Robustness of Nonparametric Regression Using Kernelized Depth Function

August 5, 2017

We propose a novel method to reduce the impact of outliers on a given data set using kernelized spatial depth (see Chen et al. (2009)) and the Nadaraya-Watson estimator (see Nadaraya (1964) and Watson (1964)). An 'outlier' is an observation that lies at an abnormal distance from other values in a random sample from a population. Generally, outliers are expected to appear more likely in outer layers with small depth values, than in inner layers with large depth values.

Spatial Median and Depth

A median is insensitive to outliers as compared to the mean. The multi-dimensional sample median for multi-dimensional data $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ is any $\mathbf{x} \in \mathbf{R}^d$ that satisfies

$$\| \sum_{i=1}^{n} S(\mathbf{x}_i - \mathbf{x}) \| = 0,$$

where $\| \cdot \|$ is the Euclidean distance. The spatial depth of $\mathbf{x}$ for an unknown cdf F is defined as follows (see Vardi and Zhand (2000) and Serfling (2002)):

$$D(x, \chi) = 1 - \frac{1}{\mid \chi \cup \{\mathbf{x}\} \mid -1} \| \sum_{y \in \chi} S(\mathbf{y} - \mathbf{x}) \|.$$

Here, χ is the collection of sample points. The spatial depth value at median is 1. Note that this depth function depends only on the sum of unit vectors from point under consideration to the remaining data set.

Kernelized Spatial Depth

We now consider a kernelized version of spatial depth as follows:

$$D_K(\mathbf{x}, \chi) = 1 - \frac{1}{\mid \chi \cup \{\mathbf{x}\} \mid -1} \left(\sum_{\mathbf{y}, \mathbf{z} \in \chi} \frac{K(\mathbf{x}, \mathbf{x}) + K(\mathbf{y}, \mathbf{z}) - K(\mathbf{x}, \mathbf{y}) - K(\mathbf{x}, \mathbf{z})}{\delta_K(\mathbf{x}, \mathbf{y}) \delta_K(\mathbf{x}, \mathbf{z})} \right)^{1/2},$$

where $K(\mathbf{x}, \mathbf{y}) = exp(-\|\mathbf{x} - \mathbf{y}\|^2 / \sigma^2)$ and $\delta_K(\mathbf{x}, \mathbf{y}) = \sqrt{K(\mathbf{x}, \mathbf{x}) + K(\mathbf{y}, \mathbf{y}) - 2K(\mathbf{x}, \mathbf{y})}$.

Note that $\dfrac{K(\mathbf{x}, \mathbf{x}) + K(\mathbf{y}, \mathbf{z}) - K(\mathbf{x}, \mathbf{y}) - K(\mathbf{x}, \mathbf{z})}{\delta_K(\mathbf{x}, \mathbf{y}) \delta_K(\mathbf{x}, \mathbf{z})} = 0$ for $\mathbf{x} = \mathbf{y}$ or $\mathbf{x} = \mathbf{z}$.

All the above calculations gives us the weight on each of the observation points. If the point lies far away from the data sets, then it has lower weight while if it lies closer to the data sets, then it has higher weight. The parameter σ determines the size of the neighbourhood that is used to compute KSPD for an observation. When σ is 0 then KSPD is constant, and when σ is infinity KSPD goes to spatial depth.

Nonparametric Regression

If we have a random sample of bivariate data $(Y_1, X_1), \ldots, (Y_n, X_n)$. The regression model is

$$Y = m(X) + e,$$

where m is an unknown function and e is the random error. Nadaraya and Watson independently proposed the following estimator:

$$m_{NW}(x) = \frac{\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) Y_i}{\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)},$$

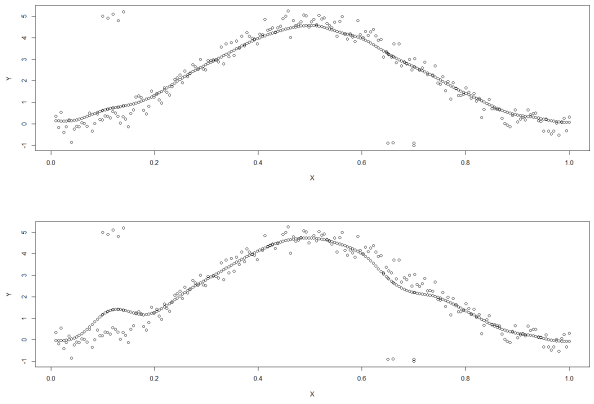
where $K(u) = exp(-u^2)$ and $h > 0$ is the smoothing parameter.
We modify the Nadaraya-Watson estimator by applying weights to each of the data points as follows:

$$m_1(x) = \frac{\sum_{i=1}^{n} D_K(\mathbf{z}_i, Z) K\left(\frac{X_i - x}{h}\right) Y_i}{\sum_{i=1}^{n} D_K(\mathbf{z}_i, Z) K\left(\frac{X_i - x}{h}\right)},$$

where $\mathbf{z}_i = (y_i, x_i)$ for $i = 1, \ldots, n$ and Z is the collection of n data points. This method is well-defined for more feature variables, i.e., multivariate Y as well as multivariate X.

Simulations

For the figure on left we have taken 209 observations in the interval [0,1] with 9 outliers and a polynomial mean function $y = 300(x^3 - 3x^4 + 3x^5 - x^6)$ with normally distributed errors with variance σ^2=0.3^2 and mean 0.
For the figure on right we have again taken 219 observations in the interval [0,1] with 19 outliers and a mean function $y = 400(100x^3 + 50sin(x) - 200)$ with normally distributed errors with variance σ^2=100 and mean 0.



The lower plots are obtained by the Nadaraya-Watson estimator, and the *bulge on the curve* shows the limitation of Nadaraya Watson method on introducing outliers to the data set while the upper plots are obtained by the proposed method.

The simulation was done by varying the parameters σ and h. We varied σ between 0 and infinity to check the impact of the outliers. The smoothing parameter h is obtained by minimizing the error term between the estimated value and the given value.

Ongoing Work

- We are still working on the part with smoothing parameters. We will estimate it by minimising the leave-one-out cross validation which is defined by

$$CV = \frac{1}{n}\sum_{i=1}^{n}(Y_i - m_{(-i)}(x_i))^2$$

  where $m_{(-i)}(x_i)$ is the estimator obtained by omitting the i^{th} pair (X_i, Y_i).

- It will be interesting to see how this method works in higher dimensions.

- We also want to investigate the performance of the $0-1$ weight function.

References

Chen, Y., Dang, X., Peng, H., and Bart, H. L. (2009) Outlier detection with the kernelized spatial depth function. IEEE Transactions on Pattern Analysis and Machine Intelligence, 31(2), 288-305.

Nadaraya, E. A. (1964) On estimating regression. Theory of Probability and Its Applications, 9(1), 141-142.

Serfling, R. (2002) A depth function and a scale curve based on spatial quantiles. In Statistical Data Analysis Based on the L1-Norm and Related Methods (pp. 25-38). Birkhuser Basel.

Vardi, Y., and Zhang, C. H. (2000). The multivariate L1-median and associated data depth. Proceedings of the National Academy of Sciences, 97(4), 1423-1426.

Watson, G. S. (1964) Smooth regression analysis. Sankhya: The Indian Journal of Statistics (Series A), 26(4), 359-372.