Appl. Comput. Harmon. Anal. ••• (••••) •••–•••

Contents lists available at ScienceDirect

Applied and Computational Harmonic Analysis

www.elsevier.com/locate/acha

Letter to the Editor

Learning with correntropy-induced losses for regression with mixture of symmetric stable noise

Yunlong Feng *, Yiming Ying

Department of Mathematics and Statistics, State University of New York at Albany, NY, USA

ARTICLE INFO

Article history:
Received 25 June 2018
Received in revised form 12 March 2019
Accepted 1 September 2019
Available online xxxx
Communicated by Naoki Saito

ABSTRACT

In recent years, correntropy and its applications in machine learning have been drawing continuous attention owing to its merits in dealing with non-Gaussian noise and outliers. However, theoretical understanding of correntropy, especially in the learning theory context, is still limited. In this study, we investigate correntropy based regression in the presence of non-Gaussian noise or outliers within the statistical learning framework. Motivated by the practical way of generating non-Gaussian noise or outliers, we introduce mixture of symmetric stable noise, which include Gaussian noise, Cauchy noise, and their mixture as special cases, to model non-Gaussian noise or outliers. We demonstrate that under the mixture of symmetric stable noise assumption, correntropy based regression can learn the conditional mean function or the conditional median function well without resorting to the finite-variance or even the finite first-order moment condition on the noise. In particular, for the above two cases, we establish asymptotic optimal learning rates for correntropy based regression estimators that are asymptotically of type $\mathcal{O}(n^{-1})$. These results justify the effectiveness of the correntropy based regression estimators in dealing with outliers as well as non-Gaussian noise. We believe that the present study makes a step forward towards understanding correntropy based regression from a statistical learning viewpoint, and may also shed some light on robust statistical learning for regression.

1. Introduction and motivation

Within the information-theoretic learning framework developed in [44], correntropy was proposed in [48, 36] and serves as a similarity measure between two random variables. Given two scalar random variables U, V, the correntropy $\mathcal{V}_\sigma$ between U and V is defined as $\mathcal{V}_\sigma(U,V) = \mathbb{E}\mathcal{K}_\sigma(U,V)$ with $\mathcal{K}_\sigma$ a Gaussian kernel given by $\mathcal{K}_\sigma(u,v) = \exp\left\{-(u-v)^2/\sigma^2\right\}$, the scale parameter $\sigma > 0$, and (u,v) a realization of (U,V). It is noticed in [36] that the correntropy $\mathcal{V}_\sigma(U,V)$ can induce a new metric between U and V. It is argued in [36,44] that this new metric could be a better option in measuring the distance between U and V than the Euclidean

* Corresponding author.
E-mail address: ylfeng@albany.edu (Y. Feng).

https://doi.org/10.1016/j.acha.2019.09.001
1063-5203/

metric when the random variable defined by the residual $U - V$ admits a non-Gaussian distribution which is frequently encountered in applications. During the past several years, the merits of correntropy have been verifying by numerous real-world applications across various fields, e.g., signal processing [36,8,9,7,68], image processing [24,26,25,22,61,62,67,63], time series forecasting [4,5,40], and many other machine learning tasks such as regression, classification, and clustering [60,52,66]. Noticing that most of the above mentioned problems can be interpreted from a regression viewpoint, recently some understanding towards correntropy based regression in statistical learning has been conducted in [18] and [17], to which the present study is closely related. We, therefore, first revisit the conclusions on correntropy based regression drawn in [18] and [17].

1.1. Formulating correntropy based regression

We start with the following frequently assumed data-generating model in nonparametric regression

$$Y = f^\star(X) + \varepsilon, \tag{1}$$

where X is the independent variable that takes values in a compact metric space $\mathcal{X} \subset \mathbb{R}^d$, Y the dependent variable that takes value in $\mathcal{Y} = \mathbb{R}$, and ε the noise variable. We assume that $\mathbb{E}(\varepsilon|X) = 0$ if it exists, otherwise, we assume that $\mathsf{median}(\varepsilon|X) = 0$. In regression problems, it is typical that we can only access a set of i.i.d. observations $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^n$ generated by (1). Our purpose in regression is to infer the unknown truth $f^\star$ while only referring to these observations.

The idea of correntropy based regression is to select the hypothesis from a hypothesis space that maximizes the empirical correntropy estimator between $\{y_i\}_{i=1}^n$ and $\{f(x_i)\}_{i=1}^n$ for any $f : \mathcal{X} \to \mathbb{R}$, which we term as the Maximum Correntropy Criterion based Regression (MCCR) [18]. Recall that the following correntropy induced loss $\ell_\sigma : \mathbb{R} \to [0, +\infty)$ is defined in [18]:

$$\ell_\sigma(t) = \sigma^2 \left(1 - e^{-\frac{t^2}{\sigma^2}}\right), \ t \in \mathbb{R}, \tag{2}$$

where $\sigma > 0$ is a tuning parameter. MCCR can be formulated into the following empirical risk minimization scheme

$$f_{\mathbf{z}} := \arg\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \ell_\sigma(y_i - f(x_i)), \tag{3}$$

where $\mathcal{H}$ is a hypothesis space that is assumed to be a compact subset of $C(\mathcal{X})$.

1.2. MCCR in statistical learning

As mentioned above, in the literature, correntropy and its applications in various fields have been investigated. However, in the statistical learning context, theoretical understanding of correntropy based regression estimators is still limited. Unlike commonly employed error metric in regression problems, the error metric induced by correntropy is non-convex and involves a scale parameter σ, which complicate the analysis. Recently, [18] investigated correntropy based regression when the scale parameter $\sigma := \sigma(n)$ goes large in correspondence to the sample size n, which was inspired by the studies in [27,16] on empirical minimum error entropy minimization algorithms. When the scale parameter $\sigma(n)$ tends to zero, [17] made some efforts in order to understand correntropy in regression problems and assess the performance of the correntropy based regression estimators from a statistical learning viewpoint. The main concerns in [18] and [17] are the learning performance of $f_{\mathbf{z}}$ when the sample size n goes to infinity, where different scenarios of the noise

Please cite this article in press as: Y. Feng, Y. Ying, Learning with correntropy-induced losses for regression with mixture of symmetric stable noise, Appl. Comput. Harmon. Anal. (2019), https://doi.org/10.1016/j.acha.2019.09.001

variable ε and the choices of the σ values were considered. Briefly, the following conclusions were drawn in the above-mentioned two studies:

- By relating the scale parameter σ to the sample size n (i.e., $\sigma := \sigma(n)$) and assuming that the noise variable ε is zero-mean, with a diverging and properly chosen σ value, $f_{\mathbf{z}}$ can approximate the conditional mean function $f^\star$ robustly. Convergence rates were established in the absence of light-tailed assumptions, which justifies the robustness of $f_{\mathbf{z}}$. Moreover, the scale parameter σ, in this case, plays a trade-off role between robustness and the approximation ability of the estimator $f_{\mathbf{z}}$.
- By relating the scale parameter σ to the sample size n and assuming a unique zero global mode of the noise ε, with a tending-to-zero and properly chosen σ value, $f_{\mathbf{z}}$ approaches the conditional mode function $f^\star$. Note that the unique zero global mode assumption on ε allows asymmetric or heavy-tailed noise, which again explains the robustness of the MCCR estimator $f_{\mathbf{z}}$ in this case.
- With a properly chosen scale parameter σ, the correntropy based regression estimator $f_{\mathbf{z}}$ is shown to be equivalent to least squares regression estimator in the presence of symmetric and bounded noise. In this case, the equivalence is claimed in the following two senses: first, similar as that of the least squares regression estimator under the same noise condition, the population version of $f_{\mathbf{z}}$ is exactly the conditional mean function $f^\star$. Second, the convergence rates of $f_{\mathbf{z}}$ to the conditional mean function are comparable to that of least squares regression estimators.

Some merits of MCCR can be observed from the above statements. For example, MCCR can learn $f^\star$ well in the absence of light-tailed noise assumptions where least squares regression estimators are not capable. On the other hand, it also performs comparable with least squares regression estimators in the presence of bounded and symmetric noise where the latter one achieves its optimal performance. We refer to Section 6 in [17] for a general picture of existing understanding on correntropy based regression in statistical learning.

1.3. Motivation and contribution

The prominent advantages of MCCR estimator lie in its resistance ability to heavy-tailed noise and outliers. As stated above, the conducted theoretical assessments on MCCR estimators in [18] and [17] justify its superior performance in dealing with heavy-tailed noise. However, several fundamental problems related to MCCR estimators in statistical learning still remain unclear. For instance:

Problem I: Learning performance of MCCR in the presence of Gaussian noise. When Gaussian noise is present, least squares regression estimators are known to achieve their optimal performance and optimal learning rates of type $\mathcal{O}(n^{-1})$ have been established in the statistical learning literature, see e.g., [59] and [19]. Under the same noise assumption, asymptotic learning rates of type $\mathcal{O}(n^{-2/3})$ can be deduced by following the work in [18], which are not comparable with that of least squares regression estimators. Notice that the correntropy induced loss ℓ_σ is Lipschitz continuous and bounded on $\mathbb{R}$, and the fact that ℓ_σ approximates the least squares loss when σ is large enough. It is natural to conjecture that optimal learning rates of MCCR estimators may be also achievable as least squares regression estimators in the presence of Gaussian noise.

Problem II: Learning performance of MCCR with heavy-tailed noise. In the presence of heavy-tailed noise with finite variance, from [18] we know that asymptotic learning rates of type $\mathcal{O}(n^{-2/3})$ for MCCR can be established under moment assumptions. If the heavy-tailed noise has infinite variance or even infinite first-order moment condition (such as Cauchy noise), asymptotic learning rates of type $\mathcal{O}(n^{-2/5})$ were established in [17] under mild assumptions. However, both of the above two types of learning rates are far from the type $\mathcal{O}(n^{-1})$, which are regarded as optimal in statistical learning.

Please cite this article in press as: Y. Feng, Y. Ying, Learning with correntropy-induced losses for regression with mixture of symmetric stable noise, Appl. Comput. Harmon. Anal. (2019), https://doi.org/10.1016/j.acha.2019.09.001

Problem III: Understanding MCCR in the presence of outliers. When outliers are presented, how MCCR estimators learn the unknown truth function $f^\star$ still remains unclear, although empirically their superior performance in dealing with outliers has been observed. As mentioned above, this is, in fact, one of the most prominent advantages of MCCR estimators over other regression estimators. The main barrier to understanding MCCR in the presence of outliers lies in the modeling of outliers in analysis. This is because for the time being there exists no distribution independent definition of outlier and more frequently, outliers are defined in association with concrete distributions, see e.g., [23,46,1].

The present study aims to address the above three concerns on correntropy based regression, especially the concern of understanding MCCR in the presence of outliers. We start with the following motivating observation: a very frequently employed technique of generating outliers in robust statistics [57,28,31,29,20], machine learning [49,21], as well as many engineering applications [32,34] is as follows

$$\varepsilon \sim \lambda_1 \mathcal{N}(\mu_1, \sigma_1^2) + \lambda_2 \mathcal{N}(\mu_2, \sigma_2^2), \tag{4}$$

where $\lambda_1 + \lambda_2 = 1$, $\lambda_1 \gg \lambda_2$, $\sigma_1^2 \ll \sigma_2^2$, and $\mathcal{N}(\mu_1, \sigma_1^2)$, $\mathcal{N}(\mu_2, \sigma_2^2)$ are two Gaussian distributions with mean μ_1, μ_2 and variance σ_1^2, σ_2^2, respectively. In (4), $\mathcal{N}(\mu_1, \sigma_1^2)$ is usually considered as background noise while $\mathcal{N}(\mu_2, \sigma_2^2)$ is regarded as the contaminating noise that generates outliers since σ_2^2 is far larger than σ_1^2. In some cases, other distributions that have heavier tails than Gaussian (such as Cauchy noise) may be also employed in (4) as contaminating noise. On the other hand, we notice that both Gaussian noise and Cauchy noise belong to the type of symmetric stable noise. These observations remind us to impose the mixture of symmetric stable noise assumption on ε and study the performance of MCCR in this case. In fact, as we shall see later, mixture of symmetric stable distributions have been frequently employed in many engineering applications to model impulsive noise. Another nice property of mixture of symmetric stable noise lies in that it can approximate the distribution of any noise arbitrarily well.

With the introduction of mixture of symmetric stable distributions in modeling heavy-tailed noise or outliers, in this paper, we make a step forward in understanding correntropy based regression in statistical learning. More detailed speaking, concerning the study of correntropy based regression estimators, in this work, we make the following contributions:

- We introduce the mixture of symmetric stable distributions to model the noise ε. The family of mixture of symmetric stable noise includes the Gaussian noise, the mixture Gaussian noise, the Cauchy noise, and many other kinds of mixture noise, and so is capable of modeling heavy-tailed noise and outliers. We notice that within the statistical learning framework, we make some first attempts in modeling outliers via mixture of symmetric stable distributions.
- Under the mixture of symmetric stable noise assumption, we demonstrate that MCCR estimators can learn the unknown truth function $f^\star$ in an unbiased way in that the population version of $f_{\mathbf{z}}$ is exactly $f^\star$. Recall that $f^\star$ is the conditional mean function or the conditional median function, and the mixture of symmetric stable noise consists of a large family of noise from light-tailed to heavy-tailed. This indicates that MCCR could be employed to learn $f^\star$ after seeing enough observations without resorting to the sub-Gaussianity of the noise.
- We establish asymptotic learning rates of type $\mathcal{O}(n^{-1})$ which are comparable with those of least squares regression estimators under the sub-Gaussianity noise assumption. As stated above, the mixture of symmetric stable noise include Gaussian noise and Cauchy noise as two special cases, and can be used to model outliers. Therefore, the present study provides direct answers to the three problems stated above. In fact, establishing almost sure convergence rates of type $\mathcal{O}(n^{-1})$ in learning theory without appealing to finite variance assumption of the noise may be of independent interest.

Please cite this article in press as: Y. Feng, Y. Ying, Learning with correntropy-induced losses for regression with mixture of symmetric stable noise, Appl. Comput. Harmon. Anal. (2019), https://doi.org/10.1016/j.acha.2019.09.001

The rest of this paper is organized as follows. In Section 2, we provide the definitions of symmetric stable distributions and mixture of symmetric stable distributions and introduce some of their applications. Section 3 is concerned with the assessments of correntropy based regression in the presence of mixture of symmetric stable noise. The performance of MCCR, in this case, will be studied in this section, and results on learning rates of MCCR estimators will be presented here. We will also give some comments on the obtained learning rates and the MCCR estimator in this section. The paper is concluded in Section 5.

2. Mixture of symmetric stable distributions and its applications

In this section, we introduce the mixture of symmetric stable distributions and its applications. To this end, we shall first introduce the symmetric stable distribution.

Definition 1 *(Symmetric stable distribution [47]).* A univariate distribution function is symmetric stable if its characteristic function takes the following form

$$\phi(t) = \exp\left\{\mathrm{i}\mu t - \gamma|t|^{\alpha}\right\}, \quad \text{for any} \quad t \in \mathbb{R},$$

where $-\infty < \mu < \infty$, $\gamma > 0$, $0 < \alpha \leq 2$, and i is the imaginary unit.

More precisely, the symmetric stable distribution defined in Definition 1 is said to be α-stable and symmetric about the location μ. As shown in Definition 1, a symmetric stable distribution has three parameters, namely, the location parameter μ, the scale parameter γ, and the characteristic exponent α. The characteristic exponent α is a shape parameter and measures the thickness of the tails of the density function. Two typical examples of symmetric stable distributions are Gaussian distribution ($\alpha = 2$) and Cauchy distribution ($\alpha = 1$). A symmetric stable distribution with $0 < \alpha < 2$ only admits absolute moments of order less than α. Therefore, all symmetric stable distributions do not have finite variance except for the Gaussian distribution. For more properties of symmetric stable distributions, we refer to [14,41,47].

When a univariate distribution P consists of different components with each of which a symmetric stable distribution and can be expressed as a convex combination of these components, it is called a mixture of symmetric stable distributions [38].

Definition 2 *(Mixture of symmetric stable distributions).* A univariate distribution P with density p is a mixture of symmetric stable distributions if it is a convex combination of symmetric stable distributions $\{P_i\}_{i=1}^K$ with density function $\{p_i\}_{i=1}^K$ and K a positive integer, i.e., there exists $\lambda_1, \cdots, \lambda_K$ with $\lambda_i > 0$ for $i = 1, \ldots, K$, and $\sum_{i=1}^K \lambda_i = 1$, such that

$$P(t) = \sum_{i=1}^{K} \lambda_i P_i(t), \quad \text{and} \quad p(t) = \sum_{i=1}^{K} \lambda_i p_i(t), \quad \text{for any} \quad t \in \mathbb{R}.$$

In Definition 2, $\lambda_1, \ldots, \lambda_K$ are called the mixing weights and $p_1, \ldots, p_K$ are component densities. It is obvious that when $K = 1$, a mixture of symmetric stable distributions is reduced to a symmetric stable distribution. In particular, if $p_1, \ldots, p_K$ are normal densities, then p is a mixture of Gaussian. A nice property of the mixture of Gaussian density is that it can approximate any density function to arbitrary accuracy with suitable choice of parameters and enough components K [56,38].

Symmetric stable distributions have been drawing continuous attention in the statistics literature [14, 15,12,41,10]. The mixture of symmetric stable distributions, which includes the mixture of Gaussian and symmetric stable distributions as special cases, has been extensively applied into many applications. As mentioned above, in robust statistics, it has been employed to mimic perturbed or heavy-tailed distributions,

Please cite this article in press as: Y. Feng, Y. Ying, Learning with correntropy-induced losses for regression with mixture of symmetric stable noise, Appl. Comput. Harmon. Anal. (2019), https://doi.org/10.1016/j.acha.2019.09.001

see e.g., [29]. In many engineering applications, especially applications in the field of signal processing, image processing, and wireless communications, it has been frequently applied to model impulsive noise [50,2,42, 30,35,13,33,6,37,54,45,43] or outliers [3,1].

3. MCCR with mixture of symmetric stable noise

The noise is mixture of symmetric stable noise if its distribution is a mixture of symmetric stable distributions. As stated in the above section, it can be employed to model non-Gaussian noise and outliers. In this section, we study MCCR from a statistical learning viewpoint in the presence of mixture of symmetric stable noise ε. We start with the introduction of several notations and assumptions.

3.1. Notations and assumptions

We denote the unknown probability distribution over $\mathcal{X} \times \mathcal{Y}$ as ρ and ρ_X as the marginal distribution of ρ over $\mathcal{X}$. For any $f \in \mathcal{H}$, the empirical error in (3) is denoted as $\mathcal{E}^\sigma_{\mathbf{z}}(f)$, that is,

$$\mathcal{E}^\sigma_{\mathbf{z}}(f) = \frac{1}{n}\sum_{i=1}^{n} \ell_\sigma(y_i - f(x_i)),$$

and its population version $\mathcal{E}^\sigma(f)$ is defined as

$$\mathcal{E}^\sigma(f) = \int_{\mathcal{X}\times\mathcal{Y}} \ell_\sigma(y - f(x))\mathrm{d}\rho.$$

The distance between f and $f^\star$ in $L^2_{\rho_X}$ is denoted as $\|f - f^\star\|^2_\rho$. Besides, for any two quantities a, b, we denote $a \lesssim b$ if there exists a positive constant c such that $a \le cb$.

Assumption 1 *(Mixture of symmetric stable noise).* The distribution of the noise ε is a mixture of symmetric stable distributions with location parameter 0, i.e., the density $p_{\varepsilon,x}$ of the noise variable ε for any $x \in \mathcal{X}$ takes the following form

$$p_{\varepsilon,x}(t) = \sum_{i=1}^{K} \lambda_i p_{\varepsilon,x,i}(t), \quad \text{for any} \quad t \in \mathbb{R},$$

where K is a positive integer, $\lambda_i > 0$ for $i = 1, \ldots, K$, $\sum_{i=1}^K \lambda_i = 1$, and $p_{\varepsilon,x,i}$ is the density function of the symmetric stable distribution $P_{\varepsilon,x,i}$ that is centered around 0 for $i = 1, \ldots, K$.

The second assumption is on the complexity of $\mathcal{H}$ in terms of the ℓ^2-empirical covering number $\mathcal{N}_2(\mathcal{H}, \eta)$, see e.g., [65,51,19], which is defined as follows.

Definition 3. Let $\mathbf{x} = \{x_1, x_2, \ldots, x_n\} \subset \mathcal{X}^n$. The ℓ^2-empirical covering number of the hypothesis space $\mathcal{H}$, which is denoted as $\mathcal{N}_2(\mathcal{H}, \eta)$ with radius $\eta > 0$, is defined by

$$\mathcal{N}_2(\mathcal{H}, \eta) := \sup_{n\in\mathbb{N}} \sup_{\mathbf{x}\in\mathcal{X}^n} \inf\Big\{\ell \in \mathbb{N} : \exists\{f_i\}_{i=1}^{\ell} \subset \mathcal{H} \text{ such that for each } f \in \mathcal{H}, \text{there exists some}$$

$$i \in \{1, 2, \ldots, \ell\} \text{ with } \frac{1}{n}\sum_{j=1}^{n} |f(x_j) - f_i(x_j)|^2 \le \eta^2\Big\}.$$

Please cite this article in press as: Y. Feng, Y. Ying, Learning with correntropy-induced losses for regression with mixture of symmetric stable noise, Appl. Comput. Harmon. Anal. (2019), https://doi.org/10.1016/j.acha.2019.09.001

Assumption 2 *(Complexity Assumption).* There exist positive constants $0 < s < 2$ and c such that

$$\log \mathcal{N}_2(\mathcal{H}, \eta) \leq c\eta^{-s}, \ \forall\ \eta > 0.$$

Throughout this paper, we also assume that there exists a positive constant M such that $\sup_{f\in\mathcal{H}} \|f\|_\infty \leq M$, and $\|f^\star\|_\infty \leq M$.

3.2. Unbiasedness of MCCR with mixture of symmetric stable noise

In the presence of mixture of symmetric stable noise, in this part, we will show that MCCR can learn $f^\star$ in an unbiased way. This is stated in the sense of the following theorem, which is established by applying techniques proposed in [16].

Theorem 1. *Suppose that Assumption 1 holds and $f^\star \in \mathcal{H}$. Then we have*

$$f^\star = \arg\min_{f\in\mathcal{H}} \mathcal{E}^\sigma(f),$$

and for any $f \in \mathcal{H}$, it holds that

$$c_{\sigma,\gamma,\alpha}\|f - f^\star\|_\rho^2 \leq \mathcal{E}^\sigma(f) - \mathcal{E}^\sigma(f^\star) \leq \|f - f^\star\|_\rho^2,$$

where $c_{\sigma,\gamma,\alpha}$ is a positive constant that will be given explicitly in the proof.

Proof. From the definitions of the notions, we know that

$$\mathcal{E}^\sigma(f) - \mathcal{E}^\sigma(f^\star) = \sigma^2 \int_{\mathcal{X}} [F_x(f(x) - f^\star(x)) - F_x(0)]\mathrm{d}\rho_X(x),$$

where $F_x : \mathbb{R} \to \mathbb{R}$ is denoted as

$$F_x(u) := 1 - \int_{-\infty}^{+\infty} \exp\left\{-\frac{(t-u)^2}{\sigma^2}\right\} p_{\varepsilon,x}(t)\mathrm{d}t, \ x \in \mathcal{X}.$$

From the Taylor's theorem, we know that

$$F_x(f(x) - f^\star(x)) - F_x(0) = F_x'(0)(f(x) - f^\star(x)) + \frac{F_x''(\zeta_x)}{2}(f(x) - f^\star(x))^2,$$

where for any $x \in \mathcal{X}$, $0 < \zeta_x < f(x) - f^\star(x)$. Due to the symmetry assumption of the noise, for any $x \in \mathcal{X}$, we have

$$F_x'(0) = -2\int_{-\infty}^{+\infty} \exp\left(-\frac{t^2}{\sigma^2}\right)\left(\frac{t}{\sigma^2}\right) p_{\varepsilon,x}(t)\mathrm{d}t = 0,$$

and

$$F_x''(\zeta_x) = 2\int_{-\infty}^{+\infty} \exp\left\{-\frac{(t-\zeta_x)^2}{\sigma^2}\right\}\left(\frac{\sigma^2 - 2(t-\zeta_x)^2}{\sigma^4}\right) p_{\varepsilon,x}(t)\mathrm{d}t, \ x \in \mathcal{X}.$$

Please cite this article in press as: Y. Feng, Y. Ying, Learning with correntropy-induced losses for regression with mixture of symmetric stable noise, Appl. Comput. Harmon. Anal. (2019), https://doi.org/10.1016/j.acha.2019.09.001

It is obvious that for any $x \in \mathcal{X}$, the following inequality

$$F_x''(u) \leq \frac{2}{\sigma^2}$$

holds uniformly for $0 < u < f(x) - f^\star(x)$. Therefore, we have

$$\begin{aligned}\mathcal{E}^\sigma(f) - \mathcal{E}^\sigma(f^\star) &= \sigma^2 \int_{\mathcal{X}} [F_x(f(x) - f^\star(x)) - F_x(0)] \mathrm{d}\rho_X(x) \\ &= \frac{\sigma^2}{2} \int_{\mathcal{X}} F_x''(\zeta_x)(f(x) - f^\star(x))^2 \mathrm{d}\rho_X(x) \\ &\leq \int_{\mathcal{X}} (f(x) - f^\star(x))^2 \mathrm{d}\rho_X(x).\end{aligned} \tag{5}$$

On the other hand, with simple computations, we have

$$\begin{aligned}\mathcal{E}^\sigma(f) - \mathcal{E}^\sigma(f^\star) &= \sigma^2 \int_{\mathcal{X}} \int_{-\infty}^{+\infty} \left[\exp\left(-\frac{t^2}{\sigma^2}\right) - \exp\left(-\frac{(t - [f(x) - f^\star(x)])^2}{\sigma^2}\right)\right] p_{\varepsilon,x}(t) \mathrm{d}t \mathrm{d}\rho_X(x) \\ &= \sigma^2 \int_{\mathcal{X}} \int_{-\infty}^{+\infty} \left[\exp\left(-\frac{t^2}{\sigma^2}\right) - \exp\left(-\frac{(t - [f(x) - f^\star(x)])^2}{\sigma^2}\right)\right] p_{\varepsilon,x}(t) \mathrm{d}t \mathrm{d}\rho_X(x) \\ &= \sigma^2 \int_{\mathcal{X}} \int_{-\infty}^{+\infty} \left[\exp\left(-\frac{t^2}{\sigma^2}\right) - \exp\left(-\frac{(t - u_x)^2}{\sigma^2}\right)\right] p_{\varepsilon,x}(t) \mathrm{d}t \mathrm{d}\rho_X(x) \\ &= \sigma^2 \int_{\mathcal{X}} \int_{-\infty}^{+\infty} \left[\exp\left(-\frac{(t + u_x)^2}{\sigma^2}\right) - \exp\left(-\frac{t^2}{\sigma^2}\right)\right] p_{\varepsilon,x}(t) \mathrm{d}t \mathrm{d}\rho_X(x),\end{aligned}$$

where for any $x \in \mathcal{X}$, $u_x = f(x) - f^\star(x)$. From Assumption 1 on the noise and recalling the linearity property of the Fourier transform, we have

$$\widehat{p_{\epsilon,x}}(\xi) = \sum_{i=1}^{K} \lambda_i \widehat{p_{\epsilon,x,i}}(\xi),$$

where $\widehat{p_{\epsilon,x}}$ is the Fourier transform of $p_{\epsilon,x}$, and $\widehat{p_{\epsilon,x,i}}$ is the Fourier transform of $p_{\epsilon,x,i}$, $i = 1, \ldots, K$. Moreover, for $i = 1, \ldots, K$, since $P_{\epsilon,x,i}$ is a symmetric stable distribution with the location parameter 0, we know that there exist $\gamma_i > 0$ and $0 < \alpha_i \leq 2$ such that

$$\widehat{p_{\epsilon,x,i}}(\xi) = e^{-\gamma_i |\xi|^{\alpha_i}}.$$

Applying the Planchel formula, we obtain

$$\begin{aligned}\mathcal{E}^\sigma(f) - \mathcal{E}^\sigma(f^\star) &= \frac{\sigma^3}{2\sqrt{\pi}} \int_{\mathcal{X}} \int_{-\infty}^{+\infty} \exp\left(-\frac{\sigma^2 \xi^2}{4}\right) \widehat{p_{\epsilon,x}}(\xi) \left[1 - \mathbf{e}^{\mathrm{i}\xi u_x}\right] \mathrm{d}\xi \mathrm{d}\rho_X(x) \\ &= \frac{\sigma^3}{\sqrt{\pi}} \sum_{i=1}^{K} \lambda_i \int_{\mathcal{X}} \int_{-\infty}^{+\infty} \exp\left(-\frac{\sigma^2 \xi^2}{4}\right) \widehat{p_{\epsilon,x,i}}(\xi) \sin^2\left(\frac{\xi(f(x) - f^\star(x))}{2}\right) \mathrm{d}\xi \mathrm{d}\rho_X(x)\end{aligned}$$

Please cite this article in press as: Y. Feng, Y. Ying, Learning with correntropy-induced losses for regression with mixture of symmetric stable noise, Appl. Comput. Harmon. Anal. (2019), https://doi.org/10.1016/j.acha.2019.09.001

$$= \frac{\sigma^3}{\sqrt{\pi}} \int_{\mathcal{X}} \sum_{i=1}^{K} \lambda_i \int_{-\infty}^{+\infty} \exp\left(-\frac{\sigma^2\xi^2}{4} - \gamma_i|\xi|^{\alpha_i}\right) \sin^2\left(\frac{\xi(f(x) - f^\star(x))}{2}\right) \mathrm{d}\xi \mathrm{d}\rho_X(x),$$

where the second equality is due to the fact that $\mathcal{E}^\sigma(f) - \mathcal{E}^\sigma(f^\star)$ is real for any $f \in \mathcal{H}$. For any $x \in \mathcal{X}$, $|u_x| = |f(x) - f^\star(x)| \leq 2M$. When $|\xi| \leq \frac{\pi}{2M}$, from Jordan's inequality, it holds that

$$\sin^2\left(\frac{\xi(f(x) - f^\star(x))}{2}\right) \geq \frac{2\xi^2(f(x) - f^\star(x))^2}{\pi^2}.$$

As a result, we come to the following conclusion

$$\begin{aligned} \mathcal{E}^\sigma(f) - \mathcal{E}^\sigma(f^\star) &\geq \frac{2\sigma^3}{\pi^{5/2}} \int_{\mathcal{X}} \sum_{i=1}^{K} \lambda_i \int_{-\frac{\pi}{2M}}^{\frac{\pi}{2M}} \xi^2 \exp\left(-\frac{\sigma^2\xi^2}{4} - \gamma_i|\xi|^{\alpha_i}\right) (f(x) - f^\star(x))^2 \mathrm{d}\xi \mathrm{d}\rho_X(x) \\ &= c_{\sigma,\gamma,\alpha} \int_{\mathcal{X}} (f(x) - f^\star(x))^2 \mathrm{d}\rho_X(x), \end{aligned} \tag{6}$$

where

$$c_{\sigma,\gamma,\alpha} = \frac{2\sigma^3}{\pi^{5/2}} \sum_{i=1}^{K} \lambda_i \int_{-\frac{\pi}{2M}}^{\frac{\pi}{2M}} \xi^2 \exp\left(-\frac{\sigma^2\xi^2}{4} - \gamma_i|\xi|^{\alpha_i}\right) \mathrm{d}\xi. \tag{7}$$

The positiveness of $c_{\sigma,\gamma,\alpha}$ implies that for any $f \in \mathcal{H}$, we have $\mathcal{E}^\sigma(f) \geq \mathcal{E}^\sigma(f^\star)$. That is,

$$f^\star = \arg\min_{f \in \mathcal{H}} \mathcal{E}^\sigma(f).$$

To prove the second assertion, we combine inequalities (5) and (6), and obtain

$$c_{\sigma,\gamma,\alpha}\|f - f^\star\|_\rho^2 \leq \mathcal{E}^\sigma(f) - \mathcal{E}^\sigma(f^\star) \leq \|f - f^\star\|_\rho^2,$$

where $c_{\sigma,\gamma,\alpha}$ is a positive constant given in (7). This completes the proof of Theorem 1. □

Theorem 1 states that in the presence of mixture of symmetric stable noise, the population version of the MCCR estimator $f_{\mathbf{z}}$ is exactly the underlying unknown truth function $f^\star$ as long as $f^\star$ belongs to $\mathcal{H}$. Therefore, in this sense, $f_{\mathbf{z}}$ can be regarded as an unbiased estimator of $f^\star$. Another implication of Theorem 1 is that under the mixture of symmetric stable noise assumption, the excess risk of MCCR can be upper and lower bounded by the $L^2_{\rho_X}$-distance between the MCCR estimator $f_{\mathbf{z}}$ and the unknown truth $f^\star$. As we shall see later, this leads to fast convergence rates of the MCCR estimator $f_{\mathbf{z}}$ to $f^\star$.

3.3. Performance of MCCR with mixture of symmetric stable noise

We are now in a position to evaluate the learning performance of MCCR in the presence of mixture of symmetric stable noise by establishing convergence rates of $\|f_{\mathbf{z}} - f^\star\|_\rho^2$.

Theorem 2. *Suppose that Assumption 1 and Complexity Assumption with $s > 0$ hold. Let $f_{\mathbf{z}}$ be produced by (3) and $f^\star \in \mathcal{H}$. For any $0 < \delta < 1$, with confidence $1 - \delta$, it holds that*

$$\|f_{\mathbf{z}} - f^\star\|_\rho^2 \lesssim \log(1/\delta) n^{-\frac{2}{2+s}}.$$

Please cite this article in press as: Y. Feng, Y. Ying, Learning with correntropy-induced losses for regression with mixture of symmetric stable noise, Appl. Comput. Harmon. Anal. (2019), https://doi.org/10.1016/j.acha.2019.09.001

When functions in $\mathcal{H}$ are sufficiently smooth, the index s could be arbitrarily small. Therefore, it is immediate to see that the convergence rates established in Theorem 2 are asymptotically of type $\mathcal{O}(n^{-1})$. Recall that in Theorem 2, the noise ε is only assumed to be a mixture of symmetric stable noise which include the mixture Gaussian and the Cauchy noise, and can be applied to model outliers. It is interesting to see that in this case the MCCR estimator $f_{\mathbf{z}}$ can learn the conditional mean function or the conditional median function $f^\star$ well. This, in fact, explains the merits of MCCR in dealing with heavy-tailed noise or outliers. Moreover, as far as we are aware, within the statistical learning framework, we present some first results on the optimal convergence rates of regression estimator without imposing finite-variance or even finite first-order moment conditions on the noise.

To prove Theorem 2, we need the following lemma established in [65].

Lemma 1. *Let $\mathcal{F}$ be a class of measurable functions on $\mathcal{Z}$. Assume that there are constants $B, c > 0$ and $\theta \in [0,1]$ such that $\|f\|_\infty \leq B$ and $\mathbb{E}f^2 \leq c(\mathbb{E}f)^\theta$ for every $f \in \mathcal{F}$. If for some $a > 0$ and $s \in (0,2)$,*

$$\log \mathcal{N}_2(\mathcal{F}, \eta) \leq a\eta^{-s}, \qquad \forall\, \eta > 0,$$

then there exists a constant α_p depending only on p such that for any $t > 0$, with probability at least $1 - e^{-t}$, there holds

$$\mathbb{E}f - \frac{1}{m}\sum_{i=1}^m f(z_i) \leq \frac{1}{2}\gamma^{1-\theta}(\mathbb{E}f)^\theta + \alpha_p\gamma + 2\left(\frac{ct}{m}\right)^{\frac{1}{2-\theta}} + \frac{18Bt}{m}, \qquad \forall\, f \in \mathcal{F},$$

where

$$\gamma := \max\left\{c^{\frac{2-s}{4-2\theta+s\theta}}\left(\frac{a}{m}\right)^{\frac{2}{4-2\theta+s\theta}}, B^{\frac{2-s}{2+s}}\left(\frac{a}{m}\right)^{\frac{2}{2+s}}\right\}.$$

Proof of Theorem 2. To prove Theorem 2, we apply Lemma 1 to the function set $\mathcal{F}_\mathcal{H}$ defined below

$$\mathcal{F}_\mathcal{H} = \left\{g \;\middle|\; g(z) = -\sigma^2\exp\left\{-(y - f(x))^2/\sigma^2\right\} + \sigma^2\exp\left\{-(y - f^\star(x))^2/\sigma^2\right\}, f \in \mathcal{H}, z \in \mathcal{Z}\right\}.$$

We first verify conditions in Lemma 1. From the definition of $\mathcal{F}_\mathcal{H}$, for any $g \in \mathcal{F}_\mathcal{H}$, we have

$$\|g\|_\infty \leq \sigma^2 + \sigma^2 = 2\sigma^2,$$

and the following Bernstein condition holds

$$\begin{aligned}
\mathbb{E}g^2 &= \int_\mathcal{Z}\left(-\sigma^2\exp\left\{-\frac{(y - f(x))^2}{\sigma^2}\right\} + \sigma^2\exp\left\{-\frac{(y - f^\star(x))^2}{\sigma^2}\right\}\right)^2 \mathrm{d}\rho \\
&\lesssim \sigma^2\int_\mathcal{Z}\left((y - f(x)) - (y - f^\star(x))\right)^2\mathrm{d}\rho \\
&= \sigma^2\int_\mathcal{X}(f(x) - f^\star(x))^2\mathrm{d}\rho \lesssim \mathbb{E}g,
\end{aligned} \tag{8}$$

where the first inequality is a consequence of the mean value theorem and the boundedness of $\|h'\|$ with $h(t) = -\sigma^2\exp(-t^2/\sigma^2)$, $t \in \mathbb{R}$, and the second inequality is due to Theorem 1. On the other hand, for any $g_1, g_2 \in \mathcal{F}_\mathcal{H}$, there exist $f_1, f_2 \in \mathcal{H}$ such that

Please cite this article in press as: Y. Feng, Y. Ying, Learning with correntropy-induced losses for regression with mixture of symmetric stable noise, Appl. Comput. Harmon. Anal. (2019), https://doi.org/10.1016/j.acha.2019.09.001

$$g_1(z) = -\sigma^2 \exp\left\{-(y - f_1(x))^2/\sigma^2\right\} + \sigma^2 \exp\left\{-(y - f^\star(x))^2/\sigma^2\right\},$$

and

$$g_2(z) = -\sigma^2 \exp\left\{-(y - f_2(x))^2/\sigma^2\right\} + \sigma^2 \exp\left\{-(y - f^\star(x))^2/\sigma^2\right\}.$$

By applying the mean value theorem and noticing again the boundedness of $\|h'\|_\infty$, we have

$$\|g_1 - g_2\|_\infty \leq \sigma^2 \|f_1 - f_2\|_\infty.$$

Under the Complexity Assumption with $0 < s < 2$, the following relation between the ℓ^2-empirical covering numbers of $\mathcal{F}_\mathcal{H}$ and $\mathcal{H}$ holds

$$\log \mathcal{N}_2(\mathcal{F}_\mathcal{H}, \eta) \leq \log \mathcal{N}_2\left(\mathcal{H}, \eta/\sigma^2\right) \lesssim \eta^{-s}.$$

Applying Lemma 1 to the function set $\mathcal{F}_\mathcal{H}$, with simple computations, we come to the conclusion that for any $0 < \delta < 1$ with confidence $1 - \delta$, there holds

$$[\mathcal{E}^\sigma(f) - \mathcal{E}^\sigma(f^\star)] - [\mathcal{E}^\sigma_\mathbf{z}(f) - \mathcal{E}^\sigma_\mathbf{z}(f^\star)] - \frac{1}{2}[\mathcal{E}^\sigma(f) - \mathcal{E}^\sigma(f^\star)] \lesssim \log(1/\delta) n^{-\frac{2}{2+s}}.$$

Noticing that $\mathcal{E}^\sigma_\mathbf{z}(f_\mathbf{z}) \leq \mathcal{E}^\sigma_\mathbf{z}(f^\star)$, we have

$$\frac{1}{2}[\mathcal{E}^\sigma(f_\mathbf{z}) - \mathcal{E}^\sigma(f^\star)] \leq [\mathcal{E}^\sigma(f_\mathbf{z}) - \mathcal{E}^\sigma(f^\star)] - [\mathcal{E}^\sigma_\mathbf{z}(f_\mathbf{z}) - \mathcal{E}^\sigma_\mathbf{z}(f^\star)] - \frac{1}{2}[\mathcal{E}^\sigma(f_\mathbf{z}) - \mathcal{E}^\sigma(f^\star)].$$

Therefore, for any $0 < \delta < 1$ with confidence $1 - \delta$, it holds that

$$\|f_\mathbf{z} - f^\star\|^2_\rho \lesssim \log(1/\delta) n^{-\frac{2}{2+s}}.$$

This completes the proof of Theorem 2. □

Remark 1. From the proof of Theorem 2, we see that the boundedness of the loss function ℓ_σ and the Bernstein condition (8) play a crucial role in establishing fast convergence rates of $f_\mathbf{z}$. The Bernstein condition holds because of the Lipschitz continuity of the loss function ℓ_σ on $\mathbb{R}$ and the fact that the $L^2_{\rho_X}$-distance between $f_\mathbf{z}$ and $f^\star$ can be upper bounded by the excess risk $\mathcal{E}^\sigma(f_\mathbf{z}) - \mathcal{E}^\sigma(f^\star)$, i.e., conclusions in Theorem 1.

3.4. Comments on MCCR with mixture of symmetric stable noise

We now give two remarks on the performance of the MCCR estimator $f_\mathbf{z}$ in the presence of mixture of symmetric stable noise by comparing with that of the least squares estimator.

The first remark is on the convergence rates of the two regression estimators. As shown in Theorem 2, in the presence of mixture of symmetric stable noise and when $f^\star \in \mathcal{H}$, $f_\mathbf{z}$ can learn the unknown truth function $f^\star$ well. The established learning rates are of type $\mathcal{O}(n^{-\frac{2}{2+s}})$ which are optimal in the sense that they are asymptotically of type $\mathcal{O}(n^{-1})$. Moreover, they are comparable with that of least squares estimators [64,11].

Our second remark is on the conditions required to established convergence rates for the two regression schemes. Recalling that for least squares regression, to establish learning theory type convergence rates, the response variable (and consequently the noise, under the data-generating model (1)) is frequently assumed to be uniformly bounded [11,55], which is usually not the case in practice. In fact, even in the presence of Gaussian noise, to establish learning theory type convergence rates for least squares regression, it is much

Please cite this article in press as: Y. Feng, Y. Ying, Learning with correntropy-induced losses for regression with mixture of symmetric stable noise, Appl. Comput. Harmon. Anal. (2019), https://doi.org/10.1016/j.acha.2019.09.001

involved due to the unboundedness of the response variable, in which case many conventional learning theory arguments and tools are not applicable. Recently, some efforts have been made to relax this assumption [59,19,39]. As far as we are aware, convergence rates for least squares regression estimators cannot be established without resorting to the finite-variance condition. When moving our attention to correntropy based regression, as shown above, in the presence of mixture of symmetric stable noise, optimal learning rates of MCCR estimator are established. Notice that symmetric stable noise with the characteristic exponent parameter $0 < \alpha < 2$ has infinite variance or even first-order moment. Moreover, as stated above, it can approximate any density function arbitrarily well with properly chosen K and consequently can be applied to model outliers. In this sense, our study presented here explains the capability of MCCR estimators in dealing with outliers.

4. Simulations

In this section, we provide simulations (1) to validate the feasibility of modeling outliers by using mixture of symmetric stable distributions and (2) to justify the robustness of MCCR to outliers by comparing with that of Huber regression estimators which are regarded as outlier robust.

Concerning the data generating model $Y = f^\star(X) + \varepsilon$, we set the truth function $f^\star$ as the following sinc function

$$f^\star(x) = \sin(\pi x)/(\pi x), x \in [-4, 4],$$

as done in [58,53]. In our simulation studies, we aim to learn $f^\star$ from observations that are contaminated by outliers. In particular, the outliers are generated by mixture of symmetric stable noise as proposed in this study. We consider the following two types of noise that belong to this category:

- Noise I: $\varepsilon \sim 0.9N(0, 0.05^2) + 0.1N(0, 0.5^2)$
- Noise II: $\varepsilon \sim 0.9N(0, 0.05^2) + 0.1\text{Cauchy}(0, 1)$

For Noise I, it is drawn from the mixture of two Gaussian distributions where the background noise is drawn from $N(0, 0.05^2)$ and the contaminating noise is drawn from $N(0, 0.5^2)$ to generate outliers. For Noise II, it is drawn from the mixture of Gaussian and Cauchy distributions where the Gaussian noise $N(0, 0.05^2)$ serves as background noise and outliers are generated by the contaminating noise $\text{Cauchy}(0, 1)$, i.e., Cauchy noise with the location parameter 0 and the scale parameter 1.

We set up our experiment by following that of [18], i.e., the hypothesis space $\mathcal{H}$ is chosen as a subset of a reproducing kernel Hilbert space which is selected automatically by means of a regularized empirical risk minimization, see formula (21) in [18]. A Gaussian kernel is utilized as the reproducing kernel. 200 samples are drawn as training data and 400 samples are drawn as test data. The bandwidth parameter, the regularization parameter, and the scale parameter in Huber's loss are tuned via a five-fold cross validation. The scale parameter σ in the loss function ℓ_σ (2) is set to 0.01.

Experimental results on the generation of outliers and the learned curves are plotted in Figs. 1 and 2. In Fig. 1, the black curves stand for the curve of the truth function $f^\star$. The blue dots from the two panels stand for samples that are contaminated by the background noise of Noise I and Noise II, respectively. The red crosses are samples contaminated by contaminating noise of the two noise types, respectively, which are regarded as outliers. In Fig. 2, the truth curve (black solid line) as well as the curves learned from MCCR (dashed red curve) and from Huber regression (dashed blue curve) are plotted when the noise are of type I and type II, respectively. Outliers are also marked in Fig. 2 for illustration.

Please cite this article in press as: Y. Feng, Y. Ying, Learning with correntropy-induced losses for regression with mixture of symmetric stable noise, Appl. Comput. Harmon. Anal. (2019), https://doi.org/10.1016/j.acha.2019.09.001

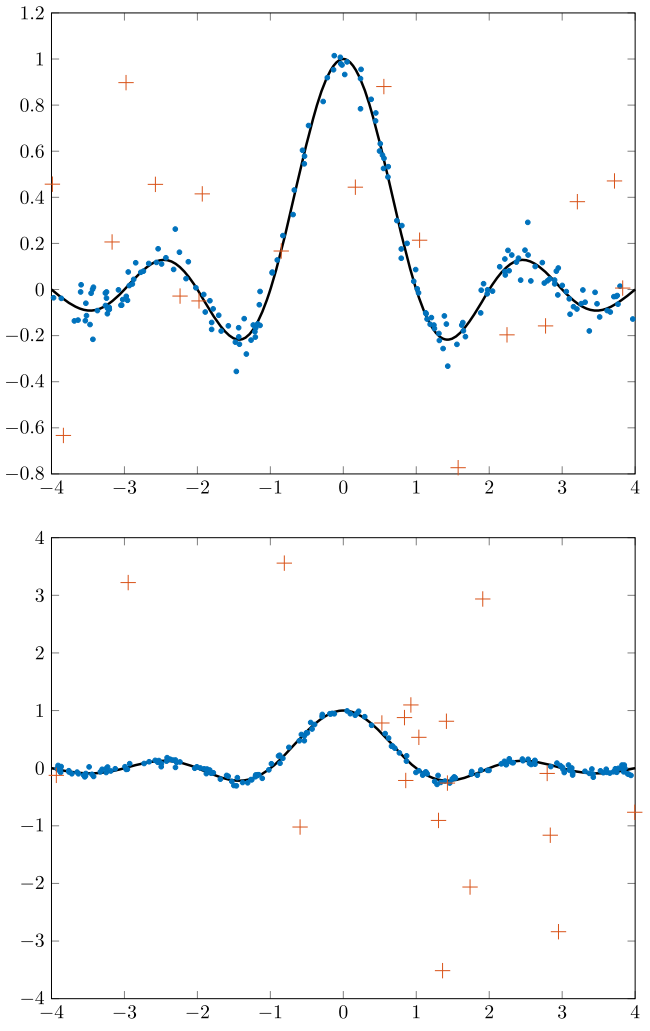

Fig. 1. Sinc function (black solid curves) and training samples. The samples with red crosses are regarded as outliers. (top) The observations are contaminated by mixture of Gaussian noise. (bottom) The observations are contaminated by mixture of Gaussian and Cauchy noise. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

From Fig. 1, it is easy to see that outliers are indeed generated when the noise are drawn from mixture of symmetric stable distributions. According to Fig. 2, MCCR is robust to outliers and performs better than Huber regression in the presence of outliers.

5. Conclusion

In this paper, we studied the correntropy based regression within the statistical learning framework by introducing the mixture of symmetric stable noise which subsume Gaussian noise, Cauchy noise, and mixture of Gaussian noise. In this study, it was introduced to model heavy-tailed noise and outliers, to which the correntropy based regression estimators have been empirically verified to be resistant. In our study, we showed that the empirical risk minimization scheme based on the correntropy induced loss can learn the underlying truth function sufficiently well while allowing the noise to be the mixture of symmetric stable noise. In particular, learning theory analysis was conducted and the learning performance of MCCR with mixture of symmetric stable noise was evaluated. It is interesting to see that, in this case, asymptotically optimal learning rates of type $\mathcal{O}(n^{-1})$ can be developed, which are comparable with that of least squares

Please cite this article in press as: Y. Feng, Y. Ying, Learning with correntropy-induced losses for regression with mixture of symmetric stable noise, Appl. Comput. Harmon. Anal. (2019), https://doi.org/10.1016/j.acha.2019.09.001

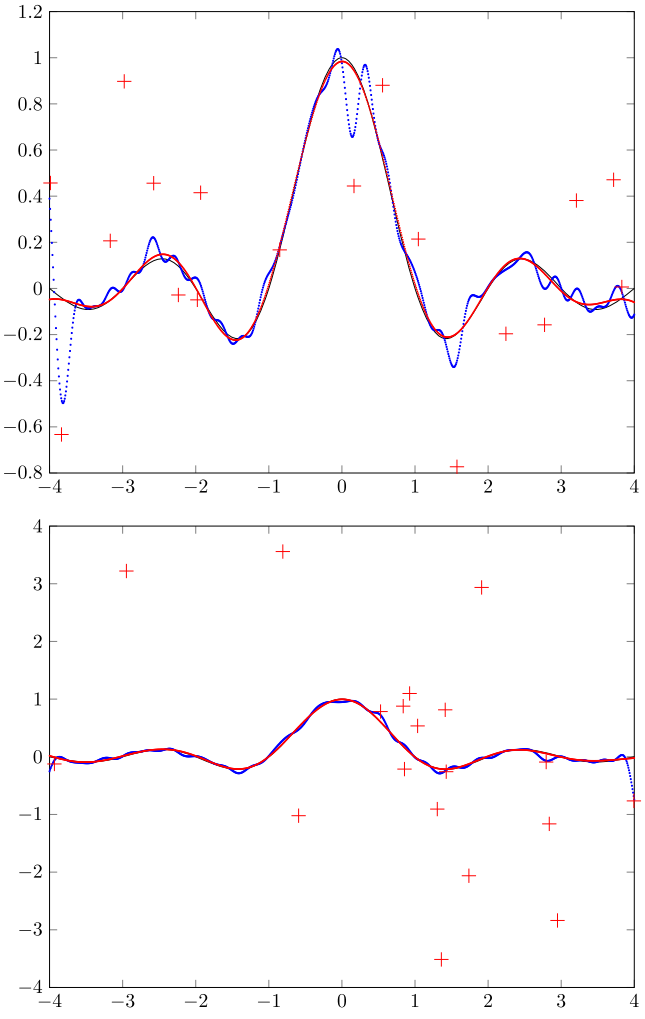

Fig. 2. Outliers (red crosses), sinc function (black solid curves) and its estimators from MCCR and Huber regression (MCCR: red dashed curve; Huber: blue dashed curve). (top) The observations are contaminated by mixture of Gaussian noise. (bottom) The observations are contaminated by mixture of Gaussian and Cauchy noise.

regression under bounded noise assumption. These theoretical findings successfully explain the efficiency and effectiveness of correntropy based regression estimators in the presence of heavy-tailed noise or outliers.

Acknowledgments

The work of Yiming Ying is supported by National Science Foundation (NSF) under Grant No. 1816227.

References

[1] Charu C. Aggarwal, Outlier Analysis, Springer, 2016.
[2] Sachin Ambike, Jacek Ilow, Dimitrios Hatzinakos, Detection for binary transmission in a mixture of Gaussian noise and impulsive noise modeled as an α-stable process, IEEE Signal Process. Lett. 1 (3) (1994) 55–57.
[3] Vic Barnett, Toby Lewis, Outliers in Statistical Data, Wiley, New York, 1994.
[4] Ricardo J. Bessa, Vladimiro Miranda, Joao Gama, Entropy and correntropy against minimum square error in offline and online three-day ahead wind power forecasting, IEEE Trans. Power Syst. 24 (4) (2009) 1657–1666.
[5] Ricardo J. Bessa, Vladimiro Miranda, José C. Príncipe, Audun Botterud, Jun Wang, Information theoretic learning applied to wind power modeling, in: The 2010 International Joint Conference on Neural Networks, IEEE, 2010, pp. 1–8.
[6] Ramon F. Brcich, Robert D. Iskander, Abdelhak M. Zoubir, The stability test for symmetric α-stable distributions, IEEE Trans. Signal Process. 53 (3) (2005) 977–986.

Please cite this article in press as: Y. Feng, Y. Ying, Learning with correntropy-induced losses for regression with mixture of symmetric stable noise, Appl. Comput. Harmon. Anal. (2019), https://doi.org/10.1016/j.acha.2019.09.001

[7] Badong Chen, Xi Liu, Haiquan Zhao, José C. Príncipe, Maximum correntropy Kalman filter, Automatica 76 (2017) 70–77.
[8] Badong Chen, José C. Príncipe, Maximum correntropy estimation is a smoothed MAP estimation, IEEE Signal Process. Lett. 19 (8) (2012) 491–494.
[9] Badong Chen, Lei Xing, Haiquan Zhao, Nanning Zheng, José C. Príncipe, Generalized correntropy for robust adaptive filtering, IEEE Trans. Signal Process. 64 (13) (2016) 3376–3387.
[10] Zhiqiang Chen, David E. Tyler, On the behavior of Tukey's depth and median under symmetric stable distributions, J. Statist. Plann. Inference 122 (1) (2004) 111–124.
[11] Felipe Cucker, Ding-Xuan Zhou, Learning Theory: An Approximation Theory Viewpoint, Cambridge University Press, 2007.
[12] William H. DuMouchel, Stable distributions in statistical inference: 1. Symmetric stable distributions compared to other symmetric long-tailed distributions, J. Amer. Statist. Assoc. 68 (342) (1973) 469–477.
[13] Yonina C. Eldar, Arie Yeredor, Finite-memory denoising in impulsive noise using Gaussian mixture models, IEEE Trans. Circuits Syst. II, Analog Digit. Signal Process. 48 (11) (2001) 1069–1077.
[14] Eugene F. Fama, Richard Roll, Some properties of symmetric stable distributions, J. Amer. Statist. Assoc. 63 (323) (1968) 817–836.
[15] Eugene F. Fama, Richard Roll, Parameter estimates for symmetric stable distributions, J. Amer. Statist. Assoc. 66 (334) (1971) 331–338.
[16] Jun Fan, Ting Hu, Qiang Wu, Ding-Xuan Zhou, Consistency analysis of an empirical minimum error entropy algorithm, Appl. Comput. Harmon. Anal. 41 (1) (2016) 164–189.
[17] Yunlong Feng, Jun Fan, Johan A.K. Suykens, A statistical learning approach to modal regression, arXiv preprint, arXiv: 1702.05960, 2017.
[18] Yunlong Feng, Xiaolin Huang, Lei Shi, Yuning Yang, Johan A.K. Suykens, Learning with the maximum correntropy criterion induced losses for regression, J. Mach. Learn. Res. 16 (2015) 993–1034.
[19] Zheng-Chu Guo, Ding-Xuan Zhou, Concentration estimates for learning with unbounded sampling, Adv. Comput. Math. 38 (1) (2013) 207–223.
[20] Frank R. Hampel, Elvezio M. Ronchetti, Peter J. Rousseeuw, Werner A. Stahel, Robust Statistics: The Approach Based on Influence Functions, John Wiley & Sons, 2011.
[21] Jiawei Han, Jian Pei, Micheline Kamber, Data Mining: Concepts and Techniques, Elsevier, 2011.
[22] Erion Hasanbelliu, Luis Sanchez Giraldo, José C. Príncipe, Information theoretic shape matching, IEEE Trans. Pattern Anal. Mach. Intell. 36 (12) (2014) 2436–2451.
[23] Douglas M. Hawkins, Identification of Outliers, Springer, 1980.
[24] Ran He, Bao-Gang Hu, Wei-Shi Zheng, Xiang-Wei Kong, Robust principal component analysis based on maximum correntropy criterion, IEEE Trans. Image Process. 20 (6) (2011) 1485–1494.
[25] Ran He, Tieniu Tan, Liang Wang, Wei-Shi Zheng, $\ell_{2,1}$-regularized correntropy for robust feature selection, in: IEEE Conference on Computer Vision and Pattern Recognition, 2012, IEEE, 2012, pp. 2504–2511.
[26] Ran He, Wei-Shi Zheng, Bao-Gang Hu, Maximum correntropy criterion for robust face recognition, IEEE Trans. Pattern Anal. Mach. Intell. 33 (8) (2011) 1561–1576.
[27] Ting Hu, Jun Fan, Qiang Wu, Ding-Xuan Zhou, Learning theory approach to minimum error entropy criterion, J. Mach. Learn. Res. 14 (2013) 377–397.
[28] Peter J. Huber, Robust estimation of a location parameter, Ann. Math. Stat. 35 (1) (1964) 73–101.
[29] Peter J. Huber, Elvezio Ronchetti, Robust Statistics, Wiley, 2009.
[30] Jacek Ilow, Dimitrios Hatzinakos, Analytic α-stable noise modeling in a Poisson field of interferers or scatterers, IEEE Trans. Signal Process. 46 (6) (1998) 1601–1611.
[31] Jana Jurečková, Jan Picek, Robust Statistical Methods with R, CRC Press, 2005.
[32] Saleem A. Kassam, Vincent H. Poor, Robust techniques for signal processing: a survey, Proc. IEEE 73 (3) (1985) 433–481.
[33] Bart Kosko, Sanya Mitaim, Robust stochastic resonance: signal detection and adaptation in impulsive noise, Phys. Rev. E 64 (5) (2001) 051110.
[34] Richard J. Kozick, Brian M. Sadler, Maximum-likelihood array processing in non-Gaussian noise with Gaussian mixtures, IEEE Trans. Signal Process. 48 (12) (2000) 3520–3535.
[35] Ercan E. Kuruoglu, William J. Fitzgerald, Peter J.W. Rayner, Near optimal detection of signals in impulsive noise modeled with a symmetric α-stable distribution, IEEE Commun. Lett. 2 (10) (1998) 282–284.
[36] Weifeng Liu, Puskal P. Pokharel, José C. Príncipe, Correntropy: properties and applications in non-Gaussian signal processing, IEEE Trans. Signal Process. 55 (11) (2007) 5286–5298.
[37] Marco J. Lombardi, Simon J. Godsill, On-line Bayesian estimation of signals in symmetric α-stable noise, IEEE Trans. Signal Process. 54 (2) (2006) 775–779.
[38] Geoffrey J. McLachlan, Kaye E. Basford, Mixture Models: Inference and Applications to Clustering, Marcel Dekker, 1988.
[39] Shahar Mendelson, Learning without concentration, J. ACM 62 (21) (2015).
[40] Joana Mendes, Ricardo J. Bessa, Hrvoje Keko, Jean Sumaili, Valdimiro Miranda, Carlos Ferreira, Joao Gama, Audun Botterud, Zhi Zhou, Jianhui Wang, Development and Testing of Improved Statistical Wind Power Forecasting Methods, Technical report, Argonne National Laboratory (ANL), 2011.
[41] Grady Miller, Properties of certain symmetric stable distributions, J. Multivariate Anal. 8 (3) (1978) 346–360.
[42] Chrysostomos L. Nikias, Min Shao, Signal Processing with Alpha-Stable Distributions and Applications, Wiley-Interscience, 1995.
[43] Jintae Park, Georgy Shevlyakov, Kiseon Kim, Maximin distributed detection in the presence of impulsive α-stable noise, IEEE Trans. Wirel. Commun. 10 (6) (2011) 1687–1691.
[44] José C. Príncipe, Information Theoretic Learning: Renyi's Entropy and Kernel Perspectives, Springer Science & Business Media, 2010.

Please cite this article in press as: Y. Feng, Y. Ying, Learning with correntropy-induced losses for regression with mixture of symmetric stable noise, Appl. Comput. Harmon. Anal. (2019), https://doi.org/10.1016/j.acha.2019.09.001

[45] Adithya Rajan, Cihan Tepedelenlioglu, Diversity combining over Rayleigh fading channels with symmetric α-stable noise, IEEE Trans. Wirel. Commun. 9 (9) (2010) 2968–2976.
[46] Peter J. Rousseeuw, Annick M. Leroy, Robust Regression and Outlier Detection, John Wiley & Sons, 2005.
[47] Gennady Samorodnitsky, Murad S. Taqqu, Stable Non-Gaussian Random Processes: Stochastic Models with Infinite Variance, CRC Press, 1994.
[48] Ignacio Santamaría, Puskal P. Pokharel, José C. Príncipe, Generalized correlation function: definition, properties, and application to blind equalization, IEEE Trans. Signal Process. 54 (6) (2006) 2187–2197.
[49] Bernhard Schölkopf, Alexander J. Smola, Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond, MIT Press, 2001.
[50] Min Shao, Chrysostomos L. Nikias, Signal processing with fractional lower order moments: stable processes and their applications, Proc. IEEE 81 (7) (1993) 986–1010.
[51] Lei Shi, Yunlong Feng, Ding-Xuan Zhou, Concentration estimates for learning with ℓ^1-regularizer and data dependent hypothesis spaces, Appl. Comput. Harmon. Anal. 31 (2) (2011) 286–302.
[52] Abhishek Singh, Rosha Pokharel, José C. Príncipe, The C-loss function for pattern classification, Pattern Recognit. 47 (1) (2014) 441–453.
[53] Alex J. Smola, Bernhard Schölkopf, A tutorial on support vector regression, Stat. Comput. 14 (3) (2004) 199–222.
[54] Michael R. Souryal, Erik G. Larsson, Bojan Peric, Branimir R. Vojcic, Soft-decision metrics for coded orthogonal signaling in symmetric α-stable noise, IEEE Trans. Signal Process. 56 (1) (2008) 266–273.
[55] Ingo Steinwart, Andreas Christmann, Support Vector Machines, Springer, New York, 2008.
[56] Michael D. Titterington, Adrian F.M. Smith, Udi E. Makov, Statistical Analysis of Finite Mixture Distributions, Wiley, 1985.
[57] John W. Tukey, A survey of sampling from contaminated distributions, Contrib. Probab. Statist. 2 (1960) 448–485.
[58] Vladimir Vapnik, Statistical Learning Theory, Wiley-Interscience, 1998.
[59] Cheng Wang, Ding-Xuan Zhou, Optimal learning rates for least squares regularized regression with unbounded sampling, J. Complexity 27 (1) (2011) 55–67.
[60] Jim Jing-Yan Wang, Xiaolei Wang, Xin Gao, Non-negative matrix factorization by maximizing correntropy for cancer clustering, BMC Bioinform. 14 (1) (2013) 107.
[61] Lingfeng Wang, Chunhong Pan, Robust level set image segmentation via a local correntropy-based K-means clustering, Pattern Recognit. 47 (5) (2014) 1917–1925.
[62] Ying Wang, Chunhong Pan, Shiming Xiang, Feiyen Zhu, Robust hyperspectral unmixing with correntropy-based metric, IEEE Trans. Image Process. 24 (11) (2015) 4027–4040.
[63] Yulong Wang, Yuan Yan Tang, Luoqing Li, Correntropy matching pursuit with application to robust digit and face recognition, IEEE Trans. Cybern. 47 (6) (2017) 1354–1366.
[64] Qiang Wu, Yiming Ying, Ding-Xuan Zhou, Learning rates of least-square regularized regression, Found. Comput. Math. 6 (2) (2006) 171–192.
[65] Qiang Wu, Yiming Ying, Ding-Xuan Zhou, Multi-kernel regularized classifiers, J. Complexity 23 (1) (2007) 108–134.
[66] Guibiao Xu, Bao-Gang Hu, José C. Príncipe, Robust C-loss kernel classifiers, IEEE Trans. Neural Netw. Learn. Syst. 29 (3) (2018) 510–522.
[67] Fei Zhu, Abderrahim Halimi, Paul Honeine, Badong Chen, Nanning Zheng, Correntropy maximization via ADMM: application to robust hyperspectral unmixing, IEEE Trans. Geosci. Remote Sens. 55 (9) (2017) 4944–4955.
[68] Cuiming Zou, Kit Ian Kou, Robust signal recovery using the prolate spherical wave functions and maximum correntropy criterion, Mech. Syst. Signal Process. 104 (2018) 279–289.

Please cite this article in press as: Y. Feng, Y. Ying, Learning with correntropy-induced losses for regression with mixture of symmetric stable noise, Appl. Comput. Harmon. Anal. (2019), https://doi.org/10.1016/j.acha.2019.09.001