An application of a robust regression method based on Gaussian kernel function

Adenice Ferreira*
Dept. of Statistics - UFPB, João Pessoa, Brazil - adne.adenice@gmail.com

Eufrásio Lima Neto
Dept. of Statistics - UFPB, João Pessoa, Brazil - eufrasio@de.ufpb.br

Marcelo Ferreira
Dept. of Statistics - UFPB, João Pessoa, Brazil - marcelo@de.ufpb.br

Rodrigo Silva
Dept. of Statistics - UFPB, João Pessoa, Brazil - rodrigo@de.ufpb.br

Francisco Carvalho
Computer Center (CIn) - UFPE, Recife, Brazil - fatc@cin.ufpe.br

Abstract

The use of robust regression methods occurs in practical situations due to the presence of outliers. This paper proposes a robust regression method that re-weighted the outliers observations considering the Gaussian kernel function (KRR). The parameter estimate algorithm presents a low computational cost and the convergence is guaranteed. An application with a real data set have showed the usefulness of the KRR method in comparison with some classical robust approaches (WLS, M-Estimator, MM-Estimator, L1 regression) and the OLS method.

Keywords: Kernel, Robust Methods, Regression Models, Outlier.

1. Introduction

Robust regression attempts to cope with outliers and with leverage points. The regression outliers are observations deviating from the linear model pattern of the majority of the data, whereas leverage points are outlying in the space of the predictor variables. Observations which are regression outliers and leverage points are called bad leverage points. A non-robust analysis of data containing such points typically leads to erroneous results.

In this paper we propose a robust regression method that re-weighted the outliers observations considering the Gaussian kernel function. The convergence of the parameter estimate algorithm is guaranteed with a low computational cost. A comparative study between the proposed kernel based robust regression method (KRR) against the following classical regression approaches is considered: Weighted least squares (WLS), M-Estimator (Huber, 1973), MM-Estimator, L1 regression and the Ordinary least squares (OLS) method.

The paper is organized as follows: Section 2 presents the kernel based regression robust method and discuss the parameter estimate algorithm. Section 3 presents an application to a real data set an illustrates the usefulness of the KRR approach in comparison with other classical robust methods. Finally, Section 4 gives the concluding remarks.

2. KRR Method

A multiple linear regression model can be presented in the matrix notation as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} = \boldsymbol{\mu} + \boldsymbol{\epsilon} \tag{1}$$

where $\mathbf{y}$ is an $n \times 1$ vector of the response variable, $\mathbf{X}$ is an $n \times p$ matrix of explanatory variables, $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression coefficients parameters, $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$ is an $n \times 1$ vector representing the means of $\mathbf{y}$ and $\boldsymbol{\epsilon}$ is a $n \times 1$ vector containing the error terms with zero mean and an unknown variance σ^2.

The ordinary least squares (OLS) estimator optimizes parameters $\boldsymbol{\beta}$ by making the sum of squared residuals as small as possible. The OLS estimator is the optimal one for data with normal error distribution. If this assumption is violated, this estimator can perform very poorly. One of the possible reasons of unsatisfactory OLS models is the presence of regression outliers. One or a few observations, which do not follow the same model as the rest of the data, can strongly influence the regression coefficients.

Hereafter we briefly recall the basic theory about kernel functions. $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ be a non-empty set where $\mathbf{x}_i \in \mathbb{R}^p$. A function $\mathcal{K} : X \times X \to \mathbb{R}$ is called a *positive definite kernel* (or *Mercer kernel*) if and only if $\mathcal{K}$ is symmetric (i.e. $\mathcal{K}(\mathbf{x}_i, \mathbf{x}_k) = \mathcal{K}(\mathbf{x}_k, \mathbf{x}_i)$) and the following inequality hold:

$$\sum_{i=1}^{n}\sum_{k=1}^{n} c_i c_k \mathcal{K}(\mathbf{x}_i, \mathbf{x}_k) \geq 0 \quad \forall n \geq 2,$$

where $c_r \in \mathbb{R}\ \forall r = 1, \ldots, n$ (Mercer, 1909).

Let $\Phi : X \to \mathcal{F}$ be a non-linear mapping from the input space X to a high dimensional feature space $\mathcal{F}$. By applying the mapping Φ, the dot product $\mathbf{x}_i^\top \mathbf{x}_k$ in the input space is mapped to $\Phi(\mathbf{x}_i)^\top \Phi(\mathbf{x}_k)$ in the feature space. The key idea in kernel algorithms is that the non-linear mapping Φ do not need to be explicitly specified because each Mercer kernel can be expressed as $\mathcal{K}(\mathbf{x}_i, \mathbf{x}_k) = \Phi(\mathbf{x}_i)^\top \Phi(\mathbf{x}_k)$ (Müller et al., 2001).

One of the most relevant aspects in applications is that it is possible to compute Euclidean distances in $\mathcal{F}$ without knowing explicitly Φ. This can be done using the so called *distance kernel trick* (Müller et al., 2001):

$$\begin{aligned} ||\Phi(\mathbf{x}_i) - \Phi(\mathbf{x}_k)||^2 &= (\Phi(\mathbf{x}_i) - \Phi(\mathbf{x}_k))^\top (\Phi(\mathbf{x}_i) - \Phi(\mathbf{x}_k)) \\ &= \Phi(\mathbf{x}_i)^\top \Phi(\mathbf{x}_i) - 2\Phi(\mathbf{x}_i)^\top \Phi(\mathbf{x}_k) + \Phi(\mathbf{x}_k)^\top \Phi(\mathbf{x}_k) \\ &= \mathcal{K}(\mathbf{x}_i, \mathbf{x}_i) - 2\mathcal{K}(\mathbf{x}_i, \mathbf{x}_k) + \mathcal{K}(\mathbf{x}_k, \mathbf{x}_k). \end{aligned}$$

An example of a commonly used kernel function is the Gaussian kernel, given by

$$\mathcal{K}(\mathbf{x}_i, \mathbf{x}_k) = \exp\left\{-\frac{||\mathbf{x}_i - \mathbf{x}_k||^2}{2\gamma^2}\right\}, \quad \gamma > 0.$$

The basic idea in the kernel based robust regression method is to minimize the following objective function:

$$S = \sum_{i=1}^{n} ||\Phi(y_i) - \Phi(\hat{\mu}_i)||^2 = \sum_{i=1}^{n} \left\{\mathcal{K}(y_i, y_i) - 2\mathcal{K}(y_i, \hat{\mu}_i) + \mathcal{K}(\hat{\mu}_i, \hat{\mu}_i)\right\}. \tag{2}$$

If we considered the Gaussian kernel we will have that $\mathcal{K}(y_i, y_i) = \mathcal{K}(\hat{\mu}_i, \hat{\mu}_i) = 1$ $(i = 1, \ldots, n)$ and the functional (2) can be re-written as:

$$S = \sum_{i=1}^{n} 2\left[1 - \mathcal{K}(y_i, \hat{\mu}_i)\right] = \sum_{i=1}^{n} 2\left\{1 - \exp\left[-\frac{(y_i - \hat{\mu}_i)^2}{2\gamma^2}\right]\right\}, \quad \gamma > 0. \tag{3}$$

Minimization of the functional given in Eq. (3) with respect to the parameter vector $\boldsymbol{\beta}$ is equivalent to iteratively re-estimate the parameters of the Eq. (4) through an iterative re-weighting least squares process where the weights are computed by means of Gaussian kernels:

$$\mathbf{y}^* = \mathbf{X}^* \boldsymbol{\beta}^K + \boldsymbol{\epsilon}^* = \boldsymbol{\mu}^* + \boldsymbol{\epsilon}^* \tag{4}$$

where $\mathbf{y}^* = \mathbf{K}^{1/2}\mathbf{y}$, $\mathbf{X}^* = \mathbf{K}^{1/2}\mathbf{X}$ and $\mathbf{K} = \text{diag}(k_{11}, k_{22}, \ldots, k_{nn})$ is an $n \times n$ diagonal weight matrix with

$$k_{ii} = k_{ii}(y_i, \hat{\mu}_i) = \exp\left\{-\frac{(y_i - \hat{\mu}_i)^2}{2\gamma^2}\right\}, \gamma > 0, \forall i = 1, 2, \ldots, n. \quad (5)$$

The estimation of the modified parameter vector $\boldsymbol{\beta}^K$ can be performed through the following algorithm:

Algorithm 1: $\boldsymbol{\beta}^K$ Parameter Estimation Process

INPUT: $\mathbf{X}$, $\mathbf{y}$, a tolerance limit ε, a maximum number of iterations T, the Gaussian kernel hyper-parameter γ^2
OUTPUT: $\hat{\boldsymbol{\beta}}^K$, $\hat{\boldsymbol{\mu}}$
INITIALIZATION:

Set $t = 0$ and $\mathbf{K}^{(0)} = \mathbf{I}_n$, where $\mathbf{I}_n$ is the $n \times n$ identity matrix

Compute $\hat{\boldsymbol{\beta}}^{K(0)} = (\mathbf{X}^\top \mathbf{K}^{(0)} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{K}^{(0)} \mathbf{y}$

Compute $\hat{\boldsymbol{\mu}}^{(0)} = \mathbf{X}\hat{\boldsymbol{\beta}}^{K(0)}$

Compute $S^{(0)}$ as given in Eq. (3)

FITTING STEP:
repeat
  Set $t = t + 1$
  Compute $\mathbf{K}^{(t)} = \text{diag}\{k_{11}, \ldots, k_{nn}\}$ where $k_{ii}^{(t)} = \mathcal{K}_{\mathcal{E}}(y_i, \hat{\mu}_i^{(t-1)}) = \mathbf{x}_i^T \hat{\boldsymbol{\beta}}^{(t-1)}$;
  Compute $\hat{\boldsymbol{\beta}}^{K(t)} = (\mathbf{X}^\top \mathbf{K}^{(t)} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{K}^{(t)} \mathbf{y}$
  Compute $\hat{\boldsymbol{\mu}}^{(t)} = \mathbf{X}\hat{\boldsymbol{\beta}}^{K(t)}$
  Compute $S^{(t)}$ as given in Eq. (3)
until $|S^{(t)} - S^{(t-1)}| \leq \varepsilon$ *or* $t \geq T$;

The Gaussian kernel hyper-parameter γ^2 is estimated as the average of the 0.1 and 0.9 percentiles of $\sum_{i=1}^{n}(y_i - \hat{\mu}_i^{\text{OLS}})^2$ (Caputo et al., 2002), where $\hat{\mu}_i^{\text{OLS}}$ is the predicted value of y_i, $i = 1, \ldots, n$, obtained using the ordinary least squares method.

3. Application to Real Data Set

This data set, named "Public Schools", is available in `R` library sandwich. The response variable represents the per capita spending on public schools (Expenditure). The unique covariate is the per capita income, by state, USA (Greene, 1993). We have removed the state of Wisconsin because information about it were missing.

Figure 1 illustrates the scatter plot including the straight lines estimated for each method assessed in this paper. Table 1 exhibit the estimated coefficients as well as the computational time spent (in seconds). We considered the Gaussian kernel for KRR method in this data set. From Figure 1 it can be noted that the state of Alaska represents a high leverage point. The methods KRR and MM-Estimator (red and yellow lines, respectively) have presented very close regression lines, while the OLS method (black line) exhibited the worst regression equation. It can be also observed that the methods MM and KRR are less influenced by the high leverage point (the Alaska observation) than the OLS, WLS, M-Estimator and L1 methods.

It can be verified that the straight lines for MM and KRR methods are almost overlapped. This can be checked inspecting the Table 1. According to Table 2, it is possible conclude that the methods MM-Estimator and KRR have presented the lower percentage of change in the parameter estimates, after suppression of the outlier observation. The KRR method presented a lowest computational time when compared with MM-Estimator method.

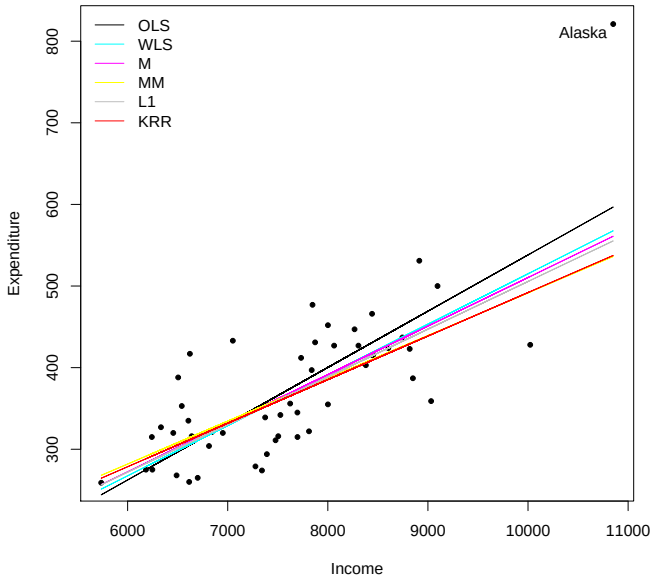

Figure 1: Public school data: Estimated straight line according to method.

Table 1: Public school data: Regression coefficient estimates and computational time spent.

Method	$\hat{\beta}_0$	$\hat{\beta}_1$	Time
OLS	−151.2651	0.0689	0.0020
WLS	−103.8318	0.0619	0.0020
M	−84.5132	0.0595	0.0040
MM	−32.7014	0.0524	0.0120
L1	−78.6607	0.0584	0.0030
KRR	−41.5718	0.0534	0.0070

Table 2: Public school data: Percentage change (%) in the parameter estimates after suppression of the outlier observation.

Method	$\hat{\beta}_0$	$\hat{\beta}_1$
OLS	82.28	24.82
WLS	30.63	7.12
M	66.04	12.80
MM	0.99	0.06
L1	20.12	3.96
KRR	19.30	1.50

4. Concluding Remarks

This paper have proposed a robust regression method considering a Gaussian kernel function. The convergence of the parameter estimate algorithm was guaranteed with a low computational cost. A comparative study between the proposed kernel based robust regression method (KRR) against some classical regression approaches in a real data set have demonstrated the usefulness of the proposed method.

Acknowledgments: The authors would like to thank CNPq and CAPES (Brazilian Agencies) for their financial support.

References

Caputo, B., Sim, K., Furesjo, F. & Smola, A. (2002). Appearance-based object recognition using SVMs: which kernel should I use? *In: Proceedings of NIPS workshop on Statitsical methods for computational experiments in visual processing and computer vision*, Whistler.

Greene, W.H. (1993). Econometric Analysis, 2nd edition. Macmillan Publishing Company, New York.

Huber, P.J. (1973). Robust regression: asymptotics, conjectures and Monte Carlo. *The Annals of Statistics*, 1 (5), 799–991.

Mercer, J. (1909). Functions of positive and negative type and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society A*, 209 (441-458): 415-446.

Müller, K.R, Mika, S., Rätsch, G.R., Tsuda, K. & Schölkopf, B. (2001). An introduction to kernel-based learning algorithms. *IEEE Trans. Neural Networks*, 12, 181–202.