Robust Regression by Self-Updating Prcocess

Ting-Li Chen [*]

Abstract

Robust regression aims to reduce the effect from outliers. One standard approach is to perform weighted regression in which the weights are iteratively updated according to the new fitted line. In this paper, we will present an iterative process to reduce the effect from outliers. It is an extension of SUP clustering algorithm (Chen and Shiu, 2007). This process updates both the weights and the data points through iterations. At each iteration, a line is fitted locally for each data point. The data point is then moved to this line. Throughout this process, all data points except outliers will gradually move to form a line. We will show results from simulation studies that our proposed method outperforms the standard approach. The success of the proposed algorithm comes from its two important properties: One is that the local estimation can reduce the effect from outliers so that the method is more robust. The other is that moving data based on the current estimation can improve the overall efficiency.

Key Words: Robust regression, iterative process, clustering.

1. Introduction

Chen and Shiu (2007) proposed a clustering algorithm which stands from the viewpoint of elements to be clustered and simulates the process of how they perform self-clustering. At the end of the process, elements converge to the same position are treated as belonging to the same cluster. The exact algorithm is as follows:

(I) $x_1^{(0)}, \dots, x_N^{(0)} \in R^p$ are data points to be clustered.

(II) At time $t+1$, every point is updated to

$$x_i^{(t+1)} = \sum_{j=1}^{N} \frac{f_t(x_i^{(t)}, x_j^{(t)})}{\sum_{k=1}^{N} f_t(x_i^{(t)}, x_k^{(t)})} x_j^{(t)}, \tag{1}$$

where f_t is some function that measures the influence between two data points at time t.

(III) Repeat (II) until every data point no longer moves.

f_t is chosen to have a compact support, so that every data point is affected by it neighbors. The strengths of this algorithm are processing the following types of data (Shiu and Chen, 2014): (i) data with noise, (ii) data with large number of clusters, and (iii) unbalanced data. With these advantages, a robust algorithm, γ-SUP, based on minimizing γ-divergence has a great success on the CRYO-EM images (Chen et al., 2014).

In this paper, we generalize the Self-Updating Process (SUP) from a clustering algorithm to a robust regression algorithm.

[*] Institute of Statistical Science, Academia Sinica, 128 Academia Road Sec.2, Nankang District, Taipei, 11529 Taiwan

2. Robust Regression

In SUP clustering algorithm, each data point can be viewed to have two roles: one is the data point itself, and the other is the cluster center. The process starts with each data point being a single-point cluster. Then some clusters merge through the process. Remind that the updating rule is:

(II) At time $t+1$, every point is updated to

$$x_i^{(t+1)} = \sum_{j=1}^{N} \frac{f_t(x_i^{(t)}, x_j^{(t)})}{\sum_{k=1}^{N} f_t(x_i^{(t)}, x_k^{(t)})} x_j^{(t)}.$$

.

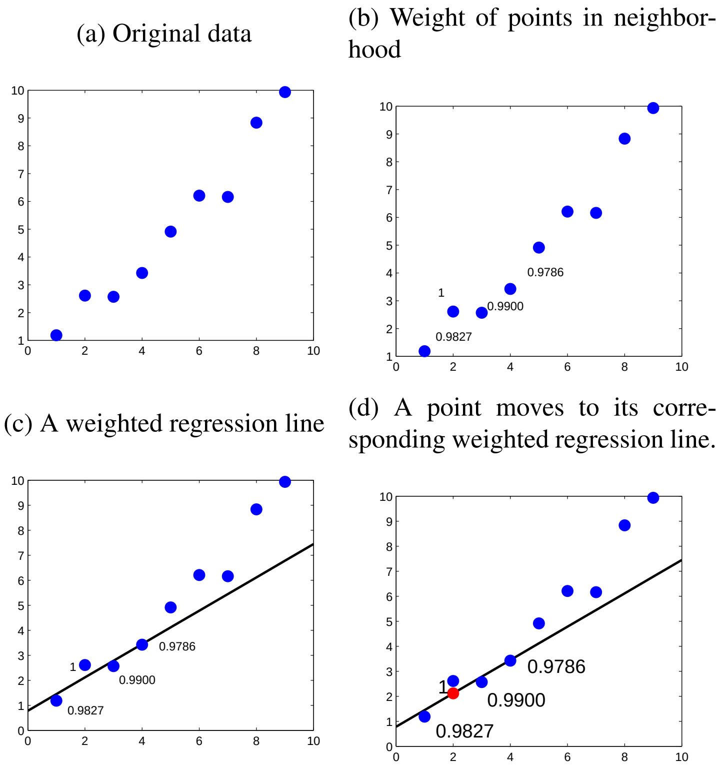
It can be rewritten as two steps

(IIa) At time $t+1$, a cluster center is calculated based on the data points in its neighborhood:

$$x_i^{(t+1)} = \sum_{j=1}^{N} \frac{f_t(x_i^{(t)}, x_j^{(t)})}{\sum_{k=1}^{N} f_t(x_i^{(t)}, x_k^{(t)})} x_j^{(t)}.$$

.

(IIb) The i-th point is moved to its corresponding cluster center.



(a) Original data

(b) Weight of points in neighborhood

(c) A weighted regression line

(d) A point moves to its corresponding weighted regression line.

Figure 1: Illustration of how the weighted regression line is fitted and how the data point is updated.

Now we can easily replace "cluster center" with a specific model. Here we use a line as an example to perform a robust linear regression. The algorithm can be formatted as follow:

(I) $x_1^{(0)}, \ldots, x_N^{(0)} \in R^p$ are data points to be regressed.

(II) At time $t+1$, for each point, a local weighted regression line is fitted based on the data points in its neighborhood.

(III) Each point is moved to its corresponding weighted regression line.

(IV) Repeat (II) and (III) until every data point no longer moves.

How the process executes in the step (II) and (III) is illustrated in Figure 1.

Assume that the weight function is a decreasing function with respect to the distance between two data points. If the weight is 0 for distance larger than a specific amount, the data points may converge into several straight lines in the end. This is similar to the clustering case that there may be several clusters if f_t has a compact support. While it is not preferred for the clustering case to have a single cluster in the end (Chen, 2014), it is meaningful to have a single overall regression line. The phenomenon is stated in the following conjecture. It should be true, but we have not been able to prove it.

Conjecture 1. *If the weight function is a positive and decreasing function with respect to the distance between any two data points, all data points will converge to a straight line by SUP.*

3. Simulation

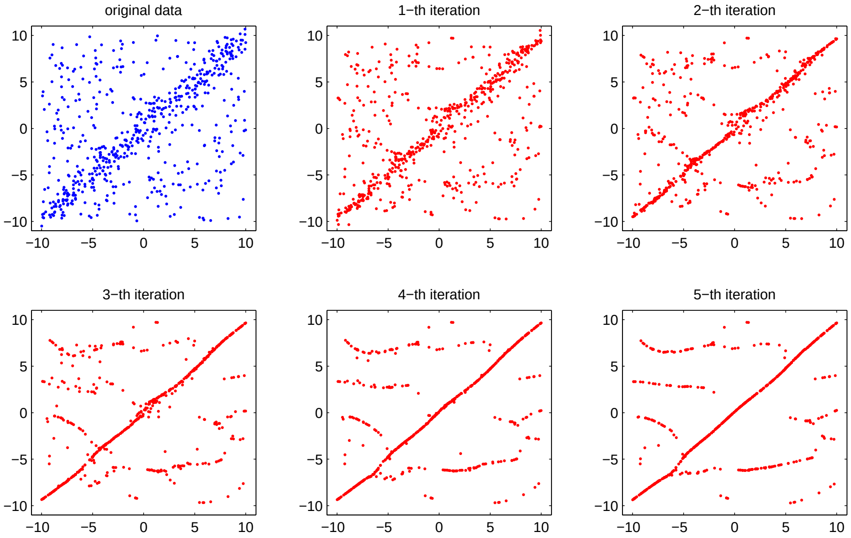


Figure 2: How data points move through iterations by SUP

Example 1:

300 points are sampled from the model $Y = X + \epsilon$, where $X \sim U[-10, 10]$ and $\epsilon \sim N(0, 1)$. Another 300 points which are treated as outliers are sampled from $U[-10, 10] \times [10, 10]$. One set of sample points is shown in Figure 2.

Figure 2 also illustrates how data points move in first five iterations by SUP. In this simulation, each data points is updated by the weighted regression line based on the points within distance 2. The weight function is $\exp(-d/T)$ where d is the Euclidean distance and $T = 100$. In this setup, the weight is close to uniform for all the points within distance 2.

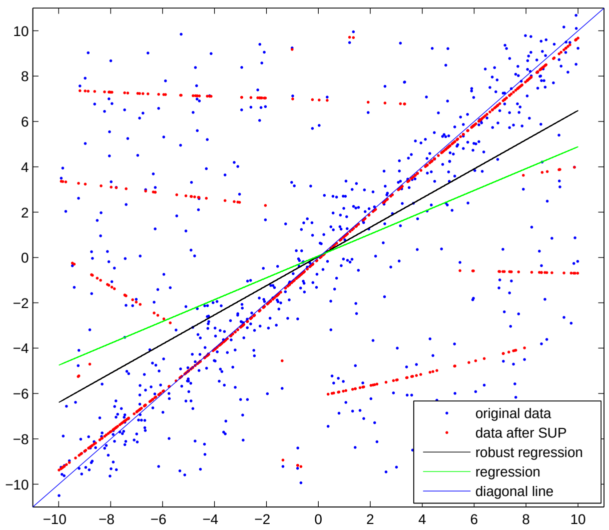


Figure 3: Comparison of regression, robust regression, and SUP

We can see that data points are updated so that they look more like lines locally. The final converged result is presented in Figure 3.

Table 1: Statistics of estimation of slope from 10000 experiments

	Regression	Robust Regression	SUP 1	SUP 2
mean	0.4997	0.8172	0.9634	0.9515
std	0.0345	0.1173	0.0465	0.0869

We compare our results with ordinary linear regression by least square and the robust regression by Huber (1981). The true slope is 1 from our simulation setting. We simulate 10,000 times and present the mean and the standard deviation by each method in Table 1. SUP2 represents the slope of the final converged line, and SUP1 represents the slope of the regression line on the original data of which converge to the final converged line. We can see that both SUP1 and SUP2 outperforms the standard linear regression and the robust regression by Huber. SUP1 is a little better than SUP2, which coincides with our past experience that the sample mean of those converged points is more accurate that the converged location when using SUP to estimate the center.

Example 2:

In this example, we consider the mixture model of five lines. 100 data points are sampled from each line added with i.i.d. noise $N(0, 1)$ in y-direction. Figure 4 (a) presents where the lines are and Figure 4 (b) displays the data sampled. With the same weight function and parameter values, the data is iteratively updated by SUP. The process is stopped at 18-th iteration, and the result is presented in Figure 4 (c). We also put everything together for comparison in Figure 4 (d).

From Figure 4 (d), the converged lines are very close to the true underlying lines. This example shows that we can apply SUP on the mixture model. Without knowing the class labels of data points, SUP can do the regression on each class separately and simultaneously.

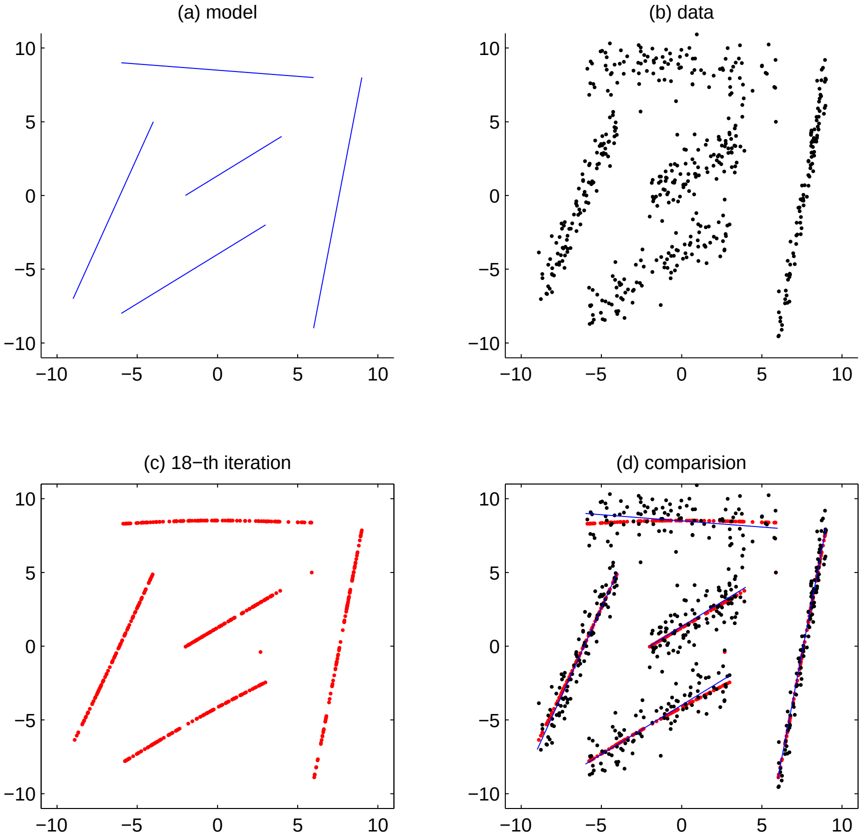


Figure 4: Data in (b) is sampled from the model in (a) with additive Gaussian noise. (c) is the result by SUP. (d) compares the model, the data, and the result

4. Discussion and Conclusion

In this paper, we extended the SUP clustering algorithm to a robust regression algorithm. From our simulation studies, the proposed method outperforms the traditional robust regression.

In fact, the extension can be applied to more general models, such as quadratic curve, hyperplane. The basic principles are local estimation and moving the data points to the estimated model. Local estimation can reduce the effect from outliers, and the iterative estimation based on the updated data can improve the overall efficiency.

Because of local estimation, SUP can handle the model fitting separately and simultaneously when there are multi models. Therefore, it can be used as a multi-feature extractor.

References

Chen, T.-L. (2014). On the convergence and consistency of the blurring mean-shift process. *Annals of the Institute of Statistical Mathematics (to appear)*.

Chen, T.-L., Hsieh, D.-N., Hung, H., Tu, I.-P., Wu, P.-S., Wu, Y.-M., Chang, W.-H., and Huang, S.-Y. (2014). γ-SUP: a clustering algorithm for cryo-electron microscopy images of asymmetric particles. *Annals of Applied Statistics*, 8(1):259–285.

Chen, T.-L. and Shiu, S.-Y. (2007). A clustering algorithm by self-updating process. *JSM Proceedings*, Statistical Computing Section, Salt Lake City, Utah; American Statistical Association, pp:2034–2038.

Huber, P. (1981). *Robust Statistics*. Wiley, New York.

Shiu, S.-Y. and Chen, T.-L. (2014). Clustering by self-updating process. eprint arXiv:1201.1979.