Advances in Applied Mathematics 应用数学进展, 2020, 9(12), 2283-2291
Published Online December 2020 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2020.912267

Hans汉斯

p-Huber损失函数及其鲁棒性研究

俞搏天

浙江师范大学数学系，浙江 金华
Email: 995521063@qq.com

收稿日期：2020年11月27日；录用日期：2020年12月22日；发布日期：2020年12月30日

摘　要

由于应用领域真实数据的复杂性，数据常常受到离群值的污染，因此研究对离群值具有鲁棒性的统计机器学习算法就显得越来越重要。本文在Huber损失的基础上提出了一种更具鲁棒性的非凸*p*-Huber损失函数，通过仿真实验比较了基于*p*-Huber损失的回归学习算法与基于L1损失、Huber损失、MCCR损失的回归算法的拟合效果。数据实验结果显示基于*p*-Huber损失函数的回归学习算法在有离群值的情况下其预测效果要优于常见的损失函数的预测效果。

关键词

回归问题，损失函数，离群值，鲁棒性

p-Huber Loss Functions and Its Robustness

Botian Yu

Department of Mathematics, Zhejiang Normal University, Jinhua Zhejiang
Email: 995521063@qq.com

Received: Nov. 27th, 2020; accepted: Dec. 22nd, 2020; published: Dec. 30th, 2020

Abstract

The data is often contaminated by outliers because of the complexity of the real data in the real applications. Hence it is getting more and more important to invent some statistical machine learning algorithms that are robust to outliers. In this paper, we propose a robust and non-convex *p*-Huber loss function based on the Huber loss. In the numerical analysis, the fitting effect of regression learning algorithm based on *p*-Huber loss and regression algorithm based on L1 loss, Huber loss and MCCR loss are compared. The numerical results show that the *p*-Huber loss function outperforms all of other common loss functions mentioned in the paper when there are outliers in the data.

文章引用: 俞搏天. *p*-Huber 损失函数及其鲁棒性研究[J]. 应用数学进展, 2020, 9(12): 2283-2291.
DOI: 10.12677/aam.2020.912267

Keywords

Regression Problem, Loss Function, Outlier, Robustness

1. 引言

在回归问题中，我们常遇到数据具有噪声或离群值的问题，即样本中的个别值其数值明显偏离它所属样本的其余观测值。一般情况下离群值产生的原因可以分为两大类：人为原因和自然原因。比如输入的时候多输入了一个 0 发生了错误，使得原本的年薪 5 万变成了 50 万，这是人为原因。一个年级前 10 名学生的成绩远远高于其他人，且这些数据没有任何错误，这是自然原因所导致的离群值。这些离群值的出现会增加误差，降低统计检验的能力，可能对具有实质性意义的估计产生偏见或影响，还会影响回归、方差统计和其他模型的基本假设。因此在数据有离群值扰动的情况下，我们要采用对离群值鲁棒的统计方法来分析数据。所谓的鲁棒性是指一个系统在受到扰动的情况下仍然能保持其功能。本文主要研究回归问题中的鲁棒性问题，我们将提出一个新的鲁棒的损失函数来提高回归算法的鲁棒性。

假设数据由下列模型生成：

$$Y = f^*(X) + \varepsilon,\quad \mathrm{E}(\varepsilon | X = x) = 0, \tag{1.1}$$

其中 $X \in \mathcal{X}$ 为解释变量，$\mathcal{X}$ 为可分度量空间，$Y \in \mathcal{Y} = \mathbb{R}$ 为响应变量。回归问题旨在通过模型(1.1)生成的观测值来训练 f^* 的一个估计 $f: \mathcal{X} \to \mathcal{Y}$。假设 $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ 上的概率分布为 ρ。在回归模型(1.1)下，量化估计函数 f 回归效率的最常用的损失函数是均方误差。均方误差也称 L2 损失，如图 1(a)，它是一种常用的损失函数，其定义如下：

$$\phi^{\mathrm{L2}}(y, f(x)) = (y - f(x))^2。 \tag{1.2}$$

基于 L2 损失的回归算法的缺点是其最优性在很大程度上依赖于高斯假设。而在许多实际应用中，数据往往比较复杂，可能会被非高斯噪声或离群值污染，若在这种情况下使用对离群值不鲁棒的损失函数，则可能使整个统计分析无用[1]。因此为了克服 L2 损失不鲁棒的特性，我们需要引进鲁棒的损失函数。

平均绝对误差也称 L1 损失，如图 1(b)，是另一种用于回归模型的损失函数，其定义如下：

$$\phi^{\mathrm{L1}}(y, f(x)) = |y - f(x)|。 \tag{1.3}$$

L1 损失比 L2 损失更具鲁棒性，但它的中心点是折点，因此不能求导，从而不易于求解。

P. J. Huber 等[2]在 1964 年提出了 Huber 损失函数，如图 1(c)，其定义如下：

$$\phi^{\mathrm{Huber}}(y, f(x)) = \begin{cases} 0.5(y - f(x))^2, & |y - f(x)| < \delta, \\ \delta |y - f(x)| - 0.5\delta^2, & |y - f(x)| \geq \delta, \end{cases} \tag{1.4}$$

其中 δ 是一个非负参数，用来控制损失函数二次与线性的范围。当残差大于等于 δ 时采用 L1 损失，残差小于 δ 时采用 L2 损失。Huber 损失函数结合了 L1 损失和 L2 损失的优点，它对离群值比 L2 损失更鲁棒，同时在中心点处可导。R. Girshick [3]在 2015 年采用 Smooth L1 损失函数在 Fast R-CNN 上取得了很好的效果，Smooth L1 是 Huber 损失中 $\delta = 1$ 的特殊情况。

冯云龙等[4]在 2015 年提出了基于最大相关熵[5]的 MCCR 损失函数，如图 1(d)，其定义如下：

$$\phi^{\mathrm{MCCR}}\left(y, f\left(x\right)\right)=\sigma^2\left(1-\mathrm{e}^{\frac{(y-f(x))^2}{\sigma^2}}\right), \tag{1.5}$$

其中 σ 是一个非负的尺度参数。该损失函数对带非高斯噪声的回归问题有帮助，对高斯噪声也能取得很好的表现。

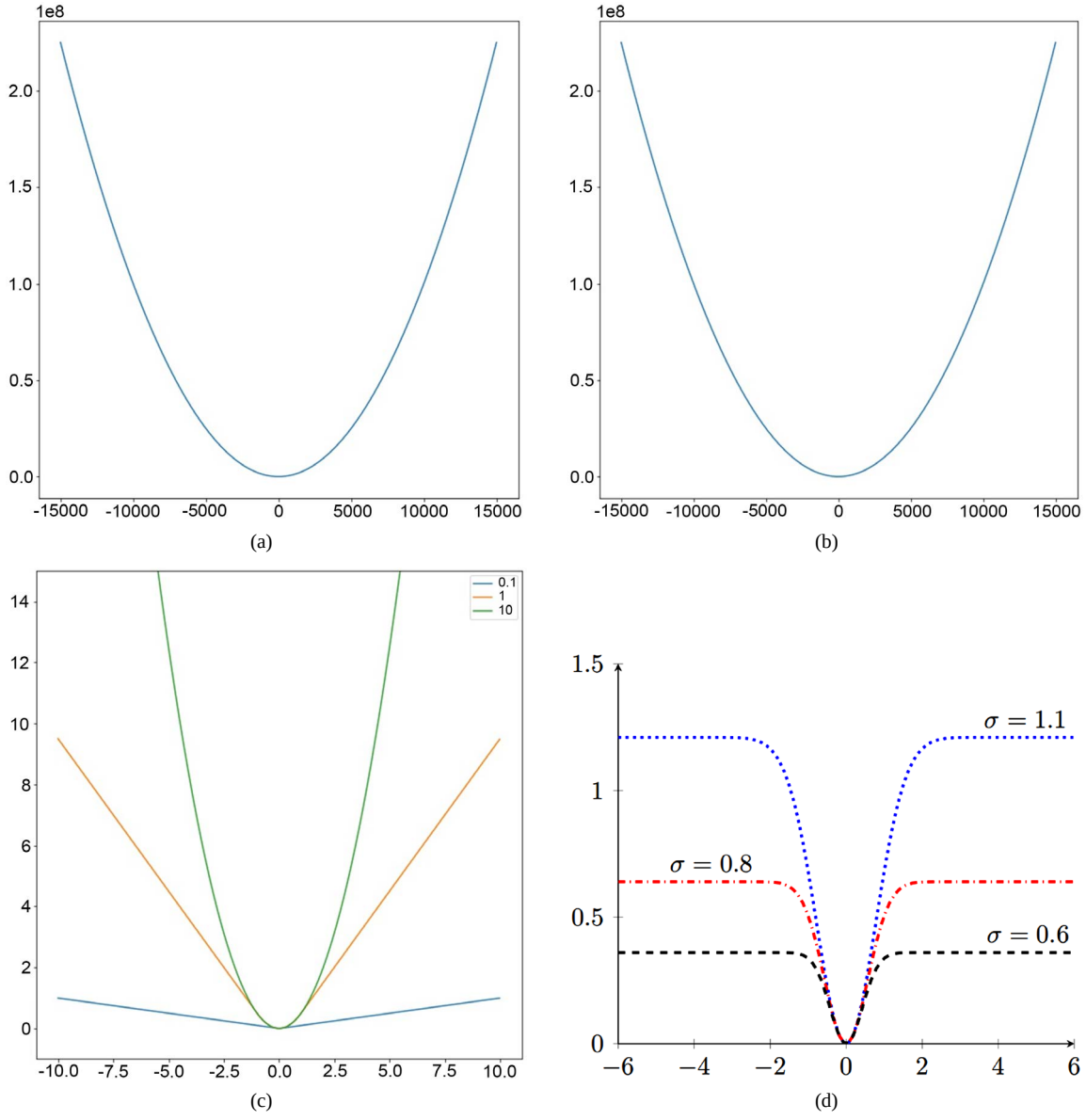

Figure 1. Diagram of four loss functions
图 1. 四种损失函数示意图

本文为进一步提高回归算法的鲁棒性，提出了一种新的基于 Huber 损失函数的非凸鲁棒损失函数 *p*-Huber 损失，该损失函数由参数 p 和 δ 控制。具体定义和性质在第二节中讨论。在第三节中，我们对采用 *p*-Huber 损失的模型进行仿真实验，并与 L1、Huber、MCCR 三种损失函数从预测效果等方面进行了比较。

2. *p*-Huber 损失函数及回归算法

2.1. *p*-Huber 损失函数

在 Huber 损失函数的基础上我们提出了如下新的非凸鲁棒损失函数 *p*-Huber 回归损失，其定义如下：

$$\phi^{p\text{-Huber}}\left(y, f\left(x\right)\right)=\begin{cases}\left(y-f\left(x\right)\right)^2, & \left|y-f\left(x\right)\right|<\delta,\\ \dfrac{2\delta^{2-p}}{p}\left|y-f\left(x\right)\right|^p-\dfrac{2-p}{p}\delta^2, & \left|y-f\left(x\right)\right|\geq\delta,\end{cases} \tag{2.1}$$

其中 p 和 δ 是两个非负的参数。当 $p=1$ 时，损失函数与 Huber 损失函数相同(只相差一个常数系数)。当 $p=2$ 时，损失函数与经典的 L2 损失相同。

如图 2(a)所示：参数 p 主要控制 *p*-Huber 损失函数的凹凸性，当 $p<1$ 时图像为非凸，当 $p>1$ 时图像为凸，本文主要研究 $p<1$ 时非凸的情况。图 2(b)显示参数 δ 主要控制 *p*-Huber 损失函数的拐点。*p*-Huber 损失函数桥接了 L2 损失函数和鲁棒性损失函数。

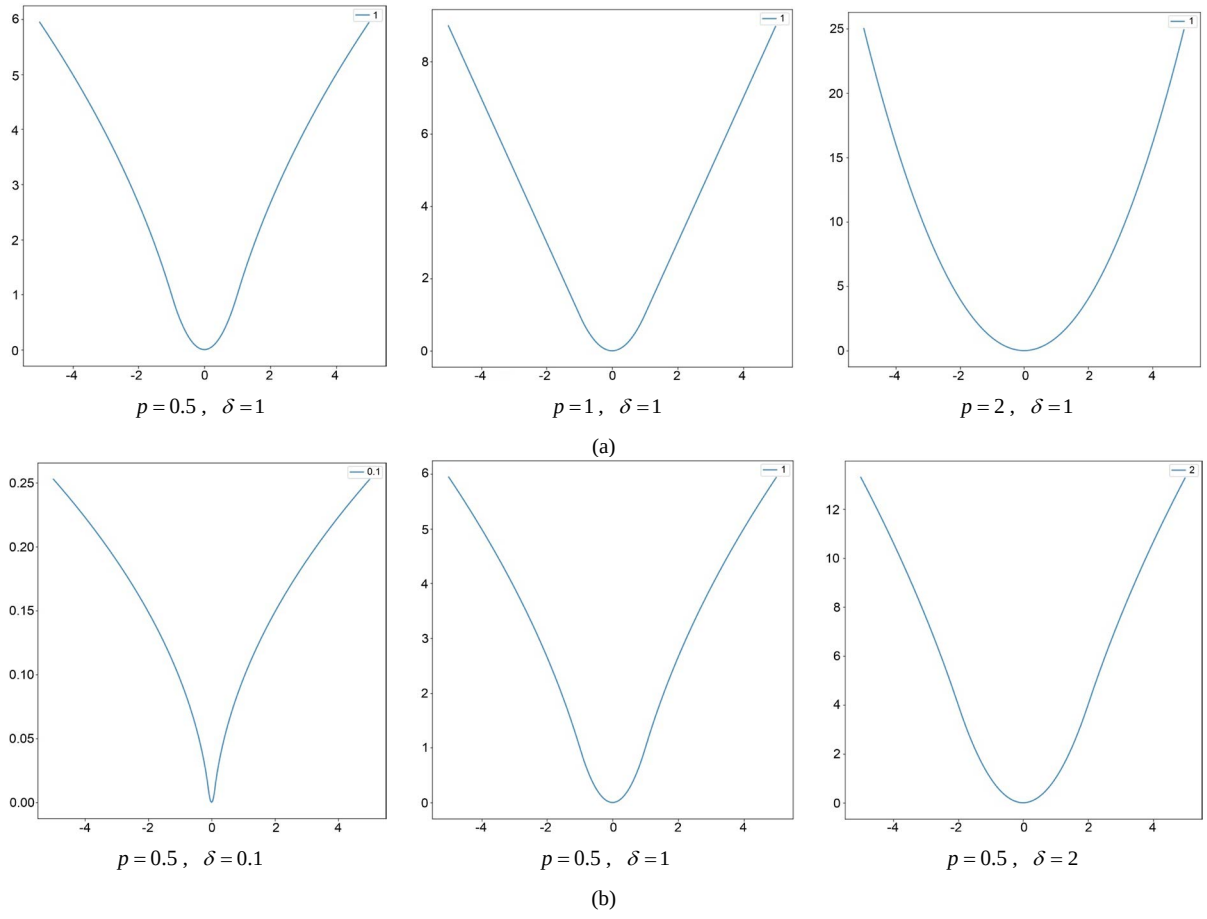

Figure 2. Diagram of different parameters of *p*-Huber loss
图 2. *p*-Huber 损失不同参数示意图

2.2. 基于 *p*-Huber 损失的学习算法

给定一个容量为 m 的独立同分布的样本 $z=\left\{\left(x_1,y_1\right),\cdots,\left(x_m,y_m\right)\right\}$，对于任何估计 $f:\mathcal{X}\to\mathcal{Y}$，其在样本点上导致的经验误差为：

$$\mathcal{E}_z(f)=\frac{1}{m}\sum_{i=1}^{m}\phi^{p\text{-Huber}}\left(y_i,f\left(x_i\right)\right)。$$

本文考虑下列经验风险最小化学习算法：

$$f_z=\arg\min_{f\in\mathcal{H}}\frac{1}{m}\sum_{i=1}^{m}\phi^{p\text{-Huber}}\left(y_i,f\left(x_i\right)\right), \tag{2.2}$$

其中假设空间 $\mathcal{H}$ 是 $C(\mathcal{X})$ 的紧子集。$C(\mathcal{X})$ 是以 $\|f\|_\infty=\sup_{x\in\mathcal{X}}|f(x)|$ 为范数的所有定义在 $\mathcal{X}$ 上的连续函数的巴拿赫空间。$\mathcal{H}$ 的紧性可保证估计函数 f_z 的存在性。模型(2.2)是一个带约束的最小化问题。当 $\mathcal{H}$ 取成由 Mercer 核 $\mathcal{K}$ 生成的某个再生核希尔伯特空间[6] $\mathcal{H}_\mathcal{K}$ 的有界子集时，模型(2.2)和下列不带约束的正则化学习算法是等价的：

$$f_z=\arg\min_{f\in\mathcal{H}_\mathcal{K}}\frac{1}{m}\sum_{i=1}^{m}\phi^{p\text{-Huber}}\left(y_i,f\left(x_i\right)\right)+\lambda\|f\|_\mathcal{K}^2, \tag{2.3}$$

其中 λ 是非负的正则化参数，用以防止算法发生过拟合。

表示定理[7]告诉我们只需要在以下函数集中最小化模型(2.3)：

$$\mathcal{H}_\mathcal{K}=\left\{\sum_{i=1}^{m}\alpha_i\mathcal{K}(x,x_i)+b,b\in\mathbb{R},\alpha_i\in\mathbb{R},i=1,\cdots,m\right\},$$

这里 b 是一个偏移量。在第三节仿真实验中我们采用如下高斯核：

$$\mathcal{K}_h\left(x_i,x_j\right)=\exp\left(-\left\|x_i-x_j\right\|^2/h^2\right),$$

其中参数 h 是待定的高斯核的尺度参数。

3. 仿真实验

3.1. 实验平台及噪声介绍

软件环境：

操作系统：Windows 10 家庭版 x64 位

开发工具：Python 3、Matlab 2015b

硬件环境：

处理器：Inter i7-6700HQ@2.60 GHz

内存：8 GB

在我们的实验中，按如下方式添加噪声：

$$\text{noise}:=\tau_1\varepsilon_1+\tau_2\varepsilon_2^p, \tag{3.1}$$

其中 ε_1 遵循标准高斯分布，ε_2^p 是一个脉冲噪声(离群值)，定义如下：

$$\text{Prob}\left(\varepsilon_2^p=t\right)=\begin{cases}1-p, & t=0,\\ p/2, & t=1,\\ p/2, & t=-1.\end{cases}$$

这里引入 τ_1 和 τ_2 来设置高斯噪声的方差和脉冲噪声的大小。在我们的实验中，我们始终设置 $p=0.1$，即 10%的样本被脉冲噪声污染。

3.2. 辛格函数下的评估

为了展现 p-Huber 损失函数的有效性，我们将其预测精度分别与 L1 损失、Huber 损失、MCCR 损失

进行比较。首先我们选择辛格函数作为回归函数。辛格函数[8] [9]经常被用来举例说明回归函数。一维辛格函数的表达式如下，图像见图 3。

$$f(x)=\sin(\pi x)/(\pi x), x\in[-4,4]。\tag{3.2}$$

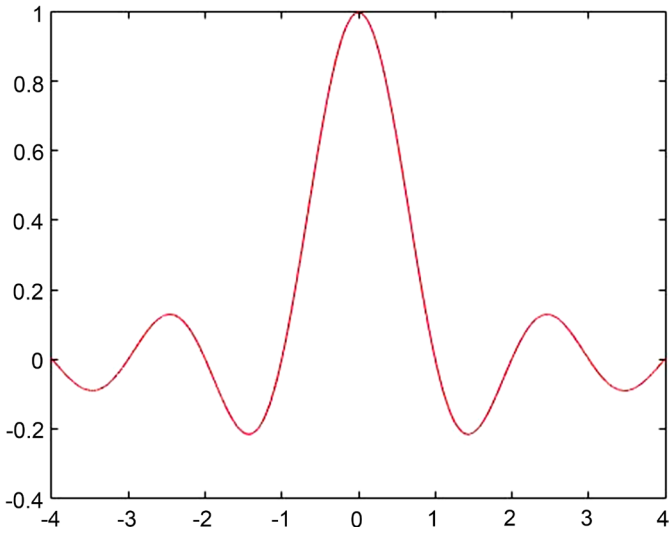

Figure 3. One-dimensional Sincfunction graph
图 3. 一维辛格函数图像

在实验中，首先从被高斯噪声破坏的辛格函数中抽取规模大小为 100 的训练集，然后我们抽取了另一个同样规模的被高斯噪声和离群值污染的训练集。在每个训练集上对辛格函数的拟合结果绘制在图 4 和图 5 中，其中黑色实线表示辛格函数，被高斯噪声污染的训练集用加号来表示，离群值用正方形来标记，橙色虚线代表 *p*-Huber 拟合的曲线，红色虚线代表 MCCR 拟合的曲线，蓝色虚线代表 Huber 拟合的曲线，绿色虚线代表 L1 拟合的曲线。

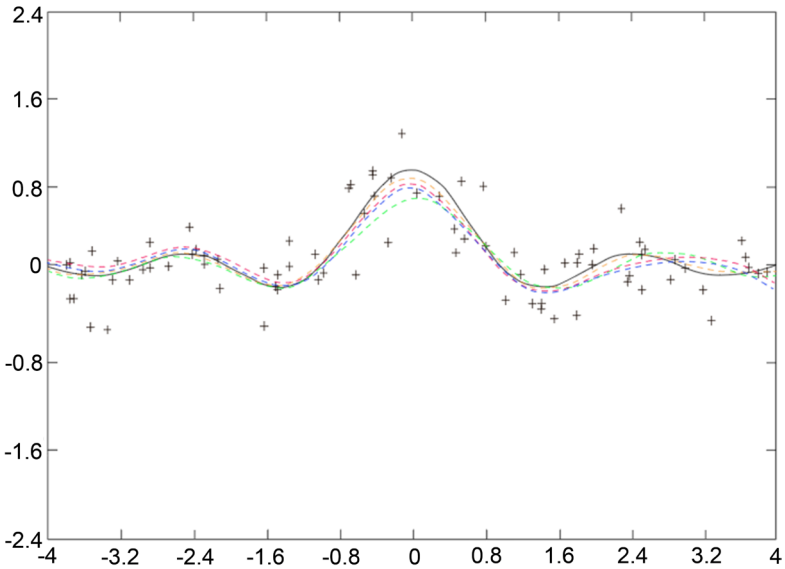

Figure 4. The fitting effect diagram of different models to the Sinc function when contaminated by Gaussian noise
图 4. 高斯噪声污染时不同模型对辛格函数的拟合效果图

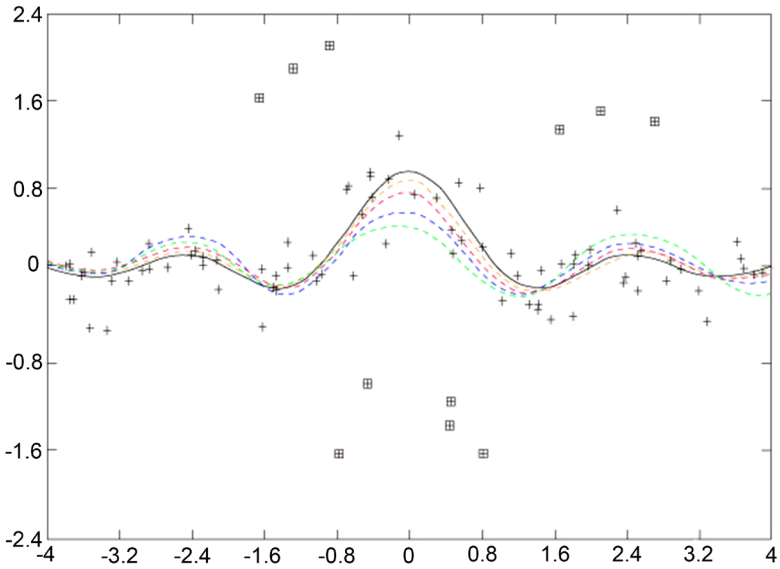

Figure 5. The fitting effect diagram of different models to the Sinc function when contaminated by Gaussian noise and outliers
图 5. 高斯噪声和离群值污染时不同模型对辛格函数的拟合效果图

从图 4 中我们可以看出，当数据仅受高斯噪声污染时，这四个模型都能很好的拟合函数图像。从图 5 中我们可以看出，当数据受到高斯噪声和离群值的污染时，这四个模型都能成功拟合曲线，但是 p-Huber 拟合效果最好。

下面我们通过实验来分析在加入噪声和离群值的情况下 p-Huber 损失函数的参数 p 和 δ 对算法的影响。经过数值实验，对于拟合结果我们绘制了如下表格。在表格里，我们记录了数据集的平方误差相对和，其定义如下：

$$\mathrm{RSSE}\left(\hat{f}\right)=\sum_{x\in T}\left(f\left(x\right)-\hat{f}\left(x\right)\right)^2\Big/\sum_{x\in T}\left(f\left(x\right)-\bar{f}_T\right)^2, \tag{3.3}$$

其中 $\bar{f}_T$ 表示 $f\left(x\right)$ 在数据集 T 上的均值。

Table 1. When $\delta=0.7$, different values p predict the result of Sinc function
表 1. $\delta=0.7$ 时，不同的取值 p 对辛格函数预测的结果

p	0.03	0.53	1.03	1.53	2.03
RSSE	0.003112	0.004117	0.004439	0.004560	0.004698
p	2.53	3.03	3.53	4.03	4.53
RSSE	0.004778	0.004999	0.005154	0.005325	0.005504

Table 2. When $p=2.53$, different values δ predict the result of Sinc function
表 2. $p=2.53$ 时，不同 δ 的取值对辛格函数预测的结果

δ	0.7	1.7	2.7	3.7	4.7
RSSE	0.004698	0.004775	0.004775	0.004781	0.004781
δ	5.7	6.7	7.7	8.7	9.7
RSSE	0.004775	0.004780	0.004783	0.004777	0.004779

表 1 显示当固定 δ 时，随着参数 p 的变化，实验拟合效果也不相同，这说明参数 p 对鲁棒性有影响。表 2 显示当固定 p 时，随着参数 δ 的变化，实验拟合结果几乎相同，这说明参数 δ 对鲁棒性几乎没有影响，δ 主要用来控制损失函数的拐点。

3.3. 弗里德曼基准函数下的评估

我们接下来考虑多维回归问题，选择了弗里德曼基准函数作为回归函数。弗里德曼基准函数[10]在研究回归问题时已被广泛采用，其定义如下：

$$f_1(x)=10\sin\left(\pi x^1x^2\right)+20\left(x^3-0.5\right)^2+10x^4+5x^5;$$

$$f_2(x)=\sqrt{\left(x^1\right)^2+\left(x^2x^3-1\Big/\left(x^2x^4\right)\right)^2};$$

$$f_3(x)=\arctan\left(1\Big/x^1\left(x^2x^3-1\Big/\left(x^2x^4\right)\right)\right)。$$

对于 f_1，$x=\left(x^1,\cdots,x^{10}\right)$，其中 $x^j, j=1,\cdots,10$ 服从 $[0,1]$ 上的均匀分布，$x^6,\cdots,x^{10}$ 是噪声变量。对于 f_2 和 f_3，$x=\left(x^1,x^2,x^3,x^4\right)$ 的每个坐标分别服从下列区间上的均匀分布：$x^1\in[0,100]$，$x^2\in[40\pi,560\pi]$，$x^3\in[0,1]$，$x^4\in[1,11]$。

对于每个函数，我们随机抽取 1000 个观测值进行训练和交叉验证，另外随机独立抽取 1000 个观测值作为测试集，然后根据(3.1)的方法添加噪声和离群值，一组加入了高斯噪声，另一组加入了高斯噪声和离群值。对于 f_1 我们设置 $\tau_1=1$。对于 f_2 和 f_3，设置 τ_1 使得信号功率与 ε_1 功率之比为 3。对于无离群值的情况下，我们设置 $\tau_2=0$。对于训练集中存在离群值的四个模型，我们设置 $\tau_2=\max_{x\in D}f(x)-\min_{x\in D}f(x)$，其中 D 是每个基准函数的定义域。对于每个回归模型，都通过均方误差下的 10 倍交叉验证对高斯核的宽度 h，正则化参数 λ 和损失函数中的比例参数(LAD 损失没有比例参数)进行了调整。我们将其残差 $\left\{y_i-f\left(x_i\right)\right\}_{i=1}^{100}$ 都记录下来。对于每个回归模型，其平方误差相对和记录在表 3。

Table 3. The prediction results of different models on Friedman's benchmark function
表 3. 不同模型对弗里德曼基准函数的预测结果

测试函数	有无离群值	L1	Huber	MCCR	p-Huber
f_1	无	0.5482	0.5481	0.5419	0.4761
	有	0.5512	0.5513	0.5505	0.5062
f_2	无	0.2824	0.2831	0.2972	0.0717
	有	0.2836	0.2841	0.2991	0.0849
f_3	无	0.0127	0.0127	0.0128	0.0036
	有	0.0163	0.0163	0.0163	0.0046

表 3 中我们可以看出，当数据仅受到高斯噪声污染时，这四个模型都能很好的拟合。但当数据被离群值破坏时，所有四个模型虽然仍能拟合，但是p-Huber 拟合效果相比其它三个更优。

3.4. 真实数据集下的评估

本节中我们用四个真实数据集对 p-Huber 损失的拟合效果进行评估。这四个真实数据集中一个来自 2018 年 12 月至 2019 年 12 月股票交易历史数据，另外三个来自机器学习数据集中的 UCI 数据集：Airfoil Self-Noise 数据集，Concrete Compressive Strength 数据集，Yacht Hydrodynamics 数据集。

我们将 2/3 数据用作了训练，其余用作测试，并多次实验，其准确性由平方误差相对和来测量。实验结果记录在表 4，其中包含了数据特征的大小 d 和数量 m。

Table 4. The prediction results of different models on real data
表 4. 不同模型对真实数据的预测结果

数据集	d	m	L1	Huber	MCCR	p-Huber
stock market	7	251	5.086e−9	5.085e−9	5.150e−9	5.050e−9
airfoil	6	1053	2.468e−7	4.948e−7	1.131e−6	2.074e−7
concrete	9	1030	0.010043	0.010356	0.010031	0.010006
yacht	7	308	0.0030	0.0039	0.0077	0.0026

表 4 显示在上述四个真实数据集中，基于 p-Huber 损失函数的学习回归算法的表现要优于其它的方法。

4. 结语

在本文中，我们提出了一种新的鲁棒损失函数 p-Huber 损失函数。通过选取不同参数，对 p-Huber 损失函数进行了图像和性质分析。为了说明 p-Huber 损失函数的有效性，我们分别在辛格函数，弗里德曼基准函数和真实数据集下进行了实验。 $p<1$ 时 p-Huber 损失函数为非凸函数，通过跟其他常见损失函数进行比较，结果显示基于 p-Huber 损失函数的回归学习算法在有离群值的情况下其预测效果要优于常见的损失函数的预测效果，其更具鲁棒性。

参考文献

[1] Davies, P.L. (1993) Aspects of Robust Linear Regression. *Annals of Statistics*, **21**, 1843-1899. https://doi.org/10.1214/aos/1176349401

[2] Huber, P.J. (1964) Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics*, **35**, 73-101. https://doi.org/10.1214/aoms/1177703732

[3] Girshick, R. (2015) Fast R-CNN. *2015 IEEE International Conference on Computer Vision* (*ICCV*), Santiago, 7-13 December 2015, 1440-1448. https://doi.org/10.1109/ICCV.2015.169

[4] Feng, Y., Huang, X., Shi, L., Yang, Y. and Suykens, J.A.K. (2015) Learning with the Maximum Correntropy Criterion Induced Losses for Regression. *Journal of Machine Learning Research*, **16**, 993-1034.

[5] Santamaria, I., Pokharel, P.P. and Principe, J.C. (2006) Generalized Correlation Function: Definition, Properties, and Application to Blind Equalization. *IEEE Transactions on Signal Processing*, **54**, 2187-2197. https://doi.org/10.1109/TSP.2006.872524

[6] Aronszajn, N. (1950) Theory of Reproducing Kernels. *Transaction of the AmericanMathematical Society*, **68**, 337-404. https://doi.org/10.1090/S0002-9947-1950-0051437-7

[7] Zhang, H. and Zhang, J. (2012) Regularized Learning in Banach Spaces as an Optimization Problem: Representer Theorems. *Journal of Global Optimization*, **48**, 1-16.

[8] Gearhart, W.B. and Schulz, H.S. (1990) The Function Sinx/x. *The College Mathematics Journal*, **21**, 90-99. https://doi.org/10.1080/07468342.1990.11973290

[9] Stenger, F. (1981) Numerical Methods Based on the Whittaker Cardinal or Sinc Functions. *SIAM Review*, **23**, 165-224. https://doi.org/10.1137/1023037

[10] Friedman, J.H. (1991) Multivariate Adaptive Regression Splines. *The Annals of Statistics*, **19**, 1-67. https://doi.org/10.1214/aos/1176347963