Robust Regression by Self-Updating Process

吳芷瑄 Chih-Hsuan Wu

r04246006@ntu.edu.tw

國立台灣大學應用數學科學研究所

Institute of Applied Mathematical Sciences, National Taiwan University

2017.06.28

Outline

Robust Regression

Let $\{x_i \in \mathbb{R}^p\}_{i=1}^n$ be the explanatory variables and $\{y_i \in \mathbb{R}\}_{i=1}^n$ be the responses. Ordinary least squares model and the estimate of coefficients can be written down as

$$y_i = x_i^T \beta + \epsilon_i, \quad i = 1, ..., n$$

$$\hat{\beta} = \arg\min \sum_{i=1}^{n} (y_i - x_i^T \beta)^2.$$

- Minimize residuals in L^1-norm. (Boscovich 1757)
- Minimize $\sum \rho(r_i)$ for some symmetric function ρ with a unique minimum at zero. To keep scale-invariant, the estimation problem becomes to solve $\sum \phi(r_i/\hat{\tau})x_i = 0$, where $\phi(u)$ is the derivative of ρ and $\hat{\tau}$ is an estimated scale. (Huber 1973)
- Minimize the median of the squared residuals. (Rousseeuw 1984)

Clustering by Self-Updating Process

- A distance-based clustering method: Self-Updating Process (Chen et al. 2007)
  (i) $x_1^{(0)}, ..., x_N^{(0)} \in \mathbb{R}^p$ are the original positions of data points to be clustered.
  (ii) At time $t+1$, every point is updated to the following new position:

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

  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.
- SUP shows advantages in clustering (i) data with noise, (ii) data with a large number of clusters, and (iii) unbalanced data.

Mean-Shift Clustering

- Mean-shift clustering derived from kernel density estimate is another iterative process.

$$y_{j+1} = \sum_{i=1}^{n} \frac{g\left(\|\frac{y_j - x_i}{h}\|^2\right) x_i}{\sum_{i=1}^{n} g\left(\|\frac{y_j - x_i}{h}\|^2\right)}$$

  where y_j is the mode estimate in the j-th iteration, and $k(x) = -g(x)$ is the kernel profile. This formulation is called non-blurring mean-shift. If we substitute y_i for x_i , then it becomes blurring type, which can be viewed as a static SUP.

- Compare mean-shift with SUP

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

Convergence of Mean-Shift

- Comaniciu, D., and Meer, P. (2002). Misuse an inequality.
- Li, X., Hu, Z., and Wu, F. (2007). Assume finite modes of estimated pdf.
- Ghassabeh, Y. A. (2015). Focus on Gaussian kernel.
- Arias-Castro, E., Mason, D., and Pelletier, B. (2016). Convergence rate

Concept of Our Method

- Iterative process
- Move data points
- Distance-based
  - Blurring: Self-updates w.r.t. updating points and depends on $d(z_i^{(t)}, z_j^{(t)})$
  - Non-blurring: Self-updates w.r.t. original points and depends on $d(z_i^{(t)}, z_j)$

Illustration of the Algorithm

Given a data set $\{z_i = (x_i^T, y_i)\}_{i=1}^n$, each $x_i \in \mathbb{R}^p$ consists of the measurements of p independent variables and $y_i \in \mathbb{R}$ is the measurement of the corresponding dependent variable.

Step 1. For each z_i , fit the locally weighted regression with weight $w_i(k) = w(z_k, z_i)$ for z_k to estimate the coefficient $\hat{\beta}_i^{(0)}$ which minimize

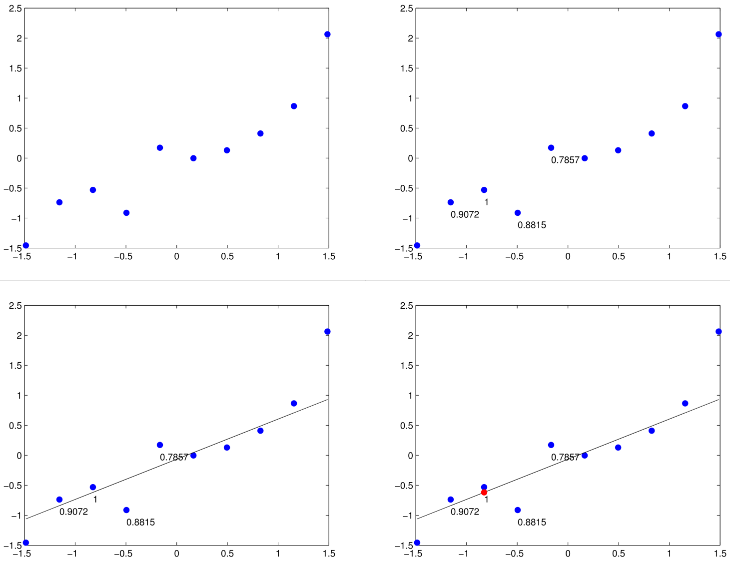
$$\sum_{k=1}^{n} w_i(k)(y_k - x_k^T \beta_i)^2.$$

Step 2. Define $z_i^{(1)} = z_i^{[1]} = X\hat{\beta}_i^{(0)}$, the locally fitted value of z_i by step 1.

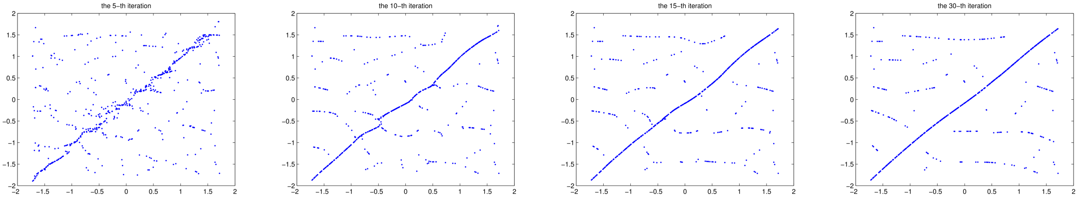
Step 3. At the t-th iteration, $t = 2, 3, \ldots$, compute the estimated coefficients for different types of SUP by fitting locally weighted least squares for each z_i with weight defined as follows:
non-blurring: $w_i^{[t]}(k) = w(z_k, z_i^{[t-1]})$ and
blurring: $w_i^{(t)}(k) = w(z_k^{(t-1)}, z_i^{(t-1)})$
And their t-th estimated coefficients $\hat{\beta}_i^{[t]}$ and $\hat{\beta}_i^{(t)}$ are defined as follows:
non-blurring: $\hat{\beta}_i^{[t]} = \arg\min \sum_{k=1}^{n} w_i^{[t]}(k)(y_k - x_k^T \beta_i)^2$, and
blurring: $\hat{\beta}_i^{(t)} = \arg\min \sum_{k=1}^{n} w_i^{(t)}(k)(y_k^{(t)} - x_k^T \beta_i)^2$.

Step 4. Update $y_i^{[t]}$ and $y_i^{(t)}$ to their locally fitted value $y_i^{[t+1]} = (X\hat{\beta}_i^{[t]})_i$, and $y_i^{(t+1)} = (X\hat{\beta}_i^{(t)})_i$ respectively.

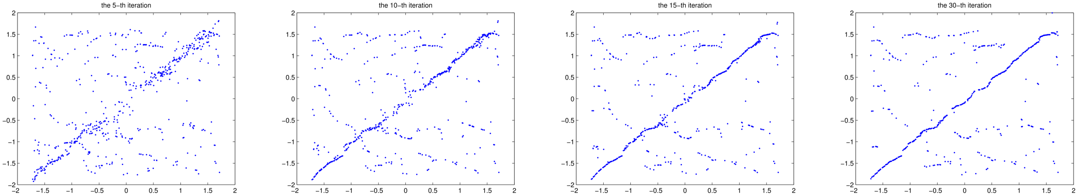
Step 5. Repeatedly carry out step 3 and 4 for limited times m or until the maximum difference of the updating points is lower than a threshold value.


0.7857
1
0.9072
0.8815

Blurring



Non-blurring



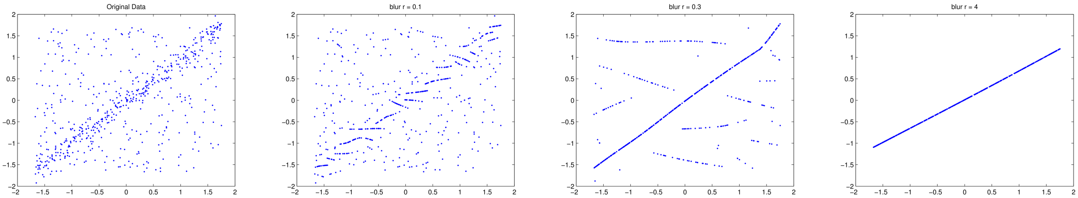
Estimation

- *Protocol 1*
  1. Fit a robust regression on the final data, and find the weight w_i of each point $(x_i^T, y_i^{(m)})$ (or $(x_i^T, y_i^{[m]})$). Collect points of top p percent weight among all points, denoting their index as $I_p = \{i : w_i \geq$ the (100-p)-th percentile of $\{w_i\}_{i=1}^n\}$
  2. Fit a linear regression on $\{(x_i^T, y_i^{(m)})\}_{i \in I_p}$ or $\{(x_i^T, y_i^{[m]})\}_{i \in I_p}$.
- *Protocol 2*
  1. This step is the same as step 1 in protocol 1.
  2. Fit a linear regression on $\{(x_i^T, y_i)\}_{i \in I_p}$.

The Parameter r

$$w(u,v) = \begin{cases} \exp\left[\frac{-d(u,v)}{T}\right] & \text{if } d(u,v) \le r. \\ 0 & \text{if } d(u,v) > r. \end{cases}$$

- r is the influence range of the weight function.
- Local structure vs global structure

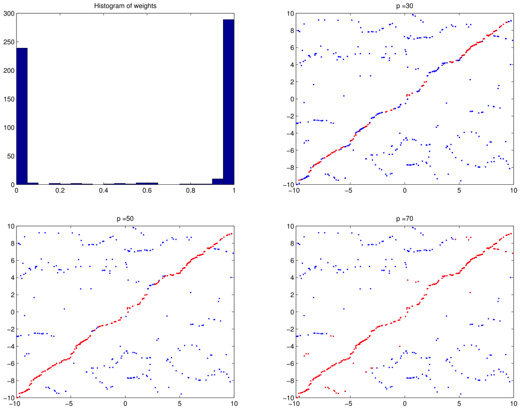


The Parameter T

- Large $T \Rightarrow$ All weights are nearly the same.
  $\Rightarrow$ Locally weighted least squares are nearly ordinary least squares
- Small $T \Rightarrow$ Only a few points are concerned.
- Choose T such that $w(u, v)$ is almost zero when $d(u, v) = r$.
  e.g. $T = r/5$ for $\exp(\frac{-d}{T})$

The Parameter p

- Small value of p results in loss of more points.
- Large p makes the estimation contain some outliers.



Data with Uniform Noise

- SUP in clustering shows a strong power in reducing the effect of noise.
- The data set consists of 300 responses $y_i = x_i + \epsilon_i$, where $\epsilon_i, i = 1, .., 300$ are i.i.d. random variables following standard normal distribution, and 300 points sampled from uniform distribution on $[-10, 10]$ x $[-10, 10]$ considered as extra uniform noise.

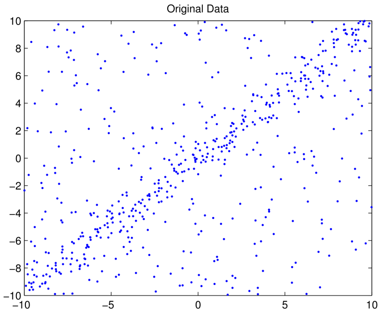


Table: Comparison of different methods

$r = 0.3$	**OLS**	**robustfit**	**robust_cut**	**nonblur1**	**nonblur2**	**blur1**	**blur2**
mean	0.4984	0.8047	0.8995	0.9929	0.9513	0.9549	0.9627
std	0.0356	0.1253	0.0664	0.027	0.0293	0.092	0.0691
MSE of slope	0.2529	5.38E-02	1.45E-02	7.78E-04	3.20E-03	1.05E-02	6.20E-03
coverage probability	0		0	0.215	0.225	0.08	0.645
mean length	0.1401		0.0411		0.0637		0.057
time	0.00027	0.00762		2.33511		2.298025	

Data with Heavy-Tailed Noise

The data are sampled from $y_i = x_i + \epsilon_i$, where $\epsilon_i, i = 1, .., 300$ are i.i.d. random variables following student-t distribution with 3 degrees of freedom.

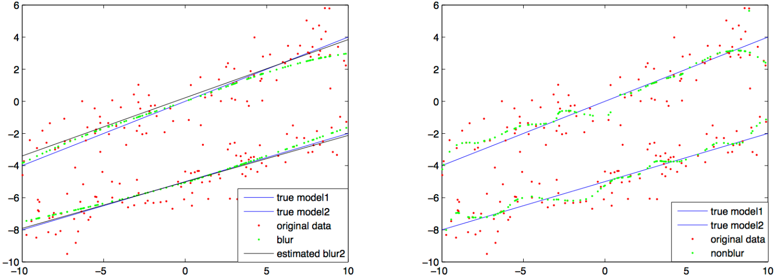
Table: Heavy-tailed

	OLS	**robust**	**robust_cut**	**nonblur1**	**nonblur2**	**blur1**	**blur2**
mean	0.9995	0.9994	0.9995	0.9995	0.9995	0.9997	0.9993
std	0.0157	0.0114	0.0126	0.0126	0.0126	0.0127	0.0123
MSE of slope	2.47E-04	9.50E-03	1.58E-04	1.60E-04	1.60E-04	1.62E-04	1.52E-04
coverage probability	0.97		0.97	0.975	0.97	0.965	0.97
mean length	0.0652		0.0537	0.0541	0.0538	0.0519	0.0527

Multiple Linear Models

There are 2 linear models considered simultaneously: $y_i = -5 + 0.3x_i + \epsilon_i$ and $y_{i+100} = 0.4x_{i+100} + \delta_i$, where $\epsilon_i, \delta_i, i = 1, .., 100$ are i.i.d. random variables following standard normal distribution.

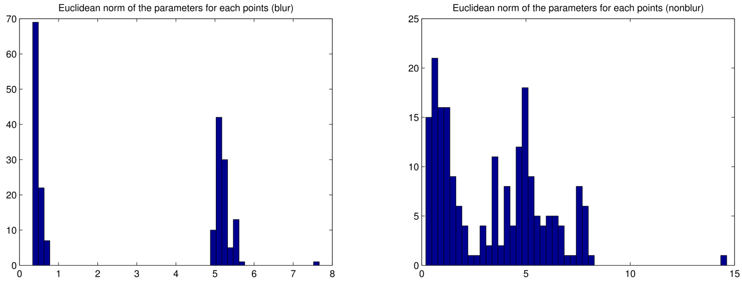
Figure: Multiple lines



Adjust the estimation protocols

- Calculate the Euclidean norm of the estimated parameters for each points.
- Furthermore, assign the points with similar parameters into the same group and continue the estimation protocols.

Figure: Histogram of Euclidean norm of parameters



Definition

The function f is positive and decreasing with respect to distance(PDD), if

(i) $0 \leq f(u,v) \leq 1$, *and* $f(u,v) = 1$ *if and only if* $u = v$.

(ii) $f(u,v)$ *depends only on* $\|u - v\|$, *the distance from u to v.*

(iii) $f(u,v)$ *is decreasing with respect to* $\|u - v\|$.

For example,

$$f_t(x_i^{(t)}, x_j^{(t)}) = \begin{cases} \exp\left[\frac{-d(x_i^{(t)}, x_j^{(t)})}{T(t)}\right] & \text{if } d(x_i^{(t)}, x_j^{(t)}) \leq r. \\ 0 & \text{if } d(x_i^{(t)}, x_j^{(t)}) > r. \end{cases}$$

Convergence of SUP Clustering

Theorem (Chen 2015)

Consider a process,

$$x_i^{(t+1)} = \sum_{j=1}^{N} \frac{f\left(x_j^{(t)} - x_i^{(t)}\right) w(x_j^{(t)}) x_j^{(t)}}{\sum_{k=1}^{N} f\left(x_k^{(t)} - x_i^{(t)}\right) w(x_k^{(t)})}$$

If f is PDD and $w(x_j^{(t)}) = w_j$ *depends only on j , there exists* $\{x_1^*, \ldots, x_N^*\}$, *such that*

$$\lim_{t\to\infty} x_i^{(t)} = x_i^* \quad \forall i = 1, \ldots, N$$

Blurring

- Since the concept of regression by blurring SUP is similar to clustering by SUP, we guess that there are similar properties in the case of regression.
- The points seem to converge locally to several lines if the weight function has a compact support.

Conjecture

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

Non-Blurring

- For non-blurring SUP, the PDD condition on f is not enough for convergence.

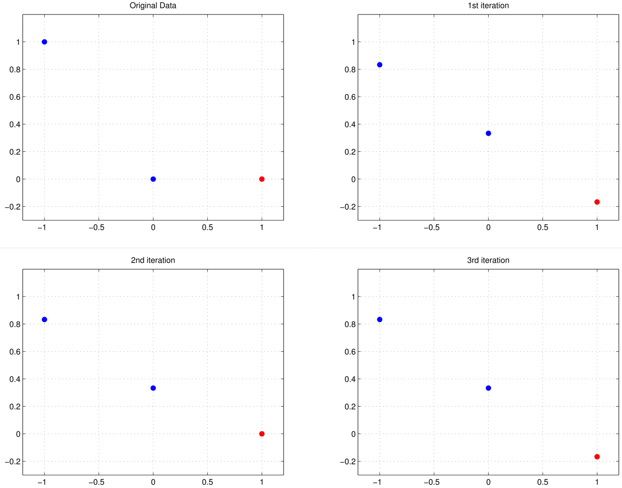
Example 1. Consider a data set $\{z_1 = (-1, 1), z_2 = (0, 0), z_3 = (1, 0)\}$ with weight function

$$w(d) = \begin{cases} 1 & \text{if } d \leq \sqrt{5} \\ 0 & \text{if } d > \sqrt{5} \end{cases}$$

Now we focus on the third point.

- At the first iteration, the distance between z_3 and $z_i, i = 1, 2$ are both less than or equal to $\sqrt{5}$. Therefore, the locally weighted least squares is actually an ordinary least squares on these three points, which causes that $z_3^{[1]} = (1, -t)$ for some $t > 0$.
- At the second iteration, $z_3^{[2]}$ will move to the weighted regression line of z_2 and z_3 because $\|z_1 - z_3^{[1]}\|$ is greater than $\sqrt{5}$. And this becomes the same situation as beginning.
- We conclude that $z_3^{[2n-1]} = (1, -t)$ and $z_3^{[2n]} = (1, 0)$ for all $n \in \mathbb{N}$.

Figure: Illustration of example 1



Example 2. Consider a simple data set with only 3 points: $\{(-1, y_1), (0, y_2), (1, y_3)\}$. Standardize the explanatory variable and the response. Therefore, $\sum_{i=1}^{3} y_i = 0$ and $\sum_{i=1}^{3} y_i^2 = 3$. And choose $w(d) = e^{-d^2/T}$ as the weight function in the non-blurring SUP.

- Let $F_i(y), i = 1, 2, 3$ be the function satisfying $y_i^{[t+1]} = F_i(y_i^{[t]})$ for any possible value $y_i^{[t]}$. In this simple case, each $F_i(y)$ can be written down explicitly.

$$F_1(y) = \frac{2u_{21}u_{31}y_2 + 4u_{11}u_{31}y_1 + u_{11}u_{21}y_1 - u_{21}u_{31}y_3}{u_{11}u_{21} + 4u_{11}u_{31} + u_{21}u_{31}}$$
$$F_2(y) = \frac{u_1u_2y_2 - 2u_1u_3y_2 + u_2u_3y_2}{u_{12}u_{22} + 4u_{12}u_{32} + u_{22}u_{32}}$$
$$F_3(y) = \frac{2u_{13}u_{23}y_2 + 4u_{13}u_{33}y_3 + u_{23}u_{33}y_3 - u_{13}u_{23}y_1}{u_{13}u_{23} + 4u_{13}u_{33} + u_{23}u_{33}}$$

where $u_{ij} = u_{ij}(y) = e^{-\frac{1}{T}[(y_i-y)^2+(x_i-x_j)^2]}, i, j = 1, 2, 3.$

After tedious calculations, we may obtain

$$F_1'(y) = \frac{2u_{11}u_{21}u_{31}[4u_{31}(2y_1^2 + 5y_1y_3 - y_3^2 + 6) + 3u_{21}(y_1^2 - y_3^2)]}{T(u_{11}u_{21} + 4u_{11}u_{31} + u_{21}u_{31})^2}$$

$$F_2'(y) = \frac{24u_{12}u_{22}u_{32}y_2(y_1u_{32} + y_3u_{12})}{T(u_{12}u_{22} + 4u_{12}u_{32} + u_{22}u_{32})^2}$$

$$F_3'(y) = \frac{2u_{13}u_{23}u_{33}[4u_{13}(2y_3^2 + 5y_3y_1 - y_1^2 + 6) + 3u_{23}(y_3^2 - y_1^2)]}{T(u_{13}u_{23} + 4u_{13}u_{33} + u_{23}u_{33})^2}$$

$$
\begin{aligned}
|F_1'(y)| &= \left|\frac{2u_{11}u_{21}u_{31}[4u_{31}(2y_1^2+5y_1y_3-y_3^2+6)+3u_{21}(y_1^2-y_3^2)]}{T(u_{11}u_{21}+4u_{11}u_{31}+u_{21}u_{31})^2}\right| \\
&\leq \left|\frac{2[4u_{31}(2y_1^2+5y_1y_3-y_3^2+6)+3u_{21}(y_1^2-y_3^2)]}{T(u_{21}+4u_{31})}\right| \\
&< \left|\frac{2[(2y_1^2+5y_1y_3-y_3^2+6)+3(y_1^2-y_3^2)]}{T}\right| \\
|F_3'(y)| &< \left|\frac{2[(2y_3^2+5y_3y_1-y_1^2+6)+3(y_3^2-y_1^2)]}{T}\right|
\end{aligned}
$$

$$
\begin{aligned}
|F_2'(y)| &= \left| \frac{24u_{12}u_{22}u_{32}y_2(y_1u_{32}+y_3u_{12})}{T(u_{12}u_{22}+4u_{12}u_{32}+u_{22}u_{32})^2} \right| \\
&\leq \left| \frac{24u_{12}u_{22}u_{32}y_2\sqrt{u_{32}^2+u_{12}^2}\sqrt{y_1^2+y_3^2}}{T(u_{12}u_{22}+4u_{12}u_{32}+u_{22}u_{32})^2} \right| \\
&\leq \left| \frac{3y_2\sqrt{u_{32}^2+u_{12}^2}\sqrt{y_1^2+y_3^2}}{2T(u_{12}+u_{32})} \right| \\
&< \left| \frac{3y_2\sqrt{u_{32}^2+2u_{32}u_{12}+u_{12}^2}\sqrt{y_1^2+y_3^2}}{2T(u_{12}+u_{32})} \right| \\
&\leq \left| \frac{3y_2\sqrt{y_1^2+y_3^2}}{2T} \right| \\
&= \left| \frac{3y_2\sqrt{3-y_2^2}}{2T} \right|
\end{aligned}
$$

- If we choose T such that

$$T \geq \max\left\{ \frac{3}{2}\left|y_2\sqrt{3-y_2^2}\right|, 2\left|(2y_1^2+5y_1y_3-y_3^2+6)+3(y_1^2-y_3^2)\right|, 2\left|(2y_3^2+5y_3y_1-y_1^2+6)+3(y_3^2-y_1^2)\right| \right\},$$

  then the iterative process is a contraction mapping. And the non-blurring SUP will converge.

- Consider a trivial situation $y_1 = -y_3$, $y_2 = 0$. Theoretically, the process should converge for any T . Hence, the points will not move in each iteration. Back to the criteria of T, $\frac{3}{2}\left|y_2\sqrt{3-y_2^2}\right| = 0$ in this case, but the other 2 values equal to $12 - 8y_1^2$ could be positive. From this point of view, the condition of T we provide to reach the convergence may be too strict.

Baseball Players' Salaries

- This data contains 337 major league baseball players' salaries (measured in thousands of dollars) in the year 1992 and their 16 performance measures from the year 1991.
- From the Bayesian variable selection analysis in Lee et al.(2016), they conclude that RBI is a crucial factor for a baseball player to achieve a high salary.

Figure: SUP on real data

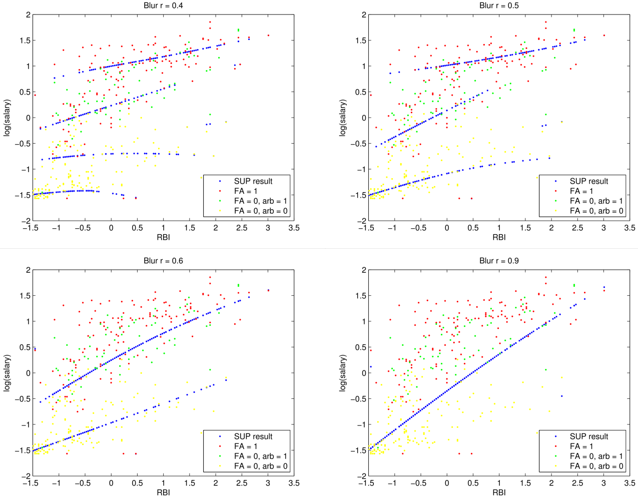
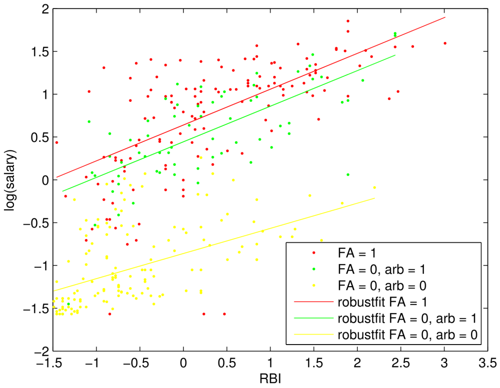


Figure: Robustfit on real data



Discussion and Future Work

1. The conjecture about convergence in blurring SUP in regression is left to be solved.
2. Deal with non-separable data.
3. Data points are not enough.
4. Generalize this approach for other statistical models.
5. Derive some theoretical properties of the estimate; moreover, give the confidence interval to make inferences.
6. Improve the computational speed.

Thank you!