A Statistical Learning Approach to Modal Regression

Yunlong Feng YUNLONG.FENG@ESAT.KULEUVEN.BE
Department of Electrical Engineering, ESAT-STADIUS, KU Leuven
Kasteelpark Arenberg 10, Leuven
B-3001, Belgium

Jun Fan JUNFAN@STAT.WISC.EDU
Department of Statistics, University of Wisconsin-Madison
1300 University Avenue, Madison
WI 53706, United States

Johan A.K. Suykens JOHAN.SUYKENS@ESAT.KULEUVEN.BE
Department of Electrical Engineering, ESAT-STADIUS, KU Leuven
Kasteelpark Arenberg 10, Leuven
B-3001, Belgium

Abstract

This paper studies the nonparametric modal regression problem systematically from a statistical learning view. Originally motivated by pursuing a theoretical understanding of the maximum correntropy criterion based regression (MCCR), our study reveals that MCCR with a tending-to-zero scale parameter is essentially modal regression. We show that nonparametric modal regression problem can be approached via the classical empirical risk minimization. Some efforts are then made to develop a framework for analyzing and implementing modal regression. For instance, the modal regression function is described, the modal regression risk is defined explicitly and its *Bayes* rule is characterized; for the sake of computational tractability, the surrogate modal regression risk, which is termed as the generalization risk in our study, is introduced. On the theoretical side, the excess modal regression risk, the excess generalization risk, the function estimation error, and the relations among the above three quantities are studied rigorously. It turns out that under mild conditions, function estimation consistency and convergence may be pursued in modal regression as in vanilla regression protocols such as mean regression, median regression, and quantile regression. However, it outperforms these regression models in terms of robustness as shown in our study from a re-descending M-estimation view. This coincides with and in return explains the merits of MCCR on robustness. On the practical side, the implementation issues of modal regression including the computational algorithm and the tuning parameters selection are discussed. Numerical assessments on modal regression are also conducted to verify our findings empirically.

Keywords: Nonparametric modal regression, empirical risk minimization, generalization bounds, robustness, kernel density estimation, statistical learning theory

1. Introduction

In this paper, we are interested in the nonparametric regression problem which aims at inferring the functional relation between input and output. Regression problems play around and are concerned with the conditional distribution, which in practice can never be known in advance. Instead, normally, what one can access is only a set of observations drawn from the joint probability distribution. To state mathematically, let us denote X as the explanatory variable that takes values in a compact metric space $\mathcal{X} \subset \mathbb{R}^d$ and $Y \in \mathcal{Y} = \mathbb{R}$ as the response variable. Typically, we consider the following data-generating model

$$Y = f^\star(X) + \epsilon,$$

where ϵ is the noise variable. In nonparametric regression problems, the purpose is to infer the unknown function $f^\star$ nonparametrically while certain assumptions on the noise variable ϵ may be imposed. As a compromise, regression estimators usually settle for learning a characterization of the conditional distribution by sifting information through observations generated above. Characterizations of the conditional distribution are versatile, where the several usual ones include the conditional mean, the conditional median, the conditional quantile, and the conditional mode. The versatility of the characterizations of the conditional distribution raises the question that which characterization we should pursue in regression problems. To answer this question, tremendous attention has been drawn in the statistics and machine learning communities. As a matter of fact, a significant part of parametric and nonparametric regression theory has been fostered to illuminate this question.

It is generally considered that each of the above-mentioned regression protocols has its own merits in its own regimes. For instance, it has been well understood that regression towards the conditional mean can be most effective if the noise is Gaussian or sub-Gaussian. Regression towards the conditional median or conditional quantile can be more robust in the absence of light-tailed noise or symmetric conditional distributions. In practice, the choice of the most appropriate regression protocol is usually decided by the type of the observations encountered. In the statistics and machine learning literature, these regression protocols have been studied extensively and understood well. In this study, we focus on a regression problem that has not been well studied in the statistical learning literature, namely, modal regression.

1.1 Modal Regression

Modal regression approaches the unknown truth $f^\star$ by regressing towards the *conditional mode function*. For a set of observations, the mode is the value that appears most frequently. While for a continuous random variable, the mode is the value at which its density function attains its peak value. The conditional mode function is denoted pointwisely as the mode of the conditional density of the independent variable conditioned on the dependent variable.

Previously proposed in Sager and Thisted (1982); Collomb et al. (1987) and studied in, e.g., Lee (1989, 1993), it is shown that one of the most appealing features of modal regression lies in its robustness to outliers, heavy-tailed noise, skewed conditional distributions, and skewed noise distributions. Moreover, regression towards the conditional mode in some cases can be a better option when predicting the trends of observations. This is also the case in

some real-world applications, as illustrated in Matzner-Løfber et al. (1998); Einbeck and Tutz (2006); Yu et al. (2014). However, it seems to us that not enough attention has been attracted to the theory and applications of modal regression, especially in the statistical learning literature. As yet another regression protocol, the above-mentioned merits of modal regression suggest that it deserves far more attention than it has received, especially in the big data era today. This motivates our study on modal regression in this paper.

1.2 Historical Notes on Modal Regression

Modal regression is concerned with the mode. Studies on the mode estimation date back to 1960s since the seminal work of Parzen (1962). It opens the door for kernel density estimation by proposing the *Parzen window* method, with the help of which the estimation of the mode can be typically proceeded. Many subsequent studies concerning the theoretical as well as practical estimation of the mode have been emerging since then, among them Chernoff (1964); Robertson and Cryer (1974); Fukunaga and Hostetler (1975); Eddy (1980); Comaniciu and Meer (2002); Dasgupta and Kpotufe (2014).

In a regression setup, the concern of the conditional mode estimate gives birth to modal regression. As far as we are aware, the idea of regression towards the conditional mode was first proposed in Sager and Thisted (1982) in an isotonic regression setup. It was then specifically investigated in Collomb et al. (1987) when dealing with dependent observations. As a theoretical study, the main conclusion drawn there was the uniform convergence of the nonparametric mode estimator to the conditional mode function. Lately, in Lee (1989, 1993), some pioneering studies of modal regression were conducted. The tractability problem of mode regression was first discussed in their studies from, say, a supervised learning and risk minimization view. By considering some specific modal regression kernels, and assuming the existence of a global conditional mode function under a linear model assumption, they established the asymptotic normality of the resulting estimator. More and more attention to the theory and applications of modal regression has been attracted since the work in Yao et al. (2012), Yao and Li (2014), and Kemp and Santos Silva (2012). In Yao and Li (2014), a global mode was assumed to exist and took a linear form. Under proper assumptions on the conditional density of the noise variable, the implementation issues and the asymptotic normality of the estimator, as well as its robustness were explored. Recently, Chen et al. (2016b) presented an interesting study towards modal regression in which the conditional mode was sought by estimating the maximum of a joint density. By assuming a factorisable modal manifold collection, results on asymptotic error bounds as well as techniques for constructing confidence sets and prediction sets were provided.

To further disentangle the literature on modal regression, we can roughly categorize existing studies by tracing the thread of global or local approaches that they follow. For local approaches, the conditional mode is sought via maximizing a conditional density or a joint density which is typically estimated non-parametrically, e.g., by using kernel density estimators. Studies in Collomb et al. (1987); Samanta and Thavaneswaran (1990); Quintela-Del-Rio and Vieu (1997); Ould-Saïd (1997); Herrmann and Ziegler (2004); Ferraty et al. (2005); Gannoun et al. (2010); Yao et al. (2012); Chen et al. (2016b); Sasaki et al. (2016); Zhou and Huang (2016); Yao and Xiang (2016) fall into this category. For global approaches, the conditional mode is usually sought by maximizing the kernel density estimator for the

variable induced by the residual and assuming that the global mode is unique and belongs to a certain hypothesis space. To name a few, studies in Lee (1989, 1993); Lee and Kim (1998); Yao and Li (2014); Kemp and Santos Silva (2012); Baldauf and Santos Silva (2012); Yu and Aristodemou (2012); Lv et al. (2014); Salah and Françoise (2016) follow this line. It should be noticed that most studies based upon global approaches assume the existence (and also the uniqueness) of a global conditional mode function that is of a parametric form. While for the studies based upon local approaches, usually only the uniqueness assumption of the conditional mode function is imposed. Loosely speaking, modal regression estimators of the former case are nonparametric, while (semi-) parametric in the latter case.

Most of the above-mentioned studies are theoretical in nature. It should be noted that some application-oriented studies on modal regression have also been conducted. Among them, Matzner-Løfber et al. (1998) carried out an empirical comparison among three regression schemes, namely, the conditional mean regression, the conditional median regression, and the conditional mode regression, in nonparametric forecasting problems. They empirically observed that for certain datasets, e.g., the Old Faithful eruption prediction dataset, the mode can be a better option in forecasting than the mean and the median; Yu et al. (2014) discussed the mode-based regression problem in the big data context. Based on empirical evaluations on the Health Survey for English dataset, they argued that the mode could be an effective alternative for pattern-finding; Einbeck and Tutz (2006) dealt with the speed-flow data in traffic engineering by applying a multi-modal regression model.

1.3 Objectives of This Study

As mentioned above, in the statistics literature, there exist some interesting studies towards modal regression from both theoretical and practical viewpoints. However, we notice that several problems related to the theoretical understanding as well as the practical implementations of modal regression remain unclear. For example:

- Modal regression regresses towards the conditional mode function, a direct estimation of which involves the estimation of a conditional or joint density. In fact, many of the existing studies on modal regression follow this approach. Notice that the explanatory variable may be high-dimensional vector-valued, which may make the estimation of the conditional or joint density infeasible. This poses an important problem: how to carry out modal regression without involving the estimation of a density function in a (possibly) high dimensional space? According to the existing studies on modal regression, assuming the existence of a global conditional mode function and imposing some prior structure assumptions on it seems to be promising in avoiding estimating such a density. However, most existing studies of this type assume that the conditional mode function possesses a certain linear or parametric form. This could be restrictive in certain circumstances;

- With a modal regression estimator at hand, how can we evaluate its statistical performance? That is, how can we measure the approximation ability of the modal regression estimator to the conditional mode function? This concern is of great importance in nonparametric statistics as well as in machine learning as it is closely related to the prediction ability of the estimator on future observations. On the other

hand, concerning the implementation issues of modal regression, how can we perform model selection in modal regression?

- It is said that modal regression is robust to outliers, heavy-tailed noise, skewed conditional distributions, and skewed noise distribution. How to understand and quantify the robustness of modal regression estimators?

1.4 Our Contribution

In our study we perform modal regression via the classical empirical risk minimization (ERM) within the statistical learning framework. A learning theory analysis is then conducted in order to assess the performance of the resulting modal regression estimator. Our contribution made in this study can be summarized as follows:

- The first main contribution of our study is that we present a first systematic statistical learning treatment on modal regression. This purpose is achieved by developing a statistical learning setup for modal regression, adapting it into the classical ERM framework, and conducting a learning theory analysis for modal regression estimators. The statistical learning approach to modal regression in this paper distinguishes our work from previous studies;

- The second main contribution of this study lies in that we develop a statistical learning framework for modal regression. To this end, the modal regression risk is devised, the *Bayes* rule of the modal regression risk is characterized, computationally tractable surrogates of the modal regression risk are introduced, and ERM schemes for modal regression are formulated;

- Following the ERM scheme, by assuming the existence of a global conditional mode function, the modal regression estimator in our study is pursued by maximizing a one-dimensional density estimator. This is more computationally tractable compared with the approaches adopted in most of the existing studies, in which the estimation of a possibly high-dimensional density is involved, as detailed in Section 4.5. This gives the third main contribution of this study;

- Another contribution made in this paper is that we present a learning theory analysis on the modal regression estimator resulted from the ERM scheme. The theoretical results in our analysis are concerned with the modal regression risk consistency, the generalization risk consistency, the function estimation ability of the modal regression estimator, and their relations, see Section 4 for details. Moreover, some perspectives and theoretical assessments on the robustness of modal regression estimators are also presented;

- It should be highlighted that, as we shall also explain below, the study in this paper is originally motivated by pursuing some further understanding towards the maximum correntropy based regression (MCCR) studied recently in Feng et al. (2015). Moreover, this study is started with the realization that MCCR with a tending-to-zero scale parameter is modal regression, see Section 2 for details. Thus, the conducted study brings us some new perspectives and deeper understanding towards MCCR.

notation	meaning
$\mathcal{X}$, $\mathcal{Y}$	the dependent variable space and the independent variable space, respectively
X, Y	random variables taking values in $\mathcal{X}$ and $\mathcal{Y}$, respectively
x, y	realizations of X and Y, respectively
$\mathcal{M}$	the function set comprised of all measurable function from $\mathcal{X}$ to $\mathbb{R}$
ϵ	the noise variable specified by the residual $Y - f^\star(X)$
$\mathbf{z}$	a set of n-size realizations of (X, Y) with $\mathbf{z} := \{(x_i, y_i)\}_{i=1}^n$
E_f	the random variable induced by the residual $Y - f(X)$
$\mathcal{H}$	a hypothesis space that is assumed to be a compact subset of $C(\mathcal{X})$
K_σ	a smoothing kernel with the bandwidth σ
ρ	the joint probability distribution of $X \times Y$
$\rho_\mathcal{X}$	the marginal distribution of X
$L^2_{\rho_\mathcal{X}}$	the function space of square-integrable functions with respect to $\rho_\mathcal{X}$
p_{E_f} or p_f	the density function of the random variable E_f
$p_{Y\|X}$	the conditional density of Y conditioned on X
$p_{X,Y}$	the joint density of X and Y
$p_{\epsilon\|X}$	the conditional density of ϵ conditioned on X
$f^\star$	the underlying truth function in modal regression, see formula (3.1)
f_M	the modal regression function or the conditional mode function, see formula (3.2)
$f_\mathbf{z}$	the empirical modal regression estimator in $\mathcal{H}$, see formula (3.4)
$f_{\mathcal{H},\sigma}$	the data-free modal regression estimator in $\mathcal{H}$, see formula (3.5)
$f_\mathcal{H}$	the data-free least squares regression estimator in $\mathcal{H}$
$\mathcal{R}(f)$	the modal regression risk for the hypothesis $f : \mathcal{X} \to \mathbb{R}$
$\mathcal{R}^\sigma(f)$	the data-free generalization risk for the hypothesis $f : \mathcal{X} \to \mathbb{R}$
$\mathcal{R}^\sigma_n(f)$	the empirical generalization risk for the hypothesis $f : \mathcal{X} \to \mathbb{R}$

Table 1: Some notations and their definitions in this paper

1.5 Structure of This Paper

This paper is organized as follows: in Section 2, we link the maximum correntropy based regression to modal regression by suggesting that MCCR with a tending-to-zero scale parameter is essentially modal regression. We also present a review on the existing understanding of MCCR in the regression context; in Section 3, we formulate the modal regression problem within the statistical learning framework. To this end, we introduce the modal regression function in Subsection 3.1. We define the modal regression risk and characterize its *Bayes* rule in Subsection 3.2. A kernel density estimation interpretation and an empirical risk minimization perspective of modal regression are provided in Subsections 3.3 and 3.4, respectively. Section 4 is devoted to developing a learning theory for modal regression. The modal regression calibration problem (see Subsection 4.2), the convergence rates of the excess generalization risk (see Subsection 4.3), and the function estimation calibration problem (see Subsection 4.4) are studied by applying standard learning theory arguments. Comparisons between our study and the existing studies are also described in this section. The robustness of modal regression, which serves as one of its most prominent features, will

also be discussed and studied quantitatively by means of the finite sample breakdown point in Section 5. Since one of the main motivations of the present study is to understand MCCR within the statistical learning framework and having realized that MCCR with a tending-to-zero scale parameter is modal regression, we, therefore, retrospect MCCR in Section 6 by applying the theory developed in Section 4 and depict a general picture of MCCR. Section 7 is concerned with the implementing issues in modal regression such as model selection and computational algorithms. Numerical validations on artificial and real datasets will also be provided in this section. We close this paper in Section 8 with concluding remarks. For the sake of readability, a table with some notations and their definitions in this paper is provided in Table 1.

2. Linking Correntropy based Regression to Modal Regression

As mentioned above, our study on modal regression in this paper is initiated to understand the so-called maximum correntropy criterion in regression problems (see Liu et al., 2007; Principe, 2010), which is a continuation of our previous study in Feng et al. (2015). As a generalized correlation measurement, the correntropy has been drawing much attention recently. Owing to its prominent merits on robustness, it has been pervasively used and has found many real-world applications in signal processing, machine learning, and computer vision (see Bessa et al., 2009; He et al., 2011, 2012; Lu et al., 2013; Chen et al., 2016a).

2.1 Correntropy and Correntropy based Regression

Mathematically speaking, correntropy is a generalized similarity measure between two scalar random variables U and V, which is defined by $\mathcal{R}^\sigma(U,V) = \mathbb{E}K_\sigma(U,V)$. Here K_σ is a Gaussian kernel given by $K_\sigma(u,v) = \exp\left\{-(u-v)^2/\sigma^2\right\}$ with the bandwidth $\sigma > 0$, (u,v) being a realization of (U,V). Given a set of i.i.d observations $\mathbf{z} = \{(x_i,y_i)\}_{i=1}^n$, for any $f : \mathcal{X} \to \mathbb{R}$, the empirical estimator of the correntropy between $f(X)$ and Y is given as

$$\mathcal{R}_n^\sigma(f) := \frac{1}{n}\sum_{i=1}^n K_\sigma(y_i, f(x_i)).$$

The maximum correntropy criterion based regression models the empirical target function by maximizing the empirical estimator of the correntropy $\mathcal{R}^\sigma$ as follows

$$f_{\mathbf{z}} = \arg\max_{f\in\mathcal{H}} \mathcal{R}_n^\sigma(f), \tag{2.1}$$

where $\mathcal{H}$ is assumed to be a compact subset of $C(\mathcal{X})$. Here, $C(\mathcal{X})$ is denoted as the Banach space of continuous functions on $\mathcal{X}$. The maximum correntropy criterion in regression problems has shown its efficiency for cases when the noise is non-Gaussian, and/or contaminated by outliers (see e.g., Liu et al., 2007; Principe, 2010; Wang et al., 2013).

2.2 Existing Understanding of MCCR Revisited and Limitations

In the literature, existing understanding towards the maximum correntropy criterion and MCCR is still limited. More frequently, the maximum correntropy criterion is roughly taken as a robustified least squares criterion, analogous to the trimmed least squares criterion.

However, the statistical performance of $f_{\mathbf{z}}$ and its relation to the least squares criterion is not clear. The barriers are mainly caused by the presence of the scale parameter σ and the non-convexity of the related model.

Recently, some theoretical understanding towards the maximum correntropy criterion was conducted in Feng et al. (2015) by introducing a distance-based regression loss, the study of which is inspired by the studies on information theoretic learning conducted in Hu et al. (2013) and Fan et al. (2016). In Feng et al. (2015), the distance-based regression loss is termed as the correntropy induced loss, $\ell_\sigma : \mathbb{R} \times \mathbb{R} \to [0, +\infty)$, which is defined as

$$\ell_\sigma(y,t) = \sigma^2\big(1 - e^{-(y-t)^2/\sigma^2}\big), \quad y, t \in \mathbb{R},$$

with $\sigma > 0$ being a scale parameter. Concerning the statistical performance of the regression estimator $f_{\mathbf{z}}$, in Feng et al. (2015), by imposing certain moment conditions on the independent variable, it is assessed in terms of its $\|\cdot\|_{L^2_{\rho_\mathcal{X}}}$-distance to the *mean regression function* $f_\rho = \mathbb{E}(Y|X)$ (i.e., the conditional mean). Here, the unknown underlying probability distribution over $\mathcal{X} \times \mathcal{Y}$ is denoted as ρ, and $\rho_\mathcal{X}$ is denoted as the marginal distribution of ρ on $\mathcal{X}$. More specifically, it is shown in Feng et al. (2015) that $f_{\mathbf{z}}$ can approximate f_ρ well under the capacity assumption of $\mathcal{H}$, when σ is properly chosen, and the moment condition $\mathbb{E}Y^4 < \infty$ holds. Thus, the regression estimator $f_{\mathbf{z}}$ is said to be robust in the sense of Györfi et al. (2006). Besides, the statistical performance of $f_{\mathbf{z}}$, the relations between MCCR and least squares regression were also investigated. A typical conclusion drawn there states that the correntropy criterion induced loss is an explicitly calibrated least squares loss in the sense of Lemma 7 in Feng et al. (2015). Moreover, it is argued there that the scale parameter σ plays a trade-off role between the robustness and the estimation ability of $f_{\mathbf{z}}$. It should be remarked that, according to Feng et al. (2015), the scale parameter σ in MCCR has to be chosen as $\sigma := n^\theta$ with $\theta > 0$, which is set to reduce the additional bias misspecified by the introduction of the robustness to the regression estimator $f_{\mathbf{z}}$ and eventually ensure its generalization and function estimation consistency. Here, the generalization consistency is referred to as the convergence of $\mathcal{R}^\sigma_n(f_{\mathbf{z}})$ to its population version $\mathbb{E}\mathcal{R}^\sigma_n(f_{\mathbf{z}})$ when the sample size n goes to infinity while the function estimation consistency refers to the convergence of $f_{\mathbf{z}}$ to f_ρ under the $L^2_{\rho_\mathcal{X}}$-distance.

The study in Feng et al. (2015) left open a problem of how to understand MCCR when $\sigma(n) \to 0$. Technically speaking, the conclusion drawn in Feng et al. (2015) is not applicable to this case due to the lack of the regression calibration relation between generalization and function estimation, i.e., Lemma 7 in Feng et al. (2015). This observation is also reported in a related study on the minimum error entropy criterion based regression in Fan et al. (2016). A counterexample was devised there showing that when $\sigma(n) \to 0$, there exist cases in which the generalization consistency and the estimation consistency cannot be pursued simultaneously. However, it remains unclear what would happen to MCCR if $\sigma(n) \to 0$. Empirical studies showed that MCCR preserves the robustness and prediction ability even if a small scale parameter σ is chosen. Therefore, some specific theoretical understanding of this case is needed, which, indeed, originally motivates us to initiate this study.

2.3 MCCR with a Tending-to-Zero Scale Parameter is Modal Regression

Our study conducted in this paper shows that when the scale parameter $\sigma(n)$ tends to zero, MCCR is, in fact, modal regression. Model regression regresses towards the mode of the conditional distribution of the independent variable conditioned on the dependent variable. In order words, the maximization of the empirical risk in (2.1) encourages the resulting estimator $f_{\mathbf{z}}$ move towards the *conditional mode function*. Recall that regression towards the conditional mode can outperform the conditional mean and the conditional quantile regression in terms of the robustness and the trends estimation of the observations. This in return explains the robustness of MCCR when $\sigma(n) \to 0$.

3. A Statistical Learning Framework for Modal Regression

3.1 Formulating the Modal Regression Problem

We first formulate the modal regression problem formally in this subsection. To this end, we first assume that we are given a set of i.i.d observations $\mathbf{z}$ that are generated by

$$Y = f^\star(X) + \epsilon, \tag{3.1}$$

where the mode of the conditional distribution of ϵ at any $x \in \mathcal{X}$ is assumed to be zero. That is, $\mathsf{mode}(\epsilon \,|\, X = x) := \arg\max_{t\in\mathbb{R}} p_{\epsilon|X}(t \,|\, X = x) = 0$ for any $x \in \mathcal{X}$, where $p_{\epsilon|X}$ is the conditional density of ϵ conditioned on X. It is obvious from (3.1) that under the zero-mode noise assumption, there holds $\mathsf{mode}(Y|X) = f^\star(X)$. We further assume that $p_{\epsilon|X}$ is continuous and bounded on $\mathbb{R}$ for any $x \in \mathcal{X}$. Here, it should be remarked that in this study we do not assume either the homogeneity or the symmetry of the distribution of the noise ϵ. In other words, the heterogeneity of the distribution of the residuals or the skewed noise distribution is allowed.

In modal regression problems, we aim at approximating the *modal regression function* (see formula (1.1) in Collomb et al. (1987)):

Definition 1 (Modal Regression Function) *The* ***modal regression function*** $f_{\mathrm{M}} : \mathcal{X} \to \mathbb{R}$ *is defined as*

$$f_{\mathrm{M}}(x) := \arg\max_{t\in\mathbb{R}} p_{Y|X}(t|\, X = x),\ x \in \mathcal{X}, \tag{3.2}$$

where $p_{Y|X}(\cdot|X)$ *denotes the conditional density of* Y *conditioned on* X*.*

Throughout this paper, we assume that the modal regression function f_{M} is well-defined on $\mathcal{X}$. That is, $\arg\max_{t\in\mathbb{R}} p_{Y|X}(t \mid X = x)$ is assumed to exist and be unique for any fixed $x \in \mathcal{X}$. Obviously, this is equivalent to assuming the existence and uniqueness of the global mode of the conditional density $p_{Y|X}$. On the other hand, due to the zero-mode assumption of the conditional distribution of ϵ in (3.1) for any $x \in \mathcal{X}$, we know that $f_{\mathrm{M}} \equiv f^\star$. Consequently, the learning for modal regression problem is equivalent to the problem of learning the modal regression function f_{M}, and thus $f^\star$. Said differently, f_{M} is the so-called target hypothesis.

From the definition, the modal regression function f_{M} is defined as the maximum of the conditional density $p_{Y|X}$ conditioned on X. Note that, maximizing the conditional

density is equivalent to maximizing the joint density $p_{X,Y}$ for any fixed realization of X. Therefore, it is direct to see that one can approximate f_{M} by maximizing the conditional density $p_{Y|X}$ or the joint density $p_{X,Y}$, both of which can be estimated via kernel density estimation. This is, in fact, what most of the existing studies on modal regression do (see e.g., Collomb et al., 1987; Chen et al., 2016b; Yao and Xiang, 2016). However, estimating the conditional density $p_{Y|X}$ or the joint density $p_{X,Y}$ via kernel density estimation suffers from the *curse of dimensionality* and is not feasible when the dimension of the input space is high. In this study, we are interested in an empirical risk minimization approach that is dimension-insensitive as formulated below.

3.2 Modeling the Modal Regression Risk and Characterizing the *Bayes* Rule

To be in position to carry out a statistical learning assessment of modal regression, besides the target hypothesis defined above, we also need to devise a fitting risk that measures the goodness-of-fit when a candidate hypothesis is considered. The newly devised fitting risk should vote the target hypothesis as the best candidate when the hypothesis space is sufficiently large. This gives the main purpose of this subsection.

Definition 2 (Modal Regression Risk) *For a measurable function $f : \mathcal{X} \to \mathbb{R}$, its **modal regression risk** $\mathcal{R}(f)$ is defined as*

$$\mathcal{R}(f) = \int_{\mathcal{X}} p_{Y|X}(f(x)|X = x)\mathrm{d}\rho_{\mathcal{X}}(x). \tag{3.3}$$

Analogous to learning for regression and classification scenarios (see, e.g., Cucker and Zhou, 2007; Steinwart and Christmann, 2008), we denote the ***Bayes* (decision) rule** of modal regression as the "best" hypothesis favored by the above modal regression risk over the measurable function set $\mathcal{M}$ (comprised of all measurable functions from $\mathcal{X}$ to $\mathbb{R}$). The following conclusion indicates that the target hypothesis f_{M} is exactly the *Bayes* rule of modal regression.

Theorem 3 *The modal regression function f_{M} in (3.2) gives the Bayes rule of modal regression. That is,*

$$f_{\mathrm{M}} = \arg\max_{f \in \mathcal{M}} \mathcal{R}(f).$$

Proof Recall that the conditional mode function f_{M} is given as

$$f_{\mathrm{M}}(x) = \arg\max_{t \in \mathbb{R}} p_{Y|X}(t \,|\, X = x),\ x \in \mathcal{X}.$$

Following the modal regression risk defined in Definition 2, for any measurable function $f \in \mathcal{M}$, we have

$$\mathcal{R}(f) = \int_{\mathcal{X}} p_{Y|X}(f(x)|X = x)\mathrm{d}\rho_{\mathcal{X}}(x) \le \int_{\mathcal{X}} p_{Y|X}(f_{\mathrm{M}}(x)|X = x)\mathrm{d}\rho_{\mathcal{X}}(x) = \mathcal{R}(f_{\mathrm{M}}),$$

which directly yields

$$f_{\mathrm{M}} = \arg\max_{f \in \mathcal{M}} \mathcal{R}(f).$$

This completes the proof of Theorem 3. ■

The plausibility of the above-defined modal regression risk stems from the fact that f_{M} is the *Bayes* rule of modal regression, as justified by Theorem 3. With the modal regression risk being defined and recalling that f_{M} maximizes the modal regression risk, the most direct way to learn f_{M} is to maximize the sample analogy of the modal regression risk. Unfortunately, this is intractable since the discretization of an unknown conditional density is involved. In the next subsection, to circumvent this problem, we introduce a surrogate of the modal regression risk.

Remark 4 *We now give a remark on the terminology "risk". For any measurable function $f : \mathcal{X} \to \mathbb{R}$, the modal regression risk $\mathcal{R}(f)$ in Definition 2 can be regarded as a measure of the extent to which the function f fits the Bayes rule f_{M} in the $\mathcal{R}(\cdot)$ sense. Therefore, the terminology "risk" is not used as what is commonly referred to in the statistical learning literature. However, in what follows, given the one-to-one correspondence between the corresponding maximization and minimization problems, we still term $\mathcal{R}(f)$ as the (modal regression) risk of f.*

3.3 Learning for Modal Regression via Kernel Density Estimation

We now show that the modal regression problem can be tackled by applying the kernel density estimation technique. To this purpose, let $f : \mathcal{X} \to \mathbb{R}$ be a measurable function and denote E_f as the random variable induced by the residual $Y - f(X)$, where the subscript f indicates its dependence on f. We also denote p_{E_f}, or simply p_f, as the density function of the random variable E_f, and denote $p_{\epsilon|X}$ as the conditional density of the random variable $\epsilon = Y - f^\star(X)$. The following theorem relates the modal regression risk of f to $p_{\epsilon|X}$ and p_{E_f}.

Theorem 5 *Let $f : \mathcal{X} \to \mathbb{R}$ be a measurable function. Then,*

$$\int_{\mathcal{X}} p_{\epsilon|X}(\cdot + f(x) - f^\star(x)|X = x)\mathrm{d}\rho_{\mathcal{X}}(x).$$

is a density of the random variable $E_f := Y - f(X)$, which we denote as p_{E_f}. Correspondingly, we have $p_{E_f}(0) = \mathcal{R}(f)$.

Proof From the model assumption that $\epsilon = Y - f^\star(X)$, we have

$$\epsilon = E_f + f(X) - f^\star(X).$$

As a result, the density function of the error variable E_f can be expressed as

$$\int_{\mathcal{X}} p_{\epsilon|X}(\cdot + f(x) - f^\star(x)|X = x)\mathrm{d}\rho_{\mathcal{X}}(x).$$

and denoted by p_{E_f}. Moreover, from the definition of the risk functional $\mathcal{R}(\cdot)$ in (3.3), we know that

$$\begin{aligned} p_{E_f}(0) &= \int_{\mathcal{X}} p_{\epsilon|X}(f(x) - f^\star(x)|X = x)\mathrm{d}\rho_{\mathcal{X}}(x) \\ &= \int_{\mathcal{X}} p_{Y|X}(f(x)|X = x)\mathrm{d}\rho_{\mathcal{X}}(x) \\ &= \mathcal{R}(f). \end{aligned}$$

This completes the proof of Theorem 5. ■

From Theorem 5, the hypothesis f that maximizes the modal regression risk $\mathcal{R}(f)$ is the one that maximizes the density of $E_f := Y - f(X)$ at 0, which can be estimated non-parametrically. In this study, the kernel density estimation technique is tailored to modal regression with the help of the modal regression kernel defined below.

Definition 6 (Modal Regression Kernel) *A kernel $K_\sigma : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ is said to be a **modal regression kernel** with the representing function ϕ and the bandwidth parameter $\sigma > 0$ if there exists a function $\phi : \mathbb{R} \to [0, \infty)$ such that $K_\sigma(u_1, u_2) = \phi\left(\frac{u_1 - u_2}{\sigma}\right)$ for any $u_1, u_2 \in \mathbb{R}$, $\phi(u) = \phi(-u)$, $\phi(u) \leq \phi(0)$ for any $u \in \mathbb{R}$, and $\int_{\mathbb{R}} \phi(u)\mathrm{d}u = 1$.*

According to Definition 6, it is easy to see that common smoothing kernels (see, e.g., Wand and Jones, 1994) such as the Naive kernel, the Gaussian kernel, the Epanechnikov kernel, and the Triangular kernel are modal regression kernels. Their corresponding representing functions can be easily deduced with simple computations. For a modal regression kernel K_σ with the representing function ϕ, throughout this paper, without loss of generality, we assume $\phi(0) = 1$.

As a consequence of Theorem 5, for any measurable function f, we know that $p_f(0) = \mathcal{R}(f)$. With the help of a modal regression kernel K_σ, it is immediate to see that an empirical kernel density estimator $\hat{p}_f$ for p_f at 0 can be formulated as follows

$$\hat{p}_f(0) = \frac{1}{n\sigma}\sum_{i=1}^{n} K_\sigma(y_i - f(x_i), 0) = \frac{1}{n\sigma}\sum_{i=1}^{n} K_\sigma(y_i, f(x_i)) := \mathcal{R}_n^\sigma(f).$$

Therefore, when confined to a hypothesis space $\mathcal{H}$, learning a function f that maximizes the modal regression risk is cast as learning the function f that maximizes the value of the empirical density estimator $\hat{p}_f$ at 0. Thus, the empirical target hypothesis is modeled as

$$\begin{aligned} f_{\mathbf{z},\sigma} :&= \arg\max_{f\in\mathcal{H}} \hat{p}_f(0) \\ &= \arg\max_{f\in\mathcal{H}} \mathcal{R}_n^\sigma(f), \end{aligned} \tag{3.4}$$

where $\mathcal{H}$ is assumed to be a compact subset of $C(\mathcal{X})$ throughout this paper. Specifically, when the smoothing kernel K_σ is chosen as the Gaussian kernel, the above risk maximization scheme retrieves MCCR. The population version of $f_{\mathbf{z},\sigma}$ can be expressed as

$$f_{\mathcal{H},\sigma} := \arg\max_{f\in\mathcal{H}} \mathcal{R}^\sigma(f), \tag{3.5}$$

where $\mathcal{R}^\sigma(\cdot)$ is the expectation of $\mathcal{R}_n^\sigma(\cdot)$ with respect to the random samples $\mathbf{z}$ and for any $f : \mathcal{X} \to \mathbb{R}$, it can be expressed as

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

The risk functional $\mathcal{R}^\sigma(f)$ defined above gives the **generalization risk** of f when a modal regression kernel K_σ with the representing function ϕ is adopted. As we shall see later, it can be seen as a surrogate of the true modal regression risk $\mathcal{R}(f)$ since $\mathcal{R}^\sigma(f)$ approximates $\mathcal{R}(f)$ when $\sigma \to 0$. The interpretation of modal regression from a kernel density estimation view explains the requirement that $\int_{\mathbb{R}} \phi(u)\mathrm{d}u = 1$ in Definition 6.

Remark 7 *In the sequel, for notational convenience, we will drop the subscript σ in $f_{\mathbf{z},\sigma}$ by denoting $f_{\mathbf{z}} := f_{\mathbf{z},\sigma}$. Additionally, given the above kernel density estimation interpretation of modal regression, the scale parameter σ in* (3.4) *is also referred to as the "bandwidth parameter".*

3.4 Modal Regression: an Empirical Risk Minimization View

In the preceding subsection, we showed that the modal regression scheme (3.4) can be interpreted from a kernel density estimation view. Maximizing the value of the kernel density estimator for E_f at 0 encourages the considered hypothesis f to approximate the projection of the *Bayes* rule onto $\mathcal{H}$, i.e., $f_{\mathcal{H},\sigma}$. In this subsection, we show that one can also interpret the modal regression scheme (3.4) within the empirical risk minimization framework.

To proceed, let us consider a modal regression kernel K_σ with the representing function ϕ and the scale parameter $\sigma > 0$. We then introduce the following distance-based modal regression loss $\phi_\sigma : \mathbb{R} \to [0, \infty)$:

$$\phi_\sigma(y - f(x)) = \sigma^{-1}\left(1 - \phi\left((y - f(x))\sigma^{-1}\right)\right). \tag{3.6}$$

Based on the newly introduced loss ϕ_σ, the modal regression scheme (3.4) can be reformulated as follows

$$f_{\mathbf{z}} = \arg\min_{f\in\mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \phi_\sigma(y_i, f(x_i)), \tag{3.7}$$

and, similarly, its data-free counterpart can be formulated as

$$f_{\mathcal{H},\sigma} = \arg\min_{f\in\mathcal{H}} \int_{\mathcal{X}\times\mathcal{Y}} \phi_\sigma(y, f(x))\mathrm{d}\rho. \tag{3.8}$$

It is easy to see that the empirical estimator (3.7) is an M-estimator and the two formulations of $f_{\mathbf{z}}$ in (3.4) and (3.7) are, in fact, equivalent. Similarly, one also obtains the same target function from the two learning paradigms (3.5) and (3.8). As we shall discuss later, the versatility of modal regression kernels entails the adaptation of robustness to the modal regression estimator (3.7).

Remark 8 *For formulation simplification, whenever referred to herein, $f_{\mathbf{z}}$ and $f_{\mathcal{H},\sigma}$ will be pointed to the estimators formulated by* (3.4) *and* (3.5)*, respectively, while keeping in mind that the conducted analysis on $f_{\mathbf{z}}$ is inspired by and within the ERM framework.*

4. A Learning Theory of Modal Regression

In this section, we aim to develop a learning theory for modal regression which explores the statistical learning performance of modal regression estimators resulted from the ERM approach formulated above.

4.1 Learning the Conditional Mode: Three Building Blocks

In Section 3, for a given hypothesis f, the modal regression risk $\mathcal{R}(f)$ is defined; moreover, it turns out that f_{M} is the *Bayes* rule of modal regression. On the other hand, we show that the modal regression estimator can be learned via maximizing the risk functional $\mathcal{R}_n^\sigma(\cdot)$. Recalling that the central concern in learning theory is risk consistency under various notions and following the clue of existing learning theory studies on the binary-classification problem, it is natural and necessary to investigate the following three problems:

1. The problem of the excess generalization risk consistency and convergence rates, i.e., the convergence from $\mathcal{R}^\sigma(f_{\mathbf{z}})$ to $\mathcal{R}^\sigma(f^\star)$;

2. The modal regression calibration problem, i.e., whether the convergence from $\mathcal{R}^\sigma(f_{\mathbf{z}})$ to $\mathcal{R}^\sigma(f^\star)$ implies the convergence from $\mathcal{R}(f_{\mathbf{z}})$ to $\mathcal{R}(f^\star)$?

3. The function estimation calibration problem, i.e., whether the convergence from $\mathcal{R}(f_{\mathbf{z}})$ to $\mathcal{R}(f^\star)$ implies the convergence from $f_{\mathbf{z}}$ to $f^\star$?

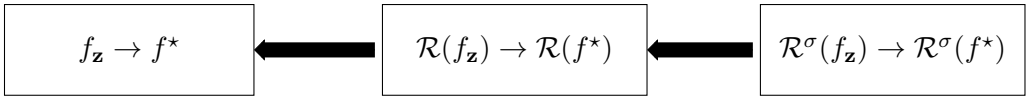

Figure 1: An illustration of the three building blocks in learning for modal regression. The left block stands for the function estimation consistency of $f_{\mathbf{z}}$, the middle block denotes the modal regression consistency of $f_{\mathbf{z}}$, while the right block represents the excess generalization risk consistency of $f_{\mathbf{z}}$.

The above three problems are fundamental in conducting a learning theory analysis on modal regression, and serve as three main building blocks. Detailed explorations will be expanded in the following subsections.

4.2 Towards the Modal Regression Calibration Problem

We first investigate the modal regression calibration problem stated in Question 1 proposed in the above subsection, i.e., whether the convergence from $\mathcal{R}^\sigma(f_{\mathbf{z}})$ to $\mathcal{R}^\sigma(f^\star)$ implies the convergence from $\mathcal{R}(f_{\mathbf{z}})$ to $\mathcal{R}(f^\star)$. To this end, we need to confine ourselves to the calibrated modal regression kernel defined below.

Definition 9 (Calibrated Modal Regression Kernel) *A modal regression kernel K_σ with the representing function ϕ is said to be a* ***calibrated modal regression kernel*** *if it satisfies the following conditions:*

(i) ϕ is bounded;

(*ii*) $\int_{\mathbb{R}} u^2\phi(u)\mathrm{d}u < \infty$.

Another restriction one needs to impose is on the conditional density $p_{\epsilon|X}$ as follows:

Assumption 1 *The conditional density of ϵ given X, namely, $p_{\epsilon|X}$, is second-order continuously differentiable, and $\|p''_{\epsilon|X}\|_\infty$ is bounded from above.*

Theorem 10 *Suppose that Assumption 1 holds and let K_σ be a calibrated modal regression kernel with the representing function ϕ and the scale parameter σ. For any measurable function $f : \mathcal{X} \to \mathbb{R}$, there holds*

$$\left|\{\mathcal{R}(f^\star) - \mathcal{R}(f)\} - \{\mathcal{R}^\sigma(f^\star) - \mathcal{R}^\sigma(f)\}\right| \leq c_1\sigma^2,$$

where $c_1 = \|p''_{\epsilon|X}\|_\infty \int_{\mathbb{R}} u^2\phi(u)\mathrm{d}u$.

Proof Recalling the definition of the risk functional $\mathcal{R}^\sigma(f)$ for any measurable function $f : \mathcal{X} \to \mathbb{R}$ and applying Taylor's Theorem to the conditional density $p_{\epsilon|X}$, we have

$$\begin{aligned}
\mathcal{R}^\sigma(f) &= \frac{1}{\sigma}\int_{\mathcal{X}\times\mathcal{Y}} \phi\left(\frac{y - f(x)}{\sigma}\right)\mathrm{d}\rho(x,y) \\
&= \frac{1}{\sigma}\int_{\mathcal{X}}\int_{\mathbb{R}} \phi\left(\frac{t - (f(x) - f^\star(x))}{\sigma}\right) p_{\epsilon|X}(t \mid X = x)\mathrm{d}t\,\mathrm{d}\rho_{\mathcal{X}}(x) \\
&= \int_{\mathcal{X}}\int_{\mathbb{R}} \phi(u) p_{\epsilon|X}(f(x) - f^\star(x) + \sigma u \mid X = x)\mathrm{d}u\mathrm{d}\rho_{\mathcal{X}}(x) \\
&= \int_{\mathcal{X}}\int_{\mathbb{R}} \phi(u) p_{\epsilon|X}(f(x) - f^\star(x) \mid X = x)\mathrm{d}u\mathrm{d}\rho_{\mathcal{X}}(x) \\
&\quad + \sigma\int_{\mathcal{X}}\int_{\mathbb{R}} u\phi(u) p'_{\epsilon|X}(f(x) - f^\star(x) \mid X = x)\mathrm{d}u\mathrm{d}\rho_{\mathcal{X}}(x) \\
&\quad + \frac{\sigma^2}{2}\int_{\mathcal{X}}\int_{\mathbb{R}} u^2\phi(u) p''_{\epsilon|X}(\eta_x \,|\, X = x)\mathrm{d}u\mathrm{d}\rho_{\mathcal{X}}(x),
\end{aligned} \tag{4.1}$$

where, for any fixed $x \in \mathcal{X}$, the point η_x lies between $f(x) - f^\star(x)$ and $f(x) - f^\star(x) + \sigma u$.

The fact that K_σ is a calibrated modal regression kernel with the representing function ϕ ensures $\int_{\mathbb{R}} \phi(u)\mathrm{d}u = 1$ and reminds the symmetry of ϕ on $\mathbb{R}$, which further indicates that $\int_{\mathbb{R}} u\phi(u)\mathrm{d}u = 0$. On the other hand, the fact that

$$\mathcal{R}(f) = \int_{\mathcal{X}} p_{\epsilon|X}(f(x) - f^\star(x)|X = x)\mathrm{d}\rho_{\mathcal{X}}(x),$$

together with Equalities (4.1) yields

$$|\mathcal{R}^\sigma(f) - \mathcal{R}(f)| \leq \frac{\sigma^2}{2}\left(\|p''_{\epsilon|X}\|_\infty \int_{\mathbb{R}} u^2\phi(u)\mathrm{d}u\right).$$

Denoting $c_1 := \|p''_{\epsilon|X}\|_\infty \int_{\mathbb{R}} u^2\phi(u)\mathrm{d}u$, we accomplish the proof of Theorem 10. ■

Remark 11 *The proof of Theorem 10 indicates that $\mathcal{R}^\sigma(f)$ is a second-order approximation (with respect to σ) of $\mathcal{R}(f)$ since $\mathcal{R}^\sigma(f) - \mathcal{R}(f) = \mathcal{O}(\sigma^2)$. In fact, if a higher-order kernel (see e.g., Section 2.8 in Wand and Jones, 1994) is used, a higher-order approximation of $\mathcal{R}(f)$ can be expected.*

From the proof of Theorem 10, we see that when K_σ is a calibrated modal regression kernel with the representing function ϕ and the scale parameter σ, for any measurable function $f : \mathcal{X} \to \mathbb{R}$, the generalization risk $\mathcal{R}^\sigma(f)$ approaches the true modal regression risk $\mathcal{R}(f)$ provided that $\sigma \to 0$. Therefore, in the above sense, $\mathcal{R}^\sigma(f)$ can be considered as a relaxation of $\mathcal{R}(f)$. On the other hand, Theorem 10 indicates that the discrepancy between the **excess modal regression risk** $\mathcal{R}(f^\star) - \mathcal{R}(f)$ and the **excess generalization risk** $\mathcal{R}^\sigma(f^\star) - \mathcal{R}^\sigma(f)$ can be upper bounded by $\mathcal{O}(\sigma^2)$. Clearly, under the assumptions of Theorem 10, when $\sigma \to 0$, $\mathcal{R}^\sigma(f^\star) - \mathcal{R}^\sigma(f)$ also approaches $\mathcal{R}(f^\star) - \mathcal{R}(f)$. In this sense, Theorem 10 establishes a *comparison theorem* akin to the one in the classification scenario (see Zhang, 2004; Bartlett et al., 2006). This elucidates the terminology—the calibrated modal regression kernel, and the terminology—the modal regression calibration problem.

4.3 Towards the Convergence Rates of the Excess Generalization Risk

One of the main focuses in learning theory is the generalization ability of a learning algorithm that measures its out-of-sample prediction ability. It plays an important role in designing learning algorithms with theoretical guarantees. In this subsection, we derive the generalization bounds for the modal regression estimator $f_{\mathbf{z}}$, i.e., the convergence rates of $\mathcal{R}^\sigma(f^\star) - \mathcal{R}^\sigma(f_{\mathbf{z}})$, by means of learning theory arguments. The following assumption is needed for this purpose:

Assumption 2 *We make the following assumptions:*

(i) *There exists a positive constant M such that $\|f^\star\|_\infty \leq M$;*

(ii) $\sup_{t\in\mathbb{R},\, x\in\mathcal{X}} p_{\epsilon|X}(t \mid X = x) = c_2 < \infty$;

(iii) *For any $\varepsilon > 0$, there exists an exponent p with $0 < p < 2$ such that the ℓ^2-empirical covering number (with radius ε) of $\mathcal{H}$, denoted as $\mathcal{N}_{2,\mathbf{x}}(\mathcal{H}, \varepsilon)$, satisfies*

$$\log \mathcal{N}_{2,\mathbf{x}}(\mathcal{H}, \varepsilon) \lesssim \varepsilon^{-p},$$

where the definition of the empirical covering number is provided in the appendix, and the notation $a \lesssim b$ for $a, b \in \mathbb{R}$ means that there exists a positive constant c such that $a \leq cb$.

Restrictions in Assumption 2 are fairly standard if we recall that the hypothesis space $\mathcal{H}$ is assumed to be a compact subset of $\mathcal{C}(\mathcal{X})$. In what follows, without loss of generality, we also assume that $\|f\|_\infty \leq M$ for any $f \in \mathcal{H}$. The following error decomposition lemma is helpful in bounding the excess generalization error.

Lemma 12 *Let $f_{\mathbf{z}}$ be produced by* (3.4) *and assume that $f^\star \in \mathcal{H}$. Then we have*

$$\mathcal{R}^\sigma(f^\star) - \mathcal{R}^\sigma(f_{\mathbf{z}}) \leq \mathcal{R}^\sigma(f_{\mathcal{H},\sigma}) - \mathcal{R}^\sigma_n(f_{\mathcal{H},\sigma}) + \mathcal{R}^\sigma_n(f_{\mathbf{z}}) - \mathcal{R}^\sigma(f_{\mathbf{z}}).$$

Proof Recalling that $f_{\mathcal{H},\sigma} = \arg\max_{f\in\mathcal{H}} \mathcal{R}^\sigma(f)$, we have

$$\begin{aligned}\mathcal{R}^\sigma(f^\star) - \mathcal{R}^\sigma(f_{\mathbf{z}}) &\leq \mathcal{R}^\sigma(f_{\mathcal{H},\sigma}) - \mathcal{R}^\sigma(f_{\mathbf{z}}) \\ &\leq \mathcal{R}^\sigma(f_{\mathcal{H},\sigma}) - \mathcal{R}_n^\sigma(f_{\mathcal{H},\sigma}) + \mathcal{R}_n^\sigma(f_{\mathcal{H},\sigma}) - \mathcal{R}_n^\sigma(f_{\mathbf{z}}) + \mathcal{R}_n^\sigma(f_{\mathbf{z}}) - \mathcal{R}^\sigma(f_{\mathbf{z}}) \\ &\leq \mathcal{R}^\sigma(f_{\mathcal{H},\sigma}) - \mathcal{R}_n^\sigma(f_{\mathcal{H},\sigma}) + \mathcal{R}_n^\sigma(f_{\mathbf{z}}) - \mathcal{R}^\sigma(f_{\mathbf{z}}),\end{aligned}$$

where the last inequality is due to the fact that the quantity $\mathcal{R}_n^\sigma(f_{\mathcal{H},\sigma}) - \mathcal{R}_n^\sigma(f_{\mathbf{z}})$ is at most zero. This completes the proof of Lemma 12. ■

The following lemma, established in Wu et al. (2007), provides a Bernstein-type concentration inequality for function-valued random variables. It was proved by applying the local Rademacher complexity arguments developed in Bartlett et al. (2005).

Lemma 13 *Let $\mathcal{F}$ be a class of bounded measurable functions. Assume that there are constants $\gamma \in [0,1]$ and $B, c_\gamma > 0$ such that $\|f\|_\infty \leq B$ and $\mathbb{E}f^2 \leq c_\gamma(\mathbb{E}f)^\gamma$ for every $f \in \mathcal{F}$. If for some $c_p > 0$ and $0 < p < 2$,*

$$\log \mathcal{N}_{2,\mathbf{x}}(\mathcal{F}, \varepsilon) \leq c_p \varepsilon^{-p}, \ \forall \varepsilon > 0,$$

then there exists a constant c_p' depending only on p such that for any $t > 0$, with probability at least $1 - e^{-t}$, there holds

$$\mathbb{E}f - \frac{1}{n}\sum_{i=1}^n f(z_i) \leq \frac{1}{2}\eta^{1-\gamma}(\mathbb{E}f)^\gamma + c_p'\eta + 2\left(\frac{c_\gamma t}{n}\right)^{\frac{1}{2-\gamma}} + \frac{18Bt}{n}, \ \forall f \in \mathcal{F},$$

where

$$\eta = \max\left\{c_\gamma^{\frac{2-p}{4-2\gamma+p\gamma}}\left(\frac{c_p}{n}\right)^{\frac{2}{4-2\gamma+p\gamma}}, B^{\frac{2-p}{2+p}}\left(\frac{c_p}{n}\right)^{\frac{2}{2+p}}\right\}.$$

Theorem 14 *Suppose that Assumption 2 holds, $f^\star \in \mathcal{H}$, and the risk functional $\mathcal{R}^\sigma(\cdot)$ is defined in association with a calibrated modal regression kernel K_σ and the representing function ϕ. Let $f_{\mathbf{z}}$ be produced by (3.4). Then for any $0 < \delta < 1$, with probability at least $1 - \delta$, there holds*

$$\mathcal{R}^\sigma(f^\star) - \mathcal{R}^\sigma(f_{\mathbf{z}}) \lesssim \frac{(1+\sigma)\log\delta^{-1}}{\sqrt{\sigma n}}.$$

Proof We prove the theorem by applying Lemma 13 to the following function-valued random variable on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$:

$$\xi(z) := \frac{1}{\sigma}\phi\left(\frac{y - f_{\mathcal{H},\sigma}(x)}{\sigma}\right) - \frac{1}{\sigma}\phi\left(\frac{y - f(x)}{\sigma}\right), \tag{4.2}$$

where $f_{\mathcal{H},\sigma}$ is given in (3.5) and $f \in \mathcal{H}$. Due to the boundedness assumption of ϕ, it is easy to see that $|\xi(z)| \leq 2\|\phi\|_\infty/\sigma$. Moreover, recalling the definition of the risk functional $\mathcal{R}^\sigma(\cdot)$, the following estimate holds

$$\begin{aligned}\mathbb{E}\xi^2 &= \mathbb{E}\left[\frac{1}{\sigma}\phi\left(\frac{Y - f_{\mathcal{H},\sigma}(X)}{\sigma}\right) - \frac{1}{\sigma}\phi\left(\frac{Y - f(X)}{\sigma}\right)\right]^2 \\ &\leq \frac{2\|\phi\|_\infty}{\sigma}(\mathcal{R}^\sigma(f_{\mathcal{H},\sigma}) + \mathcal{R}^\sigma(f)).\end{aligned} \tag{4.3}$$

From the proof of Theorem 10, we know that

$$\mathcal{R}^{\sigma}(f_{\mathcal{H},\sigma}) \leq \mathcal{R}(f_{\mathcal{H},\sigma}) + \frac{\sigma^2}{2}\left(\|p''_{\epsilon|X}\|_{\infty}\int_{\mathbb{R}} u^2\phi(u)\mathrm{d}u\right).$$

Similarly, there also holds

$$\mathcal{R}^{\sigma}(f) \leq \mathcal{R}(f) + \frac{\sigma^2}{2}\left(\|p''_{\epsilon|X}\|_{\infty}\int_{\mathbb{R}} u^2\phi(u)\mathrm{d}u\right).$$

The above two estimates together with the estimate for $\mathbb{E}\xi^2$ in (4.3) yield

$$\begin{aligned}\mathbb{E}\xi^2 &\leq \frac{2\|\phi\|_{\infty}}{\sigma}(\mathcal{R}(f_{\mathcal{H},\sigma}) + \mathcal{R}(f) + c_1\sigma^2)\\ &\leq \frac{2\|\phi\|_{\infty}}{\sigma}(p_{f_{\mathcal{H},\sigma}}(0) + p_f(0) + c_1\sigma^2)\\ &\lesssim \sigma^{-1} + \sigma,\end{aligned}$$

where the last inequality is due to the boundedness assumption of the conditional density of ϵ while the second inequality is a consequence of Theorem 5. Applying Lemma 13 to the random variable ξ with $B = 2\|\phi\|_{\infty}/\sigma$, $\gamma = 0$, and $c_{\gamma} = \sigma^{-1} + \sigma$, we see that for any $0 < \delta < 1$, with probability at least $1 - \delta$, there holds

$$\mathcal{R}^{\sigma}(f_{\mathcal{H},\sigma}) - \mathcal{R}^{\sigma}(f) - (\mathcal{R}^{\sigma}_n(f_{\mathcal{H},\sigma}) - \mathcal{R}^{\sigma}_n(f)) \lesssim \frac{(1+\sigma)\log\delta^{-1}}{\sqrt{\sigma n}}.$$

Notice that the above inequality holds for any $f \in \mathcal{H}$. Therefore, for any $0 < \delta < 1$, with probability at least $1 - \delta$, there holds

$$\mathcal{R}^{\sigma}(f_{\mathcal{H},\sigma}) - \mathcal{R}^{\sigma}(f_{\mathbf{z}}) - (\mathcal{R}^{\sigma}_n(f_{\mathcal{H},\sigma}) - \mathcal{R}^{\sigma}_n(f_{\mathbf{z}})) \lesssim \frac{(1+\sigma)\log\delta^{-1}}{\sqrt{\sigma n}}.$$

This in connection with Lemma 12 yields the desired estimate in Theorem 14. ■

The generalization bounds in Theorem 14 are derived for the case when the parameter σ goes to zero in accordance with the sample size. When σ goes large, sharper bounds can be derived as shown below.

Theorem 15 *Suppose that Assumption 2 holds, $f^{\star} \in \mathcal{H}$, and the risk functional $\mathcal{R}^{\sigma}(\cdot)$ is defined in association with a calibrated modal regression kernel K_{σ} and the corresponding representing function ϕ. Let $f_{\mathbf{z}}$ be produced by (3.4). Then for any $0 < \delta < 1$, with probability at least $1 - \delta$, there holds*

$$\mathcal{R}^{\sigma}(f^{\star}) - \mathcal{R}^{\sigma}(f_{\mathbf{z}}) \lesssim \frac{\log\delta^{-1}}{\sigma\sqrt{n}}.$$

Proof Similar to the proof of Theorem 14, the desired estimate can be proved by applying Lemma 13 to the random variable ξ in (4.2) with the only difference in bounding $\mathbb{E}\xi^2$. Recall

that for a calibrated modal regression kernel K_σ, its representing function ϕ is bounded. Therefore, we have

$$\mathbb{E}\xi^2 = \mathbb{E}\left[\frac{1}{\sigma}\phi\left(\frac{Y - f_{\mathcal{H},\sigma}(X)}{\sigma}\right) - \frac{1}{\sigma}\phi\left(\frac{Y - f(X)}{\sigma}\right)\right]^2 \lesssim \sigma^{-2}.$$

In order to accomplish the proof, it suffices to apply Lemma 13 to the random variable ξ with $B = 2\|\phi\|_\infty/\sigma$, $\gamma = 0$, and $c_\gamma = \sigma^{-2}$. By following the same procedure, the desired estimate in Theorem 15 can be obtained. ■

The ERM learning scheme (3.4) is adaptive in that the scale parameter σ may vary in correspondence to the sample size n, e.g., $\sigma = n^\theta$ with $\theta \in \mathbb{R}$. Note from Theorem 15 that, with a properly chosen σ, the ERM scheme (3.4) is generalization consistent in the sense that the generalization risk $\mathcal{R}^\sigma(f_{\mathbf{z}})$ converges to $\mathcal{R}^\sigma(f^\star)$ when the sample size n tends to infinity. It is also interesting to note that a wide range of σ values is admitted to ensure such a consistency property as shown in the following corollary.

Corollary 16 *Under the assumptions of Theorem 15, the ERM learning scheme* (3.4) *is generalization consistent when* $\sigma := n^\theta$ *with* $\theta \in (-1/2, +\infty)$.

Corollary 16 is an immediate result of Theorem 15 and its proof is omitted here. With a properly chosen σ, the following conclusion reveals that the ERM scheme (3.4) is also modal regression consistent. This gives an affirmative answer to the Question 2 proposed at the beginning of Subsection 4.1.

Theorem 17 *Suppose that Assumptions 1, 2 hold, and* $f^\star \in \mathcal{H}$. *Let* $f_{\mathbf{z}}$ *be produced by* (3.4) *which is induced by a calibrated modal regression kernel* K_σ *with* $\sigma = \mathcal{O}(n^{-1/5})$. *For any* $0 < \delta < 1$, *with probability at least* $1 - \delta$, *there holds*

$$\mathcal{R}(f_{\mathbf{z}}) - \mathcal{R}(f^\star) \lesssim \frac{\log \delta^{-1}}{n^{2/5}}.$$

Proof Since Assumption 2 holds, $f^\star \in \mathcal{H}$, and K_σ is a calibrated modal regression kernel, from Theorem 14 we know that for any $0 < \delta < 1$, with probability at least $1 - \delta$, there holds

$$\mathcal{R}^\sigma(f^\star) - \mathcal{R}^\sigma(f_{\mathbf{z}}) \lesssim \frac{\log \delta^{-1}}{\sqrt{\sigma n}}.$$

When Assumption 1 holds and K_σ is a calibrated modal regression kernel, Theorem 10 yields

$$\left|\{\mathcal{R}(f^\star) - \mathcal{R}(f)\} - \{\mathcal{R}^\sigma(f^\star) - \mathcal{R}^\sigma(f)\}\right| \lesssim \sigma^2.$$

As a result, for any $0 < \delta < 1$, with probability at least $1 - \delta$, there holds

$$\mathcal{R}(f^\star) - \mathcal{R}(f) \lesssim \sigma^2 + \frac{\log \delta^{-1}}{\sqrt{\sigma n}}.$$

With the choice $\sigma = \mathcal{O}(n^{-1/5})$, the proof of Theorem 17 can be accomplished. ■

4.4 Towards the Function Estimation Calibration Problem

We now explore the relation between the modal regression consistency of $f_{\mathbf{z}}$ and its estimation consistency, which is termed as function estimation calibration problem in our study. To this end, we need to impose some further assumptions on the conditional density $p_{\epsilon|X}$.

Definition 18 (Strongly s-Concave Density) *A density p is **strongly s-concave** if it exhibits one of the following forms:*

1. $p = \varphi_+^{1/s}$ *for some strongly concave function φ if $s > 0$, where $\varphi_+ = \max\{\varphi, 0\}$;*
2. $p = \exp(\varphi)$ *for some strongly concave function φ if $s = 0$;*
3. $p = \varphi_+^{1/s}$ *for some strongly convex function φ if $s < 0$.*

Assumption 3 *The density of ϵ conditioned on $\mathcal{X}$, denoted by $p_{\epsilon|X}(\cdot|X)$, satisfies the following conditions:*

1. $\sup_{x\in\mathcal{X}} p_{\epsilon|X}(0|X = x) = c_3$;
2. $p_{\epsilon|X}(t \mid X = x) \leq p_{\epsilon|X}(0|X = x)$, $\forall t \in \mathbb{R}, x \in \mathcal{X}$;
3. $\inf_{t\in[-2M,2M],\, x\in\mathcal{X}} p_{\epsilon|X}(t|X = x) = c_0 > 0$;
4. $p_{\epsilon|X}(\cdot \mid X)$ *denotes strongly s-concave densities for all realizations of X.*

Conditions 1 and 2 in Assumption 3 require that the global mode of the conditional density $p_{\epsilon|X}(\cdot|X)$ for any realization of X in $\mathcal{X}$ is uniquely zero while Condition 3 rules out densities that are not bounded away from below in the vicinity of this unique mode. The two conditions hold for continuous densities with a unique global mode. Condition 4 assumes the strongly s-concave density assumption on $p_{\epsilon|X}$, which is typical from a statistical view as it holds for common symmetric and skewed distributions. Several representative examples are listed below:

Example 1 (Student's t-distribution) *Let ρ be a Student's t-distribution. Its probability density function p is*

$$p(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})} \left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}},$$

where ν is the number of degrees of freedom and Γ is the gamma function. Specifically, when $\nu = 1$, it gives the density function of a typical heavy-tailed distribution, namely, Cauchy distribution; when $\nu = \infty$, it is the density function of a most common probability distribution, i.e., Gauss distribution. One can easily see that for Student's t-distributions, their densities are strongly s-concave that are of the form 3 in Definition 18.

Example 2 (Skewed normal distribution) *Let ρ be a skewed normal distribution with the probability density function*

$$p(t|\mu, \theta, \tau) = \frac{4\tau(1-\tau)}{\sqrt{2\pi\theta^2}} \exp\left\{-\frac{2(x-\mu)^2}{\sigma^2}\left(\tau - \mathsf{I}_{(x\leq\mu)}(x)\right)\right\},$$

where $\mathsf{I}_A(x)$ is the indicator function that takes the value 1 if A is true and 0, otherwise. Clearly, the above density is also strongly s-concave that is of the form 2 in Definition 18.

When Assumption 3 holds, the function estimation convergence can be elicited from the convergence of the modal regression risk, as shown in the following theorem.

Theorem 19 *Suppose that Assumption 3 holds and let $f : \mathcal{X} \to \mathcal{Y}$ be a measurable function in $\mathcal{H}$. Then, there holds*

$$\|f - f^\star\|^2_{L^2_{\rho_\mathcal{X}}} \lesssim \mathcal{R}(f^\star) - \mathcal{R}(f).$$

Proof If Assumption 3 is fulfilled, then $p_{\epsilon|X}$ is strongly s-concave. We verify the desired relation by discussing different cases of s. If $s = 0$, we know that $-\log p_{\epsilon|X}$ is strongly convex for all x. Consequently, in this case, there holds

$$\begin{aligned}\|f - f^\star\|^2_{L^2_{\rho_X}} &\lesssim \int_\mathcal{X} [-\log p_{\epsilon|X}(f(x) - f^\star(x) \mid X = x) + \log p_{\epsilon|X}(0 \mid X = x)] \mathrm{d}\rho_\mathcal{X}(x) \\ &\lesssim \int_\mathcal{X} \left[p_{\epsilon|X}(0 \mid X = x) - p_{\epsilon|X}(f(x) - f^\star(x) \mid X = x)\right] \mathrm{d}\rho_\mathcal{X}(x),\end{aligned}$$

where the last inequality is a consequence of the mean value theorem and Assumption 3. If $s > 0$, $-p^s_{\epsilon|X}$ is strongly convex for all x, then

$$\begin{aligned}\|f - f^\star\|^2_{L^2_{\rho_X}} &\lesssim \int_\mathcal{X} [-p^s_{\epsilon|X}(f(x) - f^\star(x) \mid X = x) + p^s_{\epsilon|X}(0 \mid X = x)] \mathrm{d}\rho_\mathcal{X}(x) \\ &\lesssim \max\{s c_0^{s-1}, s c_3^{s-1}\} \int_\mathcal{X} \left[p_{\epsilon|X}(0 \mid X = x) - p_{\epsilon|X}(f(x) - f^\star(x) \mid X = x)\right] \mathrm{d}\rho_\mathcal{X}(x),\end{aligned}$$

where the second inequality is due to the Lipschitz continuity of $h(t) = t^s$ and Assumption 3. If $s < 0$, $p^s_{\epsilon|X}$ is strongly convex for all x. In this case, there holds

$$\begin{aligned}\|f - f^\star\|^2_{L^2_{\rho_X}} &\lesssim \int_\mathcal{X} [p^s_{\epsilon|X}(f(x) - f^\star(x) \mid X = x) - p^s_{\epsilon|X}(0 \mid X = x)] \mathrm{d}\rho_\mathcal{X}(x) \\ &\lesssim -s c_0^{s-1} \int_\mathcal{X} \left[p_{\epsilon|X}(0 \mid X = x) - p_{\epsilon|X}(f(x) - f^\star(x) \mid X = x)\right] \mathrm{d}\rho_\mathcal{X}(x).\end{aligned}$$

where the second inequality is again due to the Lipschitz continuity of $h(t) = t^s$ and Assumption 3. Recalling the fact that

$$\mathcal{R}(f^\star) - \mathcal{R}(f) = \int_\mathcal{X} \left[p_{\epsilon|X}(0 \mid X = x) - p_{\epsilon|X}(f(x) - f^\star(x) \mid X = x)\right] \mathrm{d}\rho_\mathcal{X}(x),$$

we complete the proof of Theorem 19. ■

Combining the estimates established in the above several subsections, we are now able to answer Question 3 in Subsection 4.1 by deriving the convergence rates from $f_\mathbf{z}$ to $f^\star$.

Theorem 20 *Suppose that Assumptions 1, 2, and 3 hold, and $f^\star \in \mathcal{H}$. Let $f_\mathbf{z}$ be produced by (3.4) which is induced by a calibrated modal regression kernel K_σ with $\sigma = \mathcal{O}(n^{-1/5})$. For any $0 < \delta < 1$, with probability at least $1 - \delta$, there holds*

$$\|f_\mathbf{z} - f^\star\|^2_{L^2_{\rho_\mathcal{X}}} \lesssim \frac{\log \delta^{-1}}{n^{2/5}}.$$

Proof The theorem can be proved by combining estimates in Theorems 17, and 19. ■

4.5 Comparisons with Existing Studies

From the above learning theory analysis, it becomes clear that our study provides a different take on modal regression, i.e., a nonparametric one, but built upon the assumption of the existence of a global conditional mode function. We now go further in comparing our work with the existing studies.

On the one hand, as mentioned in the introduction, the formulated ERM approach and the conducted learning theory analysis distinguish our study with existing work. This is because the formulated ERM approach to modal regression brings us several benefits as listed below:

- Generality: In the ERM approach to modal regression problems, the hypothesis space $\mathcal{H}$ is a function space that can be infinite-dimensional. In practice, it can be specified by applying certain regularization procedures. Moreover, the prevalent kernel-based methods can be naturally integrated since the hypothesis space can be chosen as a subset of a certain reproducing kernel Hilbert space. Compared with the linear modal regression case, see, e.g., Lee (1989, 1993); Yao and Li (2014); Kemp and Santos Silva (2012), more flexibilities may be pursued since the global conditional mode function could be highly nonlinear or even non-smooth;

- Tractability: As mentioned earlier, a large part of existing studies on modal regression were conducted by resorting to maximizing the joint density estimator or the conditional density estimator, see e.g., Collomb et al. (1987); Chen et al. (2016b). However, there are two main barriers in seeking the maximizer. First, from a statistical learning view, learning the maximizer of the joint density or the conditional density is a local type learning problem, in which one has to train the model for each test point. This could be computationally expansive; second, the estimation of the joint density or conditional density is sensitive to the choice of the bandwidth. Without properly chosen bandwidths, the maximization of the joint density or the conditional density could be highly non-convex. That is, the optimization problem might be computationally prohibitive in certain cases, which makes the resulting modal regression estimator unreliable. In contrast, in our study, by following an ERM approach, the related optimization problem can be easily solved by applying iteratively re-weighted least squares algorithms and the scale parameter could be easily tuned via cross-validation;

- Scalability: The estimation of the joint density or the conditional density may suffer from the curse of dimensionality. This is because the dimensionality of the input space $\mathcal{X}$ could be quite high in real-world applications. It is well known that density estimation in high-dimensional space is in general infeasible. As opposed to that, the proposed ERM approach to modal regression in this study only involves a one-dimensional density estimation problem. In this sense, we claim that the ERM approach to modal regression is scalable to high-dimensional data.

On the other hand, it should be mentioned first that our study in this paper is most related to, but essentially different from, the studies in Yao and Li (2014); Kemp and Santos Silva (2012) and Chen et al. (2016b). Some obvious differences include: In Yao and Li (2014); Kemp and Santos Silva (2012), the modal regression problem can also be

viewed as an M-estimation problem, and they assumed that the global conditional mode function is linear. However, in our study, the modal regression estimator is learned from a certain hypothesis space $\mathcal{H}$ which could be infinite-dimensional. Moreover, the main theoretical results established in their work are the asymptotic normality of the resulting coefficient vector, while in our work we are mainly concerned with the generalization ability and function estimation convergence of the modal regression estimator. The recent study in Chen et al. (2016b) investigated the nonparametric modal regression problem under a general setup. The modal regression estimator in their study is obtained by maximizing the joint density. As explained above, this is essentially different from the empirical risk maximization approach in our study. Concerning the differences in the theoretical results, the convergence rates in Chen et al. (2016b) are established in the expectation form with respect to random samples. While in our study, learning theory convergence rates of almost sure type are developed. In fact, it is the statistical learning treatment on modal regression that gives rise to the above differences.

5. On the Robustness of Modal Regression

As mentioned earlier, robustness is regarded as one of the most appealing features of modal regression estimators. The robustness is mentioned in the sense that the estimator is resistant to outliers, and is capable of dealing with skewed conditional distributions, skewed noise distributions, or heavy-tailed noise. This subsection is dedicated to exploring the robustness of the modal regression estimator.

5.1 Some Perspectives

It is reasonable that, compared with mean or median regression, modal regression can do a better job in tackling skewed conditional distributions due to its special emphasis on the conditional mode. It, therefore, stands out from regression estimators in predicting the trends of observations. Discussion in this part will be mainly focused on the robustness of modal regression estimators to heavy-tailed noise, skewed noise, or outliers.

We note that in the above learning theory analysis, some messages related to the robustness of the modal regression estimator have been delivered. For instance, the convergence rates of $\|f_{\mathbf{z}} - f^{\star}\|^2_{L^2_{\rho_{\mathcal{X}}}}$ stated in Theorem 20 are established only under certain mild assumptions on the conditional density of the noise variable (as well as some other mild distribution-free assumptions), without assuming the boundedness or even the moment conditions of the response variable. More detailed speaking, in order to derive the generalization bounds of the estimator, the only assumption imposed on the conditional density is that its peak value is uniformly upper bounded over the input space (see Condition (ii) in Assumption 2). To obtain the modal regression calibration property, we assume that the conditional density is second-order continuously differentiable and is uniformly bounded from above (see Assumption 1). We then further restrict ourselves to some shape-restricted densities (see Assumption 3) to deduce the function estimation calibration problem. Note that, as a consequence of these light assumptions, a large family of probability distributions from light-tailed ones to heavy-tailed ones can be subsumed, with some typical examples of which include the Cauchy noise, the Gaussian noise, and the skewed ones mentioned in

Example 2. These conditions on the noise density relax greatly the usual sub-Gaussianity assumption as well as the moment conditions, which is frequently invoked in the statistical learning literature (see e.g., Cucker and Zhou (2007); Steinwart and Christmann (2008)) to study the performance of learning algorithms. These observations suggest the capability of modal regression in dealing with skewed distributions and heavy-tailed noise.

On the other hand, we recall that the modal regression problem can be easily interpreted from an ERM view as illustrated in Section 3.4. For a modal regression kernel K_σ with the representing function ϕ, this purpose can be achieved by introducing a distance-based loss ϕ_σ as in (3.6). It is obvious that the newly introduced loss is uniformly bounded due to the boundedness requirement on the representing function ϕ. This finding reminds us that the modal regression estimator (3.7) is essentially a re-descending (nonparametric) M-estimator (see Chapter 4.8 in Huber and Ronchetti (2009) for the definition of re-descending M-estimators). Its re-descending property renders it outlier-robust. Compared with other outlier-rejection techniques, the re-descending M-estimators have several advantages, e.g., without suffering from the masking effect that outlier rejection techniques often do and the efficiency enhancement (see e.g., Staudte and Sheather, 2011).

5.2 Quantifying the Robustness

As one of the main themes in robust statistics, quantifying the robustness is also inevitable in studying the modal regression estimator. In the robust statistics literature, various notions for quantifying the robustness have been proposed, e.g., the influence function (Hampel, 1974; Hampel et al., 2011), the breakdown point (Donoho and Huber, 1983; Hampel, 1968), the maxbias curve (Hampel et al., 2011), and the sensitivity curve (Tukey, 1977). We refer the reader to Chapter 10.6 in Steinwart and Christmann (2008) for a quick literature survey.

In the statistical learning literature, the robustness of learning algorithms has also been investigated under different notions, especially in the regularized learning context. For instance, among others, the robustness of weighted LS-SVM was discussed in Suykens et al. (2002) from a weighted learning viewpoint while the robustness of more general kernel-based regression was investigated in De Brabanter et al. (2009); Debruyne et al. (2010). A rigorous and systematic study towards the consistency and robustness of convex kernel-based regression schemes was conducted in Christmann and Steinwart (2007). The robustness was pursued there by imposing restrictions on the shape of the loss function and was measured in terms of the influence function. In a similar vein, Christmann and Messem (2008) and Christmann and Steinwart (2004) studied the robustness properties of support vector machine for regression and classification, respectively. In Debruyne et al. (2008), the robust model selection problem in kernel-based regression was studied with the help of the influence functions. A literature review on the robustness of support vector machines for heavy-tailed distributions was provided in Van Messem and Christmann (2010). The main message from these studies is that for convex risk minimization learning schemes, if the response variable is unbounded, the robustness of the resulting estimator is related to the growth type of the loss function (see, e.g., Christmann and Steinwart (2007) for the definition of the growth type of a loss function) and the tail behavior of the conditional distribution. We also note that, recently, a new robust regression estimator was proposed in Yu et al. (2012) where

certain nonconvex losses were allowed and the robustness was measured by using a relaxed notion of breakdown points.

In our study, noticing the non-convexity and the re-descending property of modal regression schemes, we follow the work in Huber (1984) and adopt the finite sample breakdown point to quantify the robustness of modal regression estimators. The following definition is adapted from Huber (1984).

Definition 21 *Let E be a Banach space and $\mathbf{z} = \{z_1, \cdots, z_n\}$ be an observation set with values in $Z \subset \mathbb{R}^d \times \mathbb{R}$. The finite sample breakdown point of an E-valued statistic $\mathcal{A}_{\mathbf{z}}$ is defined by*

$$\varepsilon_n(\mathcal{A}_{\mathbf{z}}) = \min_{1 \le m \le n} \left\{ \frac{m}{n+m} : \sup_{\mathbf{z}'} \|\mathcal{A}(\mathbf{z} \cup \mathbf{z}')\|_E = \infty \right\},$$

where the supremum is taken over all possible samples $\mathbf{z}' = \{z_{n+1}, \ldots, z_{n+m}\}$ that can take arbitrary values in Z.

"*The finite sample breakdown point is, roughly, the smallest amount of contamination that may cause an estimator to take on arbitrary large aberrant values*" (Donoho and Huber, 1983). In other words, it "*measures the maximum fraction of outliers which a given sample may contain without spoiling the estimate completely*" (Yohai, 1987). As a global robustness measurement, it is useful in measuring the robustness of re-descending M-estimators as shown in Huber (1984). However, in our study, the "parameter space" $\mathcal{H}$ is assumed to be a compact subset of $C(\mathcal{X})$, which limits the use of the finite sample breakdown point. Therefore, in order to explore the robustness of the modal regression estimator in terms of the finite sample breakdown point, we shall allow it to take an arbitrary value instead of assuming its boundedness. A natural way to do this is to consider an automatic model selection procedure, e.g., by applying a regularization technique. Specifically, as described below, we consider a penalized ERM scheme in a reproducing kernel Hilbert space induced by an unbounded Mercer kernel, e.g., the linear kernel, or a polynomial kernel.

To this end, we let ϕ be the representing function of a calibrated modal regression kernel K_σ and additionally assume that $\phi(t) \to 0$ when $t \to \infty$. Let $\mathcal{H}_\mathcal{K}$ be a reproducing kernel Hilbert space induced by an unbounded Mercer kernel $\mathcal{K}$. With a slight abuse of notation, we denote $f_{\mathbf{z},\lambda}$ as the output of the following penalized ERM scheme:

$$f_{\mathbf{z},\lambda} := \arg\max_{f \in \mathcal{H}_\mathcal{K}} \frac{1}{n} \sum_{i=1}^{n} \phi\left(\frac{y_i - f(x_i)}{\sigma}\right) + \lambda \|f\|_\mathcal{K}^2, \tag{5.1}$$

where $\lambda > 0$ is a regularization parameter. Based on these preparations and by adapting the idea of Huber (1984), we come to the following conclusion:

Theorem 22 *Let $f_{\mathbf{z},\lambda}$ be produced by* (5.1) *with $\mathcal{H}_\mathcal{K}$ being induced by an unbounded Mercer kernel $\mathcal{K}$. Denote*

$$A = \sum_{i=1}^{n} \phi(y_i - f_{\mathbf{z},\lambda}(x_i)).$$

Then the finite sample breakdown point of modal regression (5.1) *is*

$$\varepsilon_n(f_{\mathbf{z},\lambda}) = \frac{m^\star}{m^\star + n},$$

where $m^\star$ *is an integer satisfying* $\lceil A \rceil \leq m^\star \leq \lfloor A \rfloor + 1$. *Here,* $\lceil A \rceil$ *denotes the largest integer not greater than* A *and* $\lfloor A \rfloor$ *denotes the smallest integer not less than* A.

Proof Denote m as the size of the additionally added samples $\mathbf{z}'$ that may take arbitrary values. To prove the theorem, we first consider the case when $m < A$. Let $z_i = (x_i, y_i) \in \mathbf{z} \cup \mathbf{z}'$ with $i = 1, \ldots, n+m$. Then, there holds

$$\sum_{i=1}^{n+m} \phi\left(\frac{y_i - f_{\mathbf{z},\lambda}(x_i)}{\sigma}\right) \geq \sum_{j=1}^{n} \phi\left(\frac{y_j - f_{\mathbf{z},\lambda}(x_j)}{\sigma}\right) = A. \tag{5.2}$$

Let $\delta_1 > 0$ be such that $n\delta_1 + m < A$ and assume that $\phi(t/\sigma) \leq \delta_1$ for all $|t| \geq c_1$. We further assume that $f_1 \in \mathcal{H}_\mathcal{K}$ such that $|y_i - f_1(x_i)| \geq c_1$ for all $z_i = (x_i, y_i) \in \mathbf{z}$, $i = 1, \ldots, n$. Recalling that $\phi(u) \leq 1$ for any $u \in \mathbb{R}$, we have

$$\sum_{i=1}^{n+m} \phi\left(\frac{y_i - f_1(x_i)}{\sigma}\right) \leq n\delta_1 + m. \tag{5.3}$$

Therefore, we have

$$\sum_{i=1}^{n+m} \phi\left(\frac{y_i - f_{\mathbf{z}}(x_i)}{\sigma}\right) \geq \sum_{i=1}^{n+m} \phi\left(\frac{y_i - f_1(x_i)}{\sigma}\right). \tag{5.4}$$

Let $f_{\mathbf{z},\mathbf{z}',\lambda}$ be the empirical target function modeled by modal regression (5.1) based on observations $\mathbf{z} \cup \mathbf{z}'$. Inequality (5.4) implies that $|y - f_{\mathbf{z},\mathbf{z}',\lambda}(x)| < c_1$ for at least one point in $\mathbf{z}$. Therefore, $\|f_{\mathbf{z},\mathbf{z}',\lambda}\|_\infty$ stays bounded no matter how $\mathbf{z}'$ varies.

On the other hand, if $m > A$, let $\delta_2 > 0$ be such that $m - m\delta_2 > A$ and assume that $\phi(t/\sigma) \leq \delta_2$ for all $|t| \geq c_2$. We further assume that all points in $\mathbf{z}'$, i.e., $z_j = (x_j, y_j) \in \mathbf{z}'$, $j = n+1, \ldots, n+m$, are the same and satisfy $y_j = \hat{f}(x_j)$, $j = n+1, \ldots, n+m$, for some $\hat{f} \in \mathcal{H}_\mathcal{K}$. Suppose that $f_2 \in \mathcal{H}_\mathcal{K}$ such that $|y_{n+1} - f_2(x_{n+1})| \geq c_2$. Then, there holds

$$\sum_{i=1}^{n+m} \phi\left(\frac{y_i - f_2(x_i)}{\sigma}\right) \leq \sum_{i=1}^{n} \phi\left(\frac{y_i - f_{\mathbf{z}}(x_i)}{\sigma}\right) + \sum_{i=n+1}^{n+m} \phi\left(\frac{y_i - f_2(x_i)}{\sigma}\right) \leq A + m\delta_2. \tag{5.5}$$

Moreover, due to the assumption that $y_j = \hat{f}(x_j)$, $j = n+1, \ldots, n+m$, we have

$$\sum_{i=1}^{n+m} \phi\left(\frac{y_i - \hat{f}(x_i)}{\sigma}\right) \geq m. \tag{5.6}$$

From inequalities (5.5) and (5.6), we have

$$\sum_{i=1}^{n+m} \phi\left(\frac{y_i - f_2(x_i)}{\sigma}\right) \leq \sum_{i=1}^{n+m} \phi\left(\frac{y_i - \hat{f}(x_i)}{\sigma}\right).$$

Therefore, we have $|y_{n+1} - f_{\mathbf{z},\mathbf{z}',\lambda}(x_{n+1})| < c_2$. Varying $\mathbf{z}'$ by letting $y_{n+1} \to \infty$ while fixing x_{n+1} yields that $f_{\mathbf{z},\mathbf{z}',\lambda} \to \infty$ and thus leads to breakdown.

Following the above discussions and recalling the definition of the finite sample breakdown point, we know that the finite sample breakdown point is either $\lceil A \rceil$ or $\lfloor A \rfloor + 1$. This completes the proof of Theorem 22. ■

Some direct implications of Theorem 22 include: first, similar to the re-descending M-estimator in Huber (1984), the finite sample breakdown point of the modal regression estimator (5.1) depends on the modal regression kernel K_σ, the bandwidth σ, and the sample configuration; second, if the bandwidth σ is determined solely by the sample, the findings in Huber (1984) remind us that the finite sample breakdown point in this case is quite high and even typically exceeds 0.49. This is a "good news" situation for the robustness of the modal regression estimator since, as we shall see in the experimental section, the bandwidth parameter can be effectively chosen by applying data-driven techniques, e.g., a k-fold cross-validation approach; third, as shown below, we computationally track the modal regression estimator by iteratively solving a re-weighted least squares problem. Again according to Huber (1984), for the usual estimators with reasonable tunning, the finite sample breakdown point is usually above 0.4. Therefore, starting with an estimator that has a high finite sample breakdown point, the resulting estimator also has a high finite sample breakdown point. We will illustrate this empirically later in the experimental section.

6. A General Picture of Correntropy based Regression

Thus far, having realized that MCCR with a tending-to-zero scale parameter corresponds to modal regression, a learning theory study towards modal regression is conducted in the preceding sections, and the robustness of the modal regression estimator is also discussed. Based on these preparations and the study in Feng et al. (2015), we are now able to depict a general picture of the correntropy based regression from a statistical learning view. To this end, we exposit the correntropy based regression by considering three different cases below, namely, (1): $\sigma = \sigma(n) \to \infty$; (2): $\sigma := \sigma_0$ for some $\sigma_0 > 0$, that is, σ is fixed and independent of the sample size n; (3): $\sigma := \sigma(n) \to 0$. Before proceeding, we recall the following data-generating model

$$Y = f^\star(X) + \epsilon.$$

We first consider the case when $\sigma(n) \to \infty$. Under the zero-mean noise assumption on ϵ, i.e., $\mathbb{E}(\epsilon|X) = 0$, as revisited in Section 2.2, MCCR (2.1) with $\sigma(n) \to \infty$ encourages the approximation of $f_{\mathbf{z}}$ towards the conditional mean function $\mathbb{E}(Y|X)$ and the scale parameter σ in this case plays a trade-off role between robustness and generalization. More explicitly, in this case, the correntropy-induced loss is a calibrated least squares loss in the sense of the following theorem, see also Lemma 7 in Feng et al. (2015):

Theorem 23 (Lemma 7, Feng et al. (2015)) *Assume that $\mathbb{E}Y^4 < \infty$ and denote $f^\star = \mathbb{E}(Y|X)$. For any $f \in \mathcal{H}$, there holds*

$$\left| \|f - f^\star\|^2_{L^2_{\rho_\mathcal{X}}} - |\sigma(\mathcal{R}^\sigma(f) - \mathcal{R}^\sigma(f^\star))| \right| \lesssim \sigma^{-2}.$$

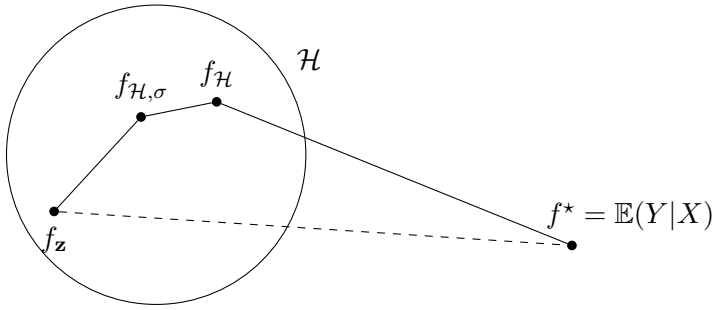

Figure 2: A schematic illustration of the mechanism in correntropy-based regression when $\sigma(n) \to \infty$ and the noise variable ϵ is assumed to be zero-mean. $f_{\mathcal{H},\sigma}$ is the data-free counterpart of $f_{\mathbf{z}}$, $f_{\mathcal{H}}$ is the data-free least squares regression estimator, and $f^\star$ is the conditional mean function.

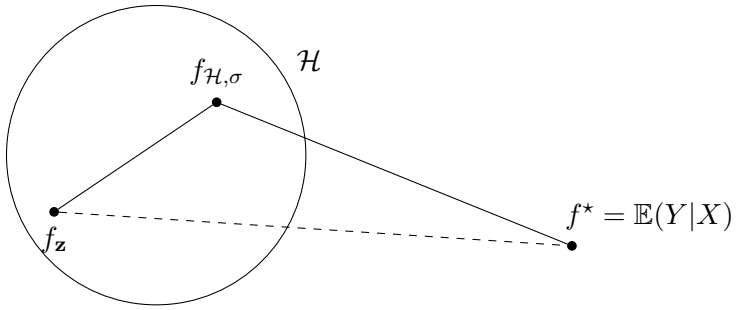

Figure 3: A schematic illustration of the mechanism in correntropy-based regression when σ is fixed and independent on n and the noise variable ϵ is assumed to be zero-mean. $f_{\mathcal{H},\sigma}$ is the data-free counterpart of $f_{\mathbf{z}}$, and $f^\star$ is the conditional mean function.

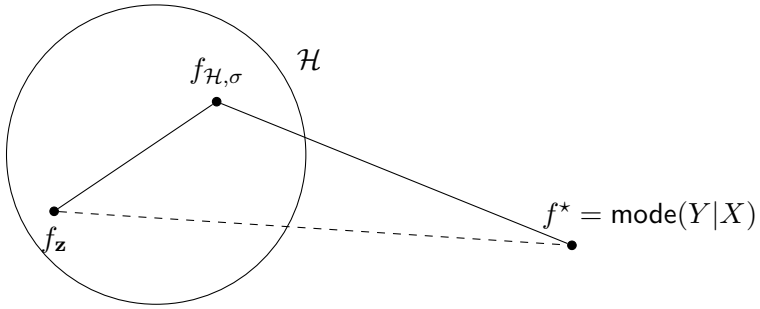

Figure 4: A schematic illustration of the mechanism in correntropy-based regression when $\sigma(n) \to 0$ and the noise variable ϵ is assumed to admit a unique global zero-mode. $f_{\mathcal{H},\sigma}$ is the data-free counterpart of $f_{\mathbf{z}}$, and $f^\star$ is the conditional mode function.

Clearly, according to the above theorem, when $\sigma(n)$ is properly chosen with $\sigma(n) \to \infty$, the consistency of $\mathcal{R}^\sigma(f_{\mathbf{z}}) - \mathcal{R}^\sigma(f^\star)$ implies the consistency of $\|f_{\mathbf{z}} - f^\star\|^2_{L^2_{\rho_\mathcal{X}}}$. In this sense, MCCR with $\sigma(n) \to \infty$ is a calibrated least squares regression. More specifically, the following theorem is established in Feng et al. (2015):

Theorem 24 *Assume that $f^\star = \mathbb{E}(Y|X) \in \mathcal{H}$ and $\mathbb{E}Y^4 < \infty$. Under a mild capacity assumption on $\mathcal{H}$, for any $0 < \delta < 1$, with confidence $1 - \delta$, there holds*

$$\|f_{\mathbf{z}} - f^\star\|^2_{L^2_{\rho_\mathcal{X}}} \lesssim \sigma^{-2} + \sigma m^{-1/(1+p)},$$

where the index $p > 0$ reflects the capacity of the hypothesis space $\mathcal{H}$.

Obviously, when σ is chosen as $\sigma := n^{-1/(3+3p)}$, the convergence rates for $\|f_{\mathbf{z}} - f^\star\|^2_{L^2_{\rho_\mathcal{X}}}$ of the type $\mathcal{O}\big(n^{-2/(3+3p)}\big)$ can be established. The robustness in the correntropy based regression can be revealed by the moment assumption $\mathbb{E}Y^4 < \infty$. Noticing that in this case the underlying truth $f^\star$ corresponds to the conditional mean, the correntropy based regression is also a robustified least squares regression. A schematic illustration of the mechanism in correntropy based regression in this case is given in Fig. 2, in which $f_{\mathcal{H},\sigma}$ is the population version of $f_{\mathbf{z}}$ and $f_\mathcal{H}$ is the data-free least squares regression estimator. As argued in Feng et al. (2015), compared with the least squares regression, an additional bias, i.e., the distance between $f_{\mathcal{H},\sigma}$ and $f_\mathcal{H}$, appears when bounding the $L^2_{\rho_\mathcal{X}}$-distance between $f_{\mathbf{z}}$ and the conditional mean function $\mathbb{E}(Y|X)$. Moreover, this bias in some sense reflects the trade-off between the convergence rates of $\|f_{\mathbf{z}} - f^\star\|^2_{L^2_{\rho_\mathcal{X}}}$ and the robustness of $f_{\mathbf{z}}$ in correntropy based regression.

The case when $\sigma = \sigma_0$, i.e., σ is fixed and independent of n, is simpler. In this case, the correntropy based regression corresponds to a special case of the least squares regression. According to Lemma 18 in Feng et al. (2015), it is also least squares regression calibrated. However, the convergence rates of $\|f_{\mathbf{z}} - f^\star\|^2_{L^2_{\rho_\mathcal{X}}}$ in this case may be established only under stringent conditions on the noise variable, e.g., the conditional density of the noise variable conditioned on X is symmetric and uniformly bounded for all realizations of X, see Theorem 6 in Feng et al. (2015). In this sense, the correntropy based regression with a fixed σ value in general loses the robustness property. On the other hand, bounding the $L^2_{\rho_\mathcal{X}}$-distance between $f_{\mathbf{z}}$ and $f^\star$ in this case can be typically done by applying the classical bias-variance decomposition as shown in Fig. 3.

The fact that MCCR can be cast as a modal regression problem when $\sigma(n) \to 0$ switches our attention from robust mean regression in Feng et al. (2015) to modal regression in this study. In retrospect, the modal regression scheme (3.4) with the Gaussian kernel as the modal regression kernel retrieves MCCR (2.1). From the arguments in the preceding sections, we know that, under the assumption that the noise variable admits a unique global zero-mode, MCCR (2.1) with $\sigma(n) \to 0$ is modal regression calibrated. That is, under proper assumptions as listed in Theorem 20, one may expect the learning theory type convergence from the MCCR estimator to the modal regression function $f_\mathrm{M} = \mathsf{mode}(Y|X)$. Our study in the above sections reveals that the modal regression problem can be also studied from an empirical risk minimization view. Therefore, a schematic illustration of the mechanism in correntropy-based regression when $\sigma(n) \to 0$ can be also presented as in Fig. 4. As explained

earlier, in this case, the robustness of MCCR comes from the re-descending property of the related loss and the built-in robustness of modal regression estimators.

	$\sigma(n) \to \infty$	σ fixed	$\sigma(n) \to 0$
resulting estimator	robustified conditional mean estimator	conditional mean estimator	conditional mode estimator
target function	$\mathbb{E}(Y\|X)$	$\mathbb{E}(Y\|X)$	$\mathsf{mode}(Y\|X)$
noise condition	moment condition	bounded and symmetric noise	possibly skewed and (or) heavy-tailed noise
established rates	$\mathcal{O}(n^{-2/(3+3p)})$	$\mathcal{O}(n^{-2/(2+p)})$	$\mathcal{O}(n^{-2/5})$

Table 2: An Overview of the Three Scenarios in Correntropy based Regression

An overview of the above-discussed three scenarios in correntropy based regression is summarized in Table 2. To sum up, in short, what makes MCCR so special is that it results an interesting walk between modal regression and robustified least squares regression by adjusting the scale parameter σ in correspondence to the sample size n. This will be verified empirically in the next section.

7. Empirical Assessments

This section is concerned with the implementation issues and empirical assessments of the modal regression estimator (MR in short throughout this section). We carry out empirical comparisons between MR and two representative robust regression estimators, namely, the estimator based on the Huber's loss (Huber in short) which approximates the conditional mean, and the estimator based on the least absolute deviation loss (LAD in short) which approximates the conditional median.

7.1 Experimental Setup

In our empirical studies, instead of considering a fixed and bounded hypothesis space $\mathcal{H}$, we consider the following regularization scheme in the reproducing kernel Hilbert space $\mathcal{H}_\mathcal{K}$ that is induced by a Mercer kernel $\mathcal{K}$:

$$f_\mathbf{z} := \arg \min_{f \in \mathcal{H}_\mathcal{K} \bigoplus \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(y_i - f(x_i)) + \lambda \|f\|_\mathcal{K}^2, \tag{7.1}$$

where $\lambda > 0$ is a tuning parameter that controls the complexity of the hypothesis space. The representer theorem ensures that $f_\mathbf{z}$ can be modeled by

$$f_\mathbf{z}(x) = \sum_{i=1}^{n} \alpha_{\mathbf{z},i} \mathcal{K}(x, x_i) + b_\mathbf{z},\ x \in \mathbb{R}.$$

For the Mercer kernel $\mathcal{K}$, we employ the Gaussian kernel $\mathcal{K}(x, x') = \exp\big(-\|x - x'\|^2/h^2\big)$ with the tuning parameter $h > 0$. We consider three different choices for the loss function

$\mathcal{L}$, namely, the correntropy-induced loss, Huber's loss, and the least absolute deviation loss. Explicit forms of the three loss functions are given as follows:

- Correntropy-induced loss: $\mathcal{L}(t) = \sigma^2(1 - \exp(-t^2/\sigma^2))$, $t \in \mathbb{R}$;

- Huber's loss:

$$\mathcal{L}(t) = \begin{cases} t^2/2, & \text{if } |t| \leq \sigma, t \in \mathbb{R}; \\ \sigma|t| - \sigma^2/2, & \text{if } |t| > \sigma, t \in \mathbb{R}; \end{cases}$$

- LAD loss: $\mathcal{L}(t) = |t|$, $t \in \mathbb{R}$.

Note that there is an additional tunning parameter σ in the correntropy-induced loss and Huber's loss.

7.2 Algorithms and Tuning Parameters Selection

We apply the iteratively re-weighted least squares algorithm to solve the regularized regression problem (7.1) when the loss function is chosen as the correntropy-induced loss, Huber's loss, or the LAD loss. The pseudo-code of the iteratively re-weighted least squares algorithm is list as in Algorithm 1. For each iteration in Algorithm 1, the weight is updated as follows:

$$\omega_i^{k+1} = \frac{|\nabla\mathcal{L}(y_i - \boldsymbol{\mathcal{K}}_i^\top \boldsymbol{\alpha}^k - b^k)|}{|y_i - \boldsymbol{\mathcal{K}}_i^\top \boldsymbol{\alpha}^k - b^k|}, \ i = 1, \ldots, n, \tag{7.2}$$

with the initial guess $\boldsymbol{\alpha}^0$, b^0 being zero. Note that for the non-differentiable LAD case, in each iteration the weight is updated as

$$\omega_i^{k+1} = \max\{\delta, |y_i - \boldsymbol{\mathcal{K}}_i^\top \boldsymbol{\alpha}^k - b^k|\}^{-1}, \ i = 1, \ldots, n, \tag{7.3}$$

where δ is a fixed small positive constant, e.g., $\delta = 10^{-4}$.

Concerning the tuning parameters selection, there are three tuning parameters in Huber and MR, i.e., the regularization parameter λ, the Gaussian kernel bandwidth h, and the scale parameter σ in the loss function. For LAD, there are two tuning parameters, i.e., the regularization parameter λ, and the Gaussian kernel bandwidth h. In our experiments, all of the above parameters are selected via a five-fold cross-validation.

Algorithm 1: Iteratively Re-weighted Least Squares Algorithm for Solving (7.1)

Input: data $\{(x_i, y_i)\}_{i=1}^n$, regularization parameter $\lambda > 0$, Gaussian kernel bandwidth $h > 0$, scale parameter $\sigma > 0$ and the initial guess $\boldsymbol{\alpha}^0 \in \mathbb{R}^n$, $b^0 \in \mathbb{R}$.

Output: the learned coefficient $\boldsymbol{\alpha}^{k+1} = (\alpha_1^{k+1}, \ldots, \alpha_n^{k+1})^\top$ and $b^{k+1} \in \mathbb{R}$.

while the stopping criterion is not satisfied **do**

- Compute $\boldsymbol{\alpha}^{k+1}$ and b^{k+1} by solving the following weighted least squares problem:

$$(\boldsymbol{\alpha}^{k+1}, b^{k+1}) = \arg \min_{\boldsymbol{\alpha} \in \mathbb{R}^n, b \in \mathbb{R}} \sum_{i=1}^n \omega_i^{k+1} (y_i - \boldsymbol{\mathcal{K}}_i^\top \boldsymbol{\alpha} - b)^2 + \lambda \boldsymbol{\alpha}^\top \boldsymbol{\mathcal{K}} \boldsymbol{\alpha},$$

where ω_i^{k+1} is specified in (7.2) or (7.3).

- Set $k := k + 1$.

end while

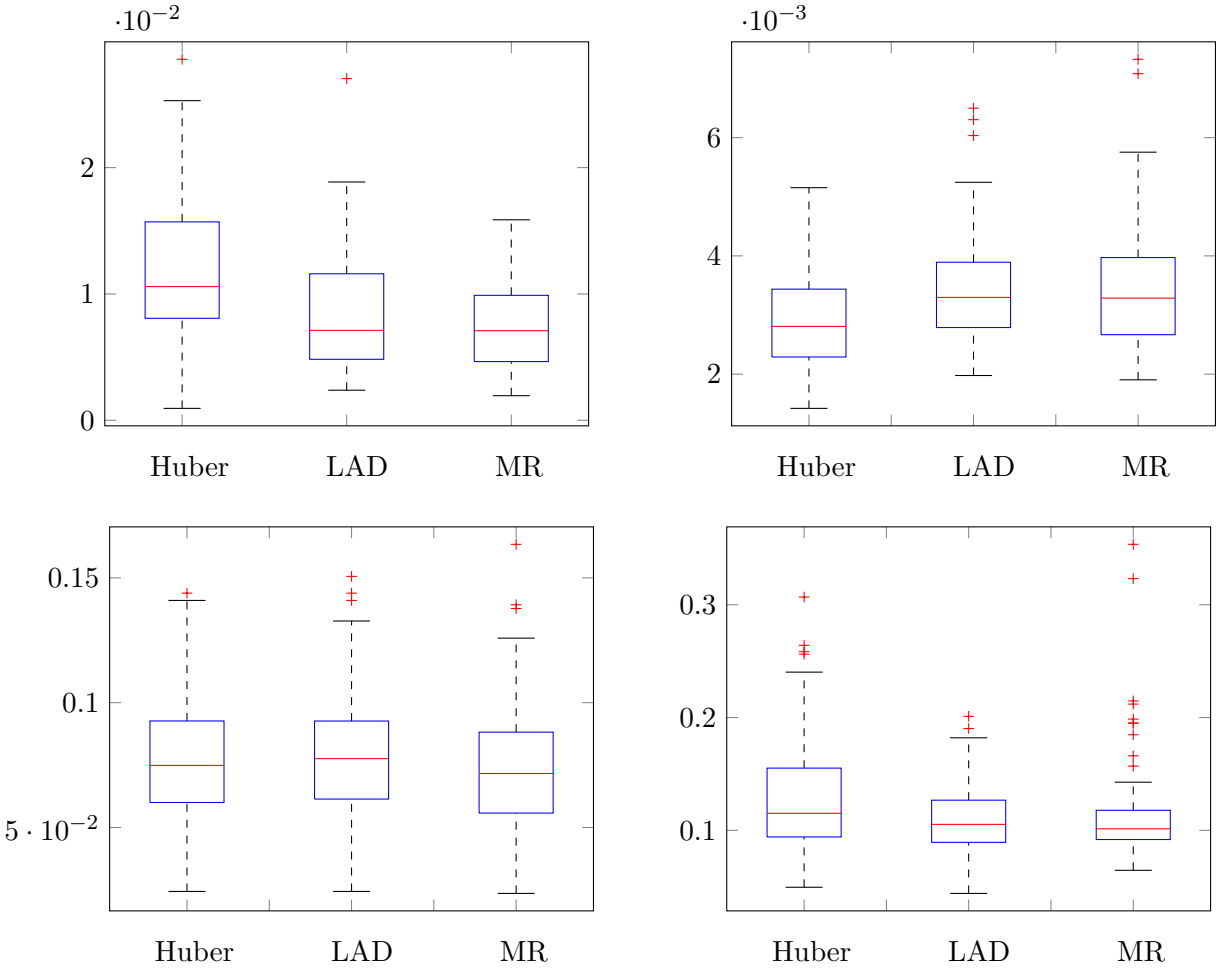

Figure 5: Box-plots of MSEs of the three regression estimators with noises from different distributions. Top left panel: MSEs of regression estimators for (7.4). Top right panel: MSEs of regression estimators for (7.5) and Case a. Bottom left panel: MSEs of regression estimators for (7.5) and Case b. Bottom right panel: MSEs of regression estimators for (7.5) and Case c.

model	Huber	LAD	MR
(7.4)	0.0117 ± 0.0056	0.0084 ± 0.0044	0.0075 ± 0.0034
(7.5)a	0.0029 ± 0.0003	0.0034 ± 0.0004	0.0034 ± 0.0002
(7.5)b	0.0773 ± 0.0006	0.0783 ± 0.0006	0.0734 ± 0.0006
(7.5)c	0.1323 ± 0.0028	0.1090 ± 0.0007	0.1136 ± 0.0019

Table 3: MSEs of the three regression estimators for models in Section 7.3

7.3 Evaluation on Toy Examples

For the first toy example, we consider the following data-generating model employed in Yao and Li (2014):

$$y = f^{\star}(x) + \kappa(x)\epsilon, \tag{7.4}$$

where $x \sim U(0,1)$, $f^{\star}(x) = 2\sin(\pi x)$, and $\kappa(x) = 1 + 2x$. The noise variable is distributed as $\epsilon \sim 0.5N(-1, 2.5^2) + 0.5N(1, 0.5^2)$. As calculated in Yao and Li (2014), the conditional mean function is

$$f_{\mathsf{ME}} = 2\sin(\pi x),$$

the conditional median function is approximately

$$f_{\mathsf{MD}} = 2\sin(\pi x) + 0.67 + 1.34x,$$

and the conditional mode function is approximately

$$f_{\mathsf{MO}} = 2\sin(\pi x) + 1 + 2x.$$

In our experiment, 200 observations are drawn from the above data-generating model and the size of the test set is also set to 200. The reconstructed curve is plotted at the test points in Fig. 6, in which the conditional mean function f_{ME}, the conditional median function f_{MD}, and the conditional mode function f_{MO} are also plotted. During the training process, we never refer to any information of the truth function, e.g., f_{ME}, f_{MD}, or f_{MO}. In our experiment, we adjust the bandwidth of the Gaussian kernel h and the regularization parameter λ via a five-fold cross-validation. For the scale parameter σ in the loss function, we set $\sigma = 0.05$ in the top panel of Fig. 6 and $\sigma = 10$ in the bottom panel of Fig. 6. In the two panels, the dotted blue curves with $\otimes$ marks are learned regression estimators from noisy observations.

From Fig. 6, it is easy to see that with a larger σ value, the resulting regression estimator tends to approximate the conditional mean function, while with a smaller σ value, the conditional mode function is approximated. This empirically justifies our theoretical observation that MCCR "results an interesting walk between modal regression and robustified least squares regression by adjusting the scale parameter σ". Moreover, from the top panel of Fig. 6, we can see that the learned regression estimator $f_{\mathbf{z}}$ approximates the conditional mode function f_{MO} well.

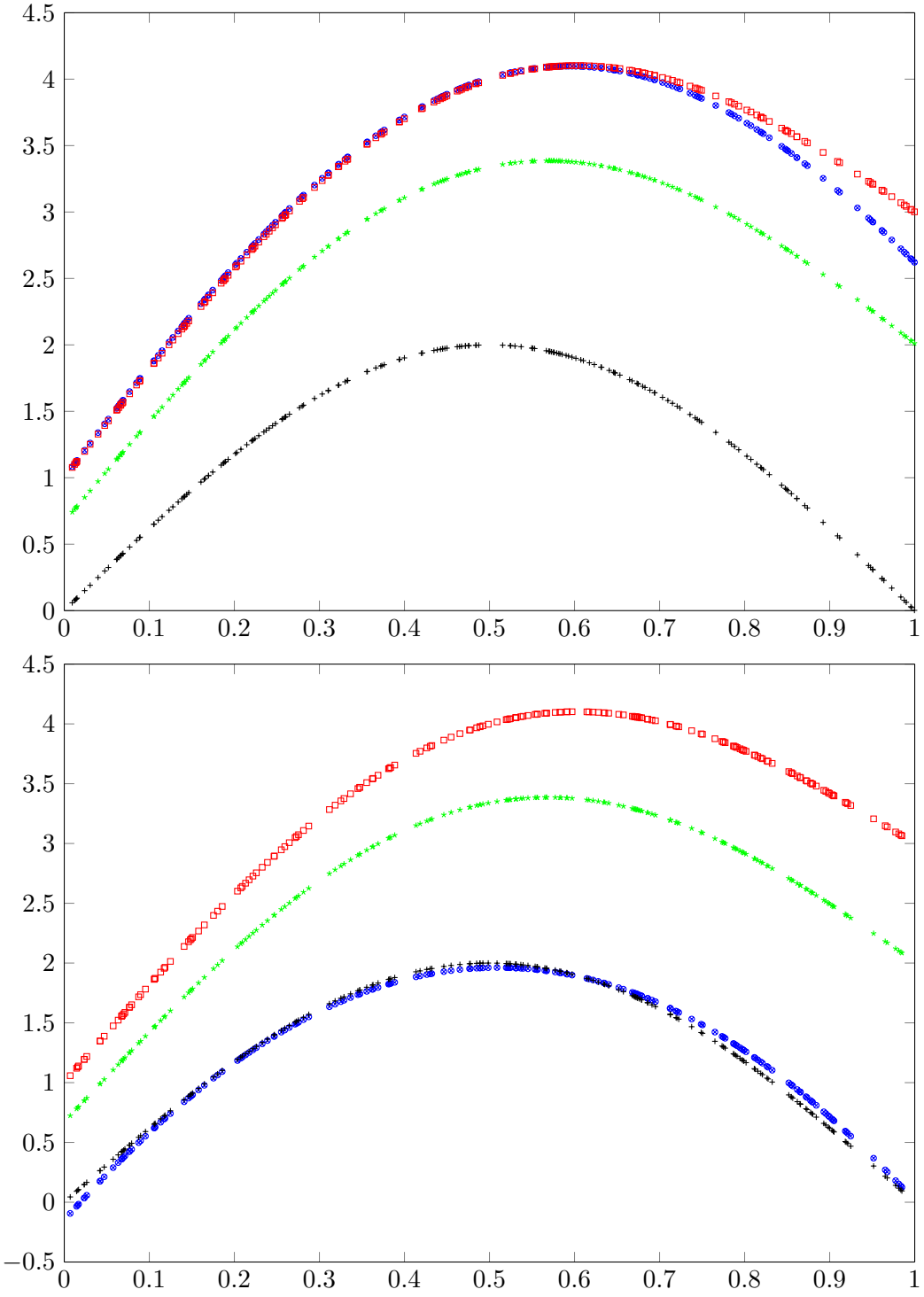

Figure 6: In the above two panels, the dotted red curve with square marks is the conditional mode function f_{MO} for observations generated by (7.4). The dotted green curve with star marks stands for their conditional median function f_{MD}. The dotted black curve with plus marks gives the conditional mean function f_{ME} for these observations. The dotted blue curve with $\otimes$ marks represents the learned estimator $f_{\boldsymbol{z}}$ from these noisy observations.

As a second toy example, we choose the *sinc* function as the truth function. That is, we consider the following data-generating model

$$y = f^\star(x) + \epsilon, \tag{7.5}$$

where $f^\star(x) = \sin(\pi x)/(\pi x)$ and x is drawn uniformly from $[-4, 4]$. Here we consider the following several different noise distributions:

a. $\epsilon \sim N(0, 1)$, i.e., Gaussian noise;

b. $\epsilon \sim t_3/\sqrt{3}$, Student's t-distribution with 3 degrees of freedom, i.e., heavy-tailed noise;

c. $\epsilon \sim 0.95N(0, 1) + 0.05N(0, 3.5^2)$, i.e., outliers and Gaussian noise.

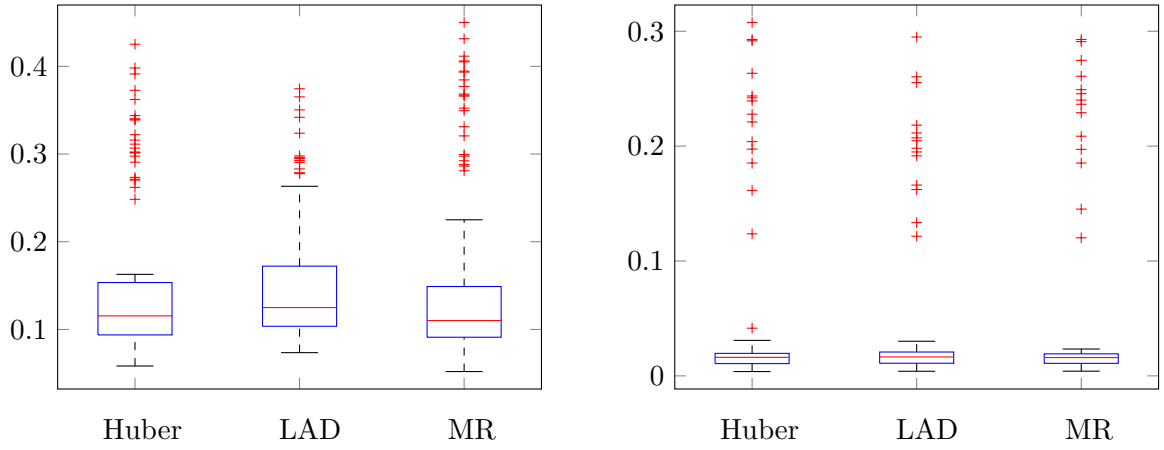

Figure 7: Box-plots of MSEs and MAEs for regression estimators for Education Expenditure Data. Left panel: MAEs for regression estimators for Education Expenditure Data. Right panel: MSEs for regression estimators with the Education Expenditure Data.

The mean squared errors (MSEs) for the above four regression estimators are reported in Table 3 and their box-plots are given in Fig. 5. From Table 3 and Fig. 5, we see that the modal regression estimator MR performs comparably with Huber and LAD in the presence of Gaussian noise. It can outperform Huber and LAD in the presence of skewed noise, outliers, or heavy-tailed noise. Moreover, from Fig. 5, we also observe that the interquartile range is usually smaller for MR than those for Huber and LAD. This in some sense may indicate the smaller confidence interval of MR, the measure of which was employed to study the effectiveness of modal regression estimators in Yao and Li (2014).

7.4 Application to Education Expenditure Data

The Education Expenditure Data, proposed in Chatterjee and Hadi (2015), deals with the education expenditure for fifty U.S. states. It was used in Rousseeuw and Leroy (2005) and Chatterjee and Mächler (1997) to illustrate robust fitting problems. Recently, it was employed in Yao et al. (2012) to demonstrate the robustness of modal regression in the presence of outliers.

The Education Expenditure Data which is outlier-contaminated contains 50 observations. In our experiment, we first preprocess the data by first scaling each feature to $[0, 1]$. Then, 40 observations are randomly chosen for training and the rest are used for testing. The experiment is repeated 100 times. The MSEs and the mean absolute errors (MAEs) are recorded in Table 4 and boxplotted in Fig. 7.

	Huber	LAD	MR
MSEs	0.0445 ± 0.0772	0.0408 ± 0.0676	0.0437 ± 0.0767
MAEs	0.1436 ± 0.0893	0.1424 ± 0.0855	0.1397 ± 0.0883

Table 4: MSEs and MAEs of regression estimators for Education Expenditure Data

From the above reported experimental results, we see the outlier-robustness of MR. Moreover, it can also perform at least comparably on well-known UCI regression datasets. We refer the reader to Section 7 in Feng et al. (2015) for more detailed numerical comparisons.

8. Conclusions

By conducting a statistical learning treatment, in this paper we investigated and explored modal regression problem which has not been well studied in the statistical learning literature. By adapting the empirical risk minimization framework to modal regression, we devised a setup for formulating the nonparametric modal regression problem and developed a learning theory. Based on these efforts, we gained some insights into nonparametric modal regression. These insights include: first, modal regression problem can be solved in the empirical risk minimization framework and can be also interpreted from a kernel density estimation view; second, learning for modal regression is generalization consistent and modal regression calibrated in the sense defined in our study; third, function estimation consistency and convergence in the sense of the $L^2_{\rho_{\mathcal{X}}}$-distance can be derived in modal regression; last but not least, modal regression is robust to outliers, heavy-tailed noise, skewed noise, and skewed conditional distributions. These insights in return unveil the working mechanism of MCCR when its scale parameter tends to zero as in this case it corresponds to a modal regression problem.

Appendix A. Definitions

Definition 25 (ℓ^2-empirical Covering Number) *Let $\mathcal{F}$ be a set of functions on $\mathcal{X}$ and $\mathbf{x} = \{x_1, \cdots, x_m\} \subset \mathcal{X}$. The metric $d_{2,\mathbf{z}}$ is defined on $\mathcal{F}$ by*

$$d_{2,\mathbf{x}}(f, g) = \left\{ \frac{1}{m} \sum_{i=1}^{m} (f(x_i) - g(x_i))^2 \right\}^{1/2}.$$

For every $\varepsilon > 0$, the ℓ^2***-empirical covering number*** *of $\mathcal{F}$ with respect to $d_{2,\mathbf{x}}$ is defined as*

$$\mathcal{N}_{2,\mathbf{x}}(\mathcal{F}, \varepsilon) = \inf \left\{ \ell \in \mathbb{N} : \exists \{f_i\}_{i=1}^{\ell} \text{ such that } \mathcal{F} = \cup_{i=1}^{\ell} \{f \in \mathcal{F} : d_{2,\mathbf{x}}(f, f_i) \leq \varepsilon\} \right\}.$$

Acknowledgments

Y.F. thanks Dr. Eryan Hu and Dr. Yuning Yang for helpful discussions. The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013) / ERC AdG A-DATADRIVE-B (290923). This paper reflects only the authors' views, the Union is not liable for any use that may be made of the contained information. Research Council KUL: GOA/10/09 MaNet, CoE PFV/10/002 (OPTEC), BIL12/11T; PhD/Postdoc grants. Flemish Government: FWO: projects: G.0377.12 (Structured systems), G.088114N (Tensor based data similarity); PhD/Postdoc grants. IWT: projects: SBO POM (100031); PhD/Postdoc grants. iMinds Medical Information Technologies SBO 2014. Belgian Federal Science Policy Office: IUAP P7/19 (DYSCO, Dynamical systems, control and optimization, 2012-2017). The research of Jun Fan was supported in part by NSF FRG Grant DMS-1265202, and NIH Grant 1-U54AI117924-01. The corresponding author is Jun Fan.

References

Markus Baldauf and Joao Santos Silva. On the use of robust regression in econometrics. *Economics Letters*, 114(1):124–127, 2012.

Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local Rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005.

Peter L. Bartlett, Michael I. Jordan, and Jon D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

Ricardo J. Bessa, Vladimiro Miranda, and Joao Gama. Entropy and correntropy against minimum square error in offline and online three-day ahead wind power forecasting. *IEEE Transactions on Power Systems*, 24(4):1657–1666, 2009.

Samprit Chatterjee and Ali S. Hadi. *Regression Analysis by Example*. John Wiley & Sons, 2015.

Samprit Chatterjee and Martin Mächler. Robust regression: a weighted least squares approach. *Communications in Statistics-Theory and Methods*, 26(6):1381–1394, 1997.

Badong Chen, Lei Xing, Haiquan Zhao, Nanning Zheng, and José C. Principe. Generalized correntropy for robust adaptive filtering. *IEEE Transactions on Signal Processing*, 64 (13):3376–3387, 2016a.

Yen-Chi Chen, Christopher R. Genovese, Ryan J. Tibshirani, and Larry Wasserman. Nonparametric modal regression. *The Annals of Statistics*, 44(2):489–514, 2016b.

Herman Chernoff. Estimation of the mode. *Annals of the Institute of Statistical Mathematics*, 16(1):31–41, 1964.

Andreas Christmann and Arnout Van Messem. Bouligand derivatives and robustness of support vector machines for regression. *Journal of Machine Learning Research*, 9:915–936, 2008.

Andreas Christmann and Ingo Steinwart. On robustness properties of convex risk minimization methods for pattern recognition. *Journal of Machine Learning Research*, 5: 1007–1034, 2004.

Andreas Christmann and Ingo Steinwart. Consistency and robustness of kernel-based regression in convex risk minimization. *Bernoulli*, 13(3):799–819, 2007.

Grard Collomb, Wolfgang Härdle, and S. Hassani. A note on prediction via estimation of the conditional mode function. *Journal of Statistical Planning and Inference*, 15(2): 227–236, 1987.

Dorin Comaniciu and Peter Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):603–619, 2002.

Felipe Cucker and Ding-Xuan Zhou. *Learning Theory: An Approximation Theory Viewpoint*. Cambridge University Press, 2007.

Sanjoy Dasgupta and Samory Kpotufe. Optimal rates for k-NN density and mode estimation. In *Advances in Neural Information Processing Systems*, pages 2555–2563, 2014.

Kris De Brabanter, Kristiaan Pelckmans, Jos De Brabanter, Michiel Debruyne, Johan A.K. Suykens, Mia Hubert, and Bart De Moor. Robustness of kernel based regression: a comparison of iterative weighting schemes. In *Artificial Neural Networks–ICANN 2009*, pages 100–110. Springer, 2009.

Michiel Debruyne, Mia Hubert, and Johan A.K. Suykens. Model selection in kernel based regression using the influence function. *Journal of Machine Learning Research*, 9:2377–2400, 2008.

Michiel Debruyne, Andreas Christmann, Mia Hubert, and Johan A.K. Suykens. Robustness of reweighted least squares kernel based regression. *Journal of Multivariate Analysis*, 101 (2):447–463, 2010.

David L. Donoho and Peter J. Huber. The notion of breakdown point. In *A Festschrift for Erich L. Lehmann*, pages 157–184. Wadsworth, Belmont, CA, 1983.

William F. Eddy. Optimum kernel estimators of the mode. *The Annals of Statistics*, 8(4): 870–882, 1980.

Jochen Einbeck and Gerhard Tutz. Modelling beyond regression functions: an application of multimodal regression to speed–flow data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 55(4):461–475, 2006.

Jun Fan, Ting Hu, Qiang Wu, and Ding-Xuan Zhou. Consistency analysis of an empirical minimum error entropy algorithm. *Applied and Computational Harmonic Analysis*, 41 (1):164–189, 2016.

Yunlong Feng, Xiaolin Huang, Lei Shi, Yuning Yang, and Johan A.K. Suykens. Learning with the maximum correntropy criterion induced losses for regression. *Journal of Machine Learning Research*, 16:993–1034, 2015.

Frédéric Ferraty, Ali Laksaci, and Philippe Vieu. Functional time series prediction via conditional mode estimation. *Comptes Rendus Mathematique*, 340(5):389–392, 2005.

Keinosuke Fukunaga and Larry Hostetler. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on Information Theory*, 21(1):32–40, 1975.

Ali Gannoun, Jerome Saracco, and Keming Yu. On semiparametric mode regression estimation. *Communications in Statistics - Theory and Methods*, 39(7):1141–1157, 2010.

László Györfi, Michael Kohler, Adam Krzyzak, and Harro Walk. *A Distribution-free Theory of Nonparametric Regression*. Springer Science & Business Media, 2006.

Frank R. Hampel. *Contributions to the theory of robust estimation*. PhD thesis, University of California Berkeley, 1968.

Frank R. Hampel. The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69(346):383–393, 1974.

Frank R. Hampel, Elvezio M. Ronchetti, Peter J. Rousseeuw, and Werner A. Stahel. *Robust Statistics: The Approach Based on Influence Functions*, volume 114. John Wiley & Sons, 2011.

Ran He, Bao-Gang Hu, Wei-Shi Zheng, and Xiang-Wei Kong. Robust principal component analysis based on maximum correntropy criterion. *IEEE Transactions on Image Processing*, 20(6):1485–1494, 2011.

Ran He, Tieniu Tan, Liang Wang, and Wei-Shi Zheng. $\ell_{2,1}$-regularized correntropy for robust feature selection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012*, pages 2504–2511. IEEE, 2012.

Eva Herrmann and Klaus Ziegler. Rates of consistency for nonparametric estimation of the mode in absence of smoothness assumptions. *Statistics & Probability Letters*, 68(4): 359–368, 2004.

Ting Hu, Jun Fan, Qiang Wu, and Ding-Xuan Zhou. Learning theory approach to minimum error entropy criterion. *Journal of Machine Learning Research*, 14:377–397, 2013.

Peter J. Huber. Finite sample breakdown of M-and P-estimators. *The Annals of Statistics*, 12(1):119–126, 1984.

Peter J. Huber and Elvezio Ronchetti. *Robust Statistics*. Wiley, 2009.

Gordon C.R. Kemp and Joao Santos Silva. Regression towards the mode. *Journal of Econometrics*, 170(1):92–101, 2012.

Myoung-Jae Lee. Mode regression. *Journal of Econometrics*, 42(3):337–349, 1989.

Myoung-Jae Lee. Quadratic mode regression. *Journal of Econometrics*, 57(1):1–19, 1993.

Myoung-Jae Lee and Hyun Ah Kim. Semiparametric econometric estimators for a truncated regression model: a review with an extension. *Statistica Neerlandica*, 52(2):200–225, 1998.

Weifeng Liu, Puskal P. Pokharel, and Jose C. Principe. Correntropy: properties and applications in non-Gaussian signal processing. *IEEE Transactions on Signal Processing*, 55 (11):5286–5298, 2007.

Canyi Lu, Jinhui Tang, Min Lin, Liang Lin, Shuicheng Yan, and Zhouchen Lin. Correntropy induced l2 graph for robust subspace clustering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1801–1808, 2013.

Zhike Lv, Huiming Zhu, and Keming Yu. Robust variable selection for nonlinear models with diverging number of parameters. *Statistics & Probability Letters*, 91:90–97, 2014.

Eric Matzner-Løfber, Ali Gannoun, and Jan G. De Gooijer. Nonparametric forecasting: a comparison of three kernel-based methods. *Communications in Statistics - Theory and Methods*, 27(7):1593–1617, 1998.

Elias Ould-Saïd. A note on ergodic processes prediction via estimation of the conditional mode function. *Scandinavian Journal of Statistics*, 24(2):231–239, 1997.

Emanuel Parzen. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076, 1962.

Jose C. Principe. *Information Theoretic Learning: Renyi's Entropy and Kernel Perspectives*. Springer Science & Business Media, 2010.

Alejandro Quintela-Del-Rio and Philippe Vieu. A nonparametric conditional mode estimate. *Journal of Nonparametric Statistics*, 8(3):253–266, 1997.

Tim Robertson and Jonathan D. Cryer. An iterative procedure for estimating the mode. *Journal of the American Statistical Association*, 69(348):1012–1016, 1974.

Peter J. Rousseeuw and Annick M. Leroy. *Robust Regression and Outlier Detection*, volume 589. John Wiley & Sons, 2005.

Thomas W. Sager and Ronald A. Thisted. Maximum likelihood estimation of isotonic modal regression. *The Annals of Statistics*, 10(3):690–707, 1982.

Khardani Salah and Yao Anne Françoise. Nonlinear parametric mode regression. *Communications in Statistics - Theory and Methods*, 46(6):3006–3024, 2016.

Mrityunjay Samanta and Aerambamoorthy Thavaneswaran. Non-parametric estimation of the conditional mode. *Communications in Statistics-Theory and Methods*, 19(12):4515–4524, 1990.

Hiroaki Sasaki, Yurina Ono, and Masashi Sugiyama. Modal regression via direct log-density eerivative estimation. In *International Conference on Neural Information Processing*, pages 108–116. Springer, 2016.

Robert G. Staudte and Simon J. Sheather. *Robust Estimation and Testing*, volume 918. John Wiley & Sons, 2011.

Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Springer, New York, 2008.

Johan A.K. Suykens, Jos De Brabanter, Lukas Lukas, and Joos Vandewalle. Weighted least squares support vector machines: robustness and sparse approximation. *Neurocomputing*, 48(1):85–105, 2002.

John W. Tukey. *Exploratory Data Analysis*. Reading, Mass., 1977.

Arnout Van Messem and Andreas Christmann. A review on consistency and robustness properties of support vector machines for heavy-tailed distributions. *Advances in Data Analysis and Classification*, 4(2-3):199–220, 2010.

Matt P. Wand and Chris M. Jones. *Kernel Smoothing*. Chapman & Hall, London, 1994.

Xueqin Wang, Yunlu Jiang, Mian Huang, and Heping Zhang. Robust variable selection with exponential squared loss. *Journal of the American Statistical Association*, 108(502): 632–643, 2013.

Qiang Wu, Yiming Ying, and Ding-Xuan Zhou. Multi-kernel regularized classifiers. *Journal of Complexity*, 23(1):108–134, 2007.

Weixin Yao and Longhai Li. A new regression model: modal linear regression. *Scandinavian Journal of Statistics*, 41(3):656–671, 2014.

Weixin Yao and Sijia Xiang. Nonparametric and varying coefficient modal regression. *arXiv preprint arXiv:1602.06609*, 2016.

Weixin Yao, Bruce G. Lindsay, and Runze Li. Local modal regression. *Journal of Nonparametric Statistics*, 24(3):647–663, 2012.

Victor J. Yohai. High breakdown-point and high efficiency robust estimates for regression. *The Annals of Statistics*, 15(2):642–656, 1987.

Keming Yu and Katerina Aristodemou. Bayesian mode regression. *arXiv preprint arXiv:1208.0579*, 2012.

Keming Yu, Katerina Aristodemou, Frauke Becker, and Joann Lord. Fast mode regression in big data analysis. In *Proceedings of the 2014 International Conference on Big Data Science and Computing*, page 24. ACM, 2014.

Yaoliang Yu, Özlem Aslan, and Dale Schuurmans. A polynomial-time form of robust regression. In *Advances in Neural Information Processing Systems*, pages 2483–2491, 2012.

Tong Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32(1):56–85, 2004.

Haiming Zhou and Xianzheng Huang. Nonparametric modal regression in the presence of measurement error. *Electronic Journal of Statistics*, 10(2):3579–3620, 2016.