; Huang, H. Regularized Modal Regression with Applications in Cognitive Impairment Prediction. In Proceedings of the Advances in Neural Information Processing Systems (NIPS), Long Beach, CA, USA, 4–9 December 2017; Volume 30, pp. 1448–1458.
16. Chen, Y.C.; Genovese, C.R.; Tibshirani, R.J.; Wasserman, L. Nonparametric Modal Regression. *Ann. Stat.* **2014**, *44*, 489–514. [CrossRef]
17. Feng, Y.; Fan, J.; Suykens, J. A Statistical Learning Approach to Modal Regression. *J. Mach. Learn. Res.* **2020**, *21*, 1–35.
18. Collomb, G.; Härdle, W.; Hassani, S. A Note on Prediction via Estimation of the Conditional Mode Function. *J. Stat. Plan. Inference* **1986**, *15*, 227–236. [CrossRef]
19. Chen, H.; Wang, Y.; Zheng, F.; Deng, C.; Huang, H. Sparse Modal Additive Model. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, 1–15. [CrossRef] [PubMed]
20. Li, J.; Ray, S.; Lindsay, B.G. A Nonparametric Statistical Approach to Clustering via Mode Identification. *J. Mach. Learn. Res.* **2007**, *8*, 1687–1723.
21. Einbeck, J.; Tutz, G. Modeling beyond Regression Function: An Application of Multimodal Regression to Speed-flow Data. *J. R. Stat. Soc. Ser. C Appl. Stat.* **2006**, *55*, 461–475. [CrossRef]
22. Tibshirani, R. Regression Shrinkage and Selection Via the Lasso. *J. R. Stat. Soc. Ser. B (Methodol.)* **1996**, *58*, 267–288. [CrossRef]
23. Feng, Y.; Huang, X.; Shi, L.; Yang, Y.; Suykens, J.A. Learning with the Maximum Correntropy Criterion Induced Losses for Regression. *J. Mach. Learn. Res.* **2015**, *16*, 993–1034.
24. Lv, F.; Fan, J. Optimal learning with Gaussians and Correntropy Loss. *Anal. Appl.* **2019**, *19*, 107–124. [CrossRef]
25. Yao, W.; Lindsay, B.G.; Li, R. Local Modal Regression. *J. Nonparametr. Stat.* **2012**, *24*, 647–663. [CrossRef]
26. Chen, Y. Modal Regression using Kernel Density Estimation: A Review. *Wiley Interdiscip. Rev. Comput. Stat.* **2018**, *10*, e1431. [CrossRef]
27. Steinwart, I.; Christmann, A. *Support Vector Machines*; Springer Science and Business Media: Berlin/Heidelberg, Germany, 2008.
28. Lv, S.; Lin, H.; Lian, H.; Huang, J. Oracle Inequalities for Sparse Additive Quantile Regression in Reproducing Kernel Hilbert Space. *Ann. Stat.* **2018**, *46*, 781–813. [CrossRef]
29. Huang, J.; Horowitz, J.L.; Wei, F. Variable Selection in Nonparametric Additive Models. *Ann. Stat.* **2010**, *38*, 2282–2313. [CrossRef] [PubMed]
30. Christmann, A.; Zhou, D.X. Learning Rates for the Risk of Kernel based Quantile Regression Estimators in Additive Models. *Anal. Appl.* **2016**, *14*, 449–477. [CrossRef]
31. Yuan, M.; Zhou, D.X. Minimax Optimal Rates of Estimation in High Dimensional Additive Models. *Ann. Stat.* **2016**, *44*, 2564–2593. [CrossRef]

32. Nikolova, M.; Ng, M.K. Analysis of Half-quadratic Minimization Methods for Sgnal and Image Recovery. *SIAM J. Sci. Comput.* **2006**, *27*, 937–966. [CrossRef]
33. Alizadeh, F.; Goldfarb, D. Second-Order Cone Programming. *Math. Program.* **2003**, *95*, 3–51. [CrossRef]
34. Guo, C.; Song, B.; Wang, Y.; Chen, H.; Xiong, H. Robust Variable Selection and Estimation Based on Kernel Modal Regression. *Entropy* **2019**, *21*, 403. [CrossRef]
35. Wang, Y.; Tang, Y.Y.; Li, L.; Chen, H. Modal Regression-based Atomic Representation for Robust Face Recognition and Reconstruction. *IEEE Trans. Cybern.* **2020**, *50*, 4393–4405. [CrossRef]
36. Suzuki, T.; Sugiyama, M. Fast learning rate of multiple kernel learning: Trade-off between sparsity and smoothness. *Ann. Stat.* **2013**, *41*, 1381–1405. [CrossRef]
37. Schlköpf, B.; Smola, A.J. *Learning with Kernels*; The MIT Press: Cambridge, MA, USA, 2002.
38. Aronszajn, N. Theory of Reproducing Kernels. *Trans. Am. Math. Soc.* **1950**, *68*, 337–404. [CrossRef]
39. Bartlett, P.L.; Bousquet, O.; Mendelson, S. Localized Rademacher Complexities. In Proceedings of the Conference on Computational Learning Theory (COLT), Sydney, Australia, 8–10 July 2002; Volume 2373, pp. 44–58.
40. Mendelson, S. Geometric Parameters of Kernel Machines. In Proceedings of the Conference on Computational Learning Theory (COLT), Sydney, Australia, 8–10 July 2002; Volume 2375, pp. 29–43.
41. Koltchinskii, V.; Yuan, M. Sparsity in Multiple Kernel Learning. *Ann. Stat.* **2010**, *38*, 3660–3695. [CrossRef]
42. Van De Geer, S. *Empirical Processes in M-Estimation*; Cambridge University Press: Cambridge, UK, 2000.
43. Löfberg, J. Automatic robust convex programming. *Optim. Methods Softw.* **2012**, *27*, 115–129. [CrossRef]