MULTIPLE LINEAR REGRESSION VIEWPOINTS
VOLUME 17, NUMBER 2, FALL 1990

Two Stage Smoothing of Scatterplots

Timothy H. Lee, Ph.D.
California State University

Donald T. Searls, Ph.D.
University of Northern Colorado

Abstract

Scatterplot smoothing is a simple but a very useful tool for data analysis. A smooth curve superimposed on the scatterplot greatly enhances the visual information, especially, the bivariate association between the prediction variable and the response variable. In this article some smoothers are reviewed with respect to consistency and sensitivity to discontinuities on the underlying functions. Robust centered span smoothers produce smooth and consistent curves but they tend to smooth over or blur the discontinuities. Non-centered span smoothers are sensitive to the discontinuities but they tend to be rough and lack consistency. Two stage smoothing is proposed as a technique that provides consistency as well as sensitivity to discontinuities.

Key words: smoother, underlying function, discontinuity, consistency, centered span, non-centered span

1. Introduction.

Scatterplots are a very useful tool for analyzing a bivariate relationship between two variables, say X and Y. The observed bivariate data points,

$$(x_1, y_1),\ (x_2, y_2),\ \ldots,\ (x_n, y_n),$$

constitute acatterplots. They visually explain the relationship. It was pointed out by Cleveland (1979) that the extreme points in the point cloud of acatterplots diatract the eyes and they tend to miss the structure of the bulk of the data. As a remedy,scatterplots are smoothed, then the visual information is enhanced and the association between the two variables is clarified. Unfortunately, if discontinuities are present the smooth curve may tend to conceal this fact. If the smoothers are sensitive to discontinuities they tend to be somewhat rough. Two stage smoothing is proposed as a technique that tends to provide smooth fits with detection of discontinuitiea.

Scatterplot smoothing is a procedure that operates over the bivariate data points to decompose the observed y_i values into two parts, System (or Smooth) and Noise (or Rough). That is, the i-th observed value of Y can be written as

$$y_i = s(x_i) + r_i,$$

where s is a system or a smoothing function and r_i is a residual (or rough). Here, we assume that y_i is generated from an underlying function and noise with a certain distribution. That is,

$$y_i = f(x_i) + e_i .$$

The underlying function $f(x_i)$ is estimated by $s(x_i)$ in the smoothing procedure. The requirement of a good smoother is that it should not be affected by occasional outliers and the output results should be smooth regardless of the input data. In this regard, Cleveland (1979) proposed Locally Weighted Regression Scatterplot Smoothing ("LOWESS") which meets the robustness condition of good smoothers. Friedman (1984) proposed a variable span smoother in which local cross validation is used to estimate the optimal span as a function of the abscissa value. McDonald and Owen (1984) proposed a split linear fit smoothing algorithm that can produce discontinuous output. It can be used for smoothing with edge detection. One feature of the split linear fit method that distinguishes it from most of the other smoothers is that it uses non-centered spans.

One of the problems encountered in smoothing scatterplots is how to estimate, as closely as possible, the $f(x)$ by $s(x)$ using the given scatterplots. Therefore, a good smoother should be robust and consistent. When the underlying function, $f(x)$, is smooth (continuous) most of the centered span smoothers perform well. However, if $f(x)$ is discontinuous or kinked, the centered span smoothers usually blur the discontinuous points and produce a smooth curve; while the non-centered span smoothers are quite sensitive to discontinuities.

In this study, the smoothers sensitive to the discontinuities, namely, the non-centered span smoother, running medians of three, and Tukey's 3RSSH, are compared for consistency. Also, an exploration was made of a two-stage smoother that is more consistent but at the same time can produce a discontinuous curve.

For computational economy, the updating formula of the sample variance proposed by Chan, T. et al (1980) were used to update the regression parameter estimations.

Next, we discuss smoothers with two different types of spans and consider detection of the discontinuities of $f(x)$.

2. Centered Span Smoother.

The centered span smoother is the most commonly used smoother. To estimate $f(x_i)$ take a number of observations around x_i so that x_i is a center of the observations. These observations constitute a span for x_i. Cleveland's LOWESS, Running Median, Moving Average, and 3RSSH are examples of the centered span smoother. Here, as a centered span smoother, we use a robust fixed span smoother which is similar to LOWESS. The basic procedure is:

(a) Find initial fitted value $\hat{y}_i$ for x_i by using local linear regression.

Fit a simple local straight line to the data in the span for x_i, $i = 1,\ldots,n$.
Then, find the initial smooth value $\hat{y}_i$, $i = 1,\ldots,n$. (Updating formula can be used with unit weight.)

(b) Depending on the residual ($r_i = y_i - \hat{y}_i$) for each x_i, assign a weight.

A weight for each x_i is based on each r_i.

Let $m = \text{Median}\{|r_i|, i = 1,\ldots,n\}$, and let $d_i = r_i/(6*m)$.

Then, the weight for the k-th observation in the span for x_i will be

$$w_k(x_i) = \begin{cases} (1 - d_i^2)^2 & \text{for } |d_i| \le 1 \\ 0 & \text{otherwise.} \end{cases}$$

(c) Based on the new weight, fit a locally weighted straight regression line.

(d) Repeat steps (b) and (c) until the convergence criterion, $|\hat{y}_{old} - \hat{y}_{new}|/|\hat{y}_{old}| < \sigma$ is satisfied.
In this study, $\sigma = 10^{-5}$ is used.

This procedure is applied for three different sizes of spans in order to give points on the boundaries of the span less weight than the points in the center. So, three values (i.e., $\hat{y}1_i$, $\hat{y}2_i$, $\hat{y}3_i$) for x_i are computed. The weight for each estimate is given depending on the span size. Let w1, w2, and w3 be weights for each of 3 spans. Then, the final smooth value for x_i will be obtained by,

$$\hat{y}_i = w1\hat{y}1_i + w2\hat{y}2_i + w3\hat{y}3_i,$$

where $w1 + w2 + w3 = 1$,

and

$w1 > w2 > w3$,

if the relationships among the spans are

span 1 < span 2 < span 3 .

In this study, the three spans used are 18, 20, and 22, respectively.

The advantages of this procedure are:

(a) It is computationally effective in terms of number of operations.

(b) It is more robust than a simple local straight line fit.

(c) Using a straight line reduces computational cost and makes the updating easier.

As seen in Figure 1, this smoother blurs the discontinuous points and produces an overall smooth curve. Running medians of three (referred to as "3R") and 3RSSH are also simple centered span smoothers. They are quite sensitive to discontinuities but produce rough (or bumpy) fits to the data.

3. Non-centered Span Smoother.

Unlike most of the smoothers, spans for x_i are not set up such that x_i is the center of a span. For example, McDonald and Owen's (1984) split linear fit smoother is such a smoother. They pointed out the weakness of the centered span smoothers and proposed a smoother that can be used for smoothing with edge detection. The idea is to make several linear fits for x_i; some of them are left-sided fits, some are central fits, and some are right-sided fits. In practice, three linear fits (one for each type of fit) are enough. Then, the three estimated values from the three types of fits are assessed depending on the basis of the mean squared residual about the line fitted over all of the data except x_i (referred to as "PMSE"). Any fitted value with PMSE greater than the average PMSE for x_i is ignored. Weights for the remaining fitted values are based on the squared differences between each PMSE and the average PMSE. Using these remaining fitted values and their respective weights, a weighted average is computed as a fitted value for x_i.

This smoother is very sensitive to discontinuities but there is a tendency for this smoother to produce a curve with a somewhat jagged appearance. This problem can be solved to some extent by applying the above algorithm repetitively to its own output. In this study, it is repeated once to avoid possible digression of the fitted curve from the underlying function f(x). See Figure 2. In this study, the span size for this smoother is 20.

4. Measurement of Consistencies.

To compare the consistencies of smoothers it is necessary to quantify them. A possible candidate to measure consistency is the average of the sample variances of the B fitted values for each x_i. Efron (1990) presented an example for a bootstrap estimate for the variance of regression coefficients. A similar idea is applied in this study as follows.

First, assuming that the underlying function is not known, apply a smoother on a generated data set and find

$s(x_i)$ and $r_i = y_i - s(x_i)$, $i = 1,\ldots,n$.

Then,

(a) Construct $\hat{F}$ by assigning $1/n$ as the weight for the residual, r_i.

(b) Draw a bootstrap data set

$y_i^* = s(x_i) + r_i^*$, $i = 1,\ldots,n$,

where r_i^*'s are i.i.d. from $\hat{F}$.

Then,

$s^*(x_i)$, $i = 1,\ldots,n$

are computed on

y_i^*, $i = 1,\ldots,n$.

(c) Independently repeat step (b) B times, obtaining bootstrap replications,

$s^{*1}(x_i), s^{*2}(x_i), \ldots, s^{*B}(x_i)$, $i = 1, \ldots, n$.

Then, compute

$$CM1 = \frac{1}{Bn}\sum_{b=1}^{B}\sum_{i=1}^{n}[s^{*b}(x_i) - s^{*}(x_i)]^2,$$

where

$$s^{*}(x_i) = \frac{1}{B}\sum_{b=1}^{B}[s^{*b}(x_i)].$$

And

$$CM2 = \frac{1}{Bn}\sum_{b=1}^{B}\sum_{i=1}^{n}[s^{*b}(x_i) - f(x_i)]^2$$

where f is the underlying function.

CM1 measures the consistencies (variation) of the smooth curve around the mean smooth curve and CM2 measures the consistencies around the underlying function. CM2 is measurable only when the underlying function is known. If the underlying function is known, it is more reasonable to use the e_i's rather than r_i's and $f(x_i)$ rather than $s(x_i)$ for step (c) in the above procedure to compare consistency. The reason is that the values of the r_i's depend on the sensitivity of smoothers to discontinuities. In Tables 1 - 4, such measures are computed for comparison of the consistency of smoothers.

5.Smoothing with Detection of the Discontinuities and Improved Consistency

We have seen that the non-centered span smoother is sensitive to the discontinuities, while the centered span smoothers blur them. By using this fact we can detect discontinuities simply by plotting the differences of the two smooth values estimated by the non-centered span smoother and by the centered span smoother. Figure 3 presents the two smooth curves for the purpose of visual comparison. The underlying function in Figure 3 is a sawtooth function. Figure 4 presents the difference plot. A discontinuity is suspected at the local maxima or minima. In the figure, a discontinuity is suspected around x = 50. Also, the difference plot shows the overall pattern of the discontinuity.

We are interested in consistency and, at the same time, in the detection of discontinuities. If a smoother has both properties, the computed values of CM1 and CM2 for that smoother will be lower than those of other smoothers. From Tables 1 - 4, we see that the robust centered span smoother has better consistency than the non-centered span smoother, but the latter has more sensitivity to discontinuities. The problem is how to combine the two desirable properties. One solution is to use two-stage smoothing. In the first step, discontinuities are located and the original data set is split such that each discontinuity serves as a splitting point. In the second step, the robust centered span smoother is applied to each of the split data sets. The consistency measurements of this smoother are shown in Tables 2 and 3 and the smooth curves produced by this method is shown in Figure 5.

6. Discussion.

In this study, the consistency measures of various smoothers are compared. The results show that:

(1) The non-centered span smoother is sensitive to discontinuities and less consistent than the robust centered span smoother,

(2) The robust centered span smoother lacks sensitivity to discontinuities but it is very consistent;

(3) Other sensitive smoothers, such as running medians of three or 3RSSH, produce quite rough curves and lack consistency; and

(4) The two-stage smoother is consistent and produces smooth curves with edge detection.

The detection and the location of the discontinuities on the x-axis are dependent upon the span size of the smoother. The determination of the span size is very important. If the span size is large, then the robust centered span smoother will blur the discontinuities. If x_i is close to a discontinuity, then the difference between the values estimated by the non-centered span smoother and the robust centered span smoother will be large. If the non-centered span smoother has a wide span it tends to ignore the discontinuities, while a narrow span will make it unnecessarily sensitive and may result in false detection of discontinuities. If there are more than one discontinuity on the underlying function the distance between any two discontinuities must be larger than the span size in order to be detected.

Sometimes outliers make the detection of discontinuity very difficult. Outliers near the discontinuities may cause confusion and lead to poor decisions. One possible remedy is to apply the running medians as a filter before the two-stage smoother is applied. The two-stage smoother works well when the discontinuities are separated enough and the functional form of the underlying function is not complicated. It works best when the underlying function is smooth but broken by discontinuities, for example, a saw tooth function. When no discontinuities are detected the two-stage smoother is the same as the robust centered span smoother. The two stage smoother has the advantages of being able to detect discontinuities as well as being very consistent.

Acknowledgements

The authors wish to acknowledge Dr.Sam Houston of University of Northern Colorado for his careful review of the manuscript. The authors also appreciate Arline Nakanishi and John Lichtenstein for their help in the preparation of the document.

References

Chan, T. F., Golub, G. H., & Leveque, R. J. (1980),"Updating formulae and pairwise algorithm for computing sample variances," DAAG29-78-G-0179.

Cleveland, W. S. (1979),"Robust locally weighted regression and smoothing scatterplots," Journal of the American Statistical Association, 74, 829-836.

Efron, B. (1990), "More Efficient Bootstrap Computations", Journal of the American Statistical Association, 85, 79-89.

Friedman, J. H. (1984), "A variable span smoother," LCS Technical Report No. 5.

Hoaglin, D. C., & Velleman, P. F. (1981). Applications, Basics, and computing of exploratory data analysis. Boston: Duxbury Press.

McDonald, J. A., & Owen, A. B. (1984). "Smoothing with split linear fits." Technical Report, Stanford University, Dept. of Statistics.

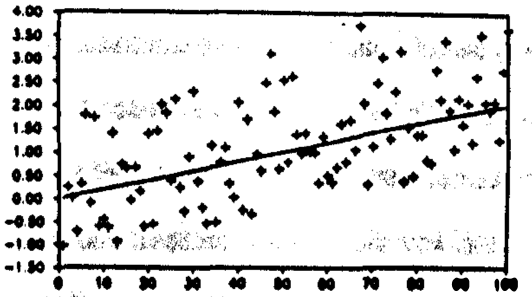
Mosteller, F., & Tukey, J. W. (1977). Data Analysis and Regression, Menlo, CA: Addison Wesley, Inc.

Appendix

A.Tables

Table-1.

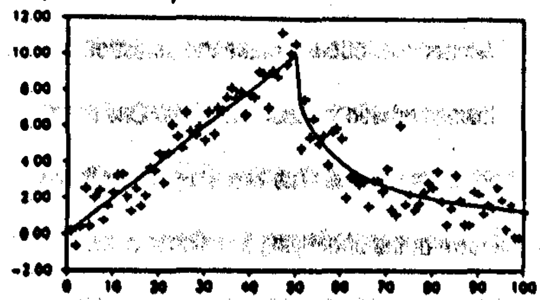
(f(x) =0.02x, monotone increasing function)



Smoother	CM1	CM2
Robust Centered Span	0.05721	0.07972
Non-centered Span	0.13593	0.17323
3RSSH	0.47333	0.64696
3R	0.49667	0.71916

Table-3.

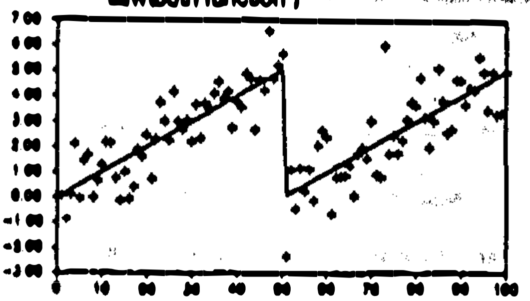
(f(x)=0.2x, for x≤50, f(x)=80/(x-40), for x>50, fin-shaped function)



Smoother	CM1	CM2
Robust Centered Span	0.06246	0.47491
Non-centered Span	0.20740	0.23504
3RSSH	0.37346	0.49477
3R	0.35296	0.51365
Two-stage	0.07766	0.10650

Table-2.

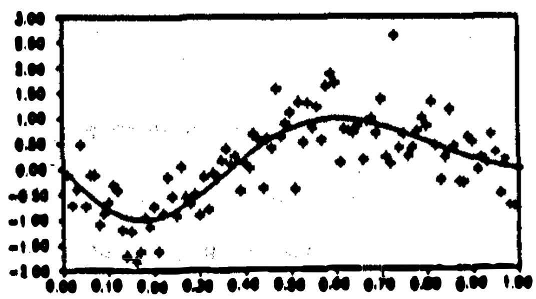
(f(x)=0.1x, for x≤50, f(x)=0.1x(x-50), for x>50, sawtooth function)



Smoother	CM1	CM2
Robust Centered Span	0.06462	0.76905
Non-centered Span	0.20362	0.39049
3RSSH	1.63010	2.14755
3R	1.66922	2.34655
Two-stage	0.05671	0.31377

Table-4.

(f(x)=sin[2π(1-x)$^{1.5}$], expandig cyclical function)



Smoother	CM1	CM2
Robust Centered Span	0.06246	0.47491
Non-centered Span	0.20740	0.23504
3RSSH	0.37346	0.49477
3R	0.35296	0.51365

B. Figures

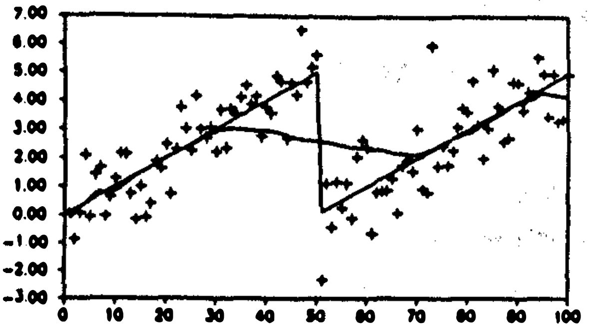


Figure 1. Smooth by Robust Centered Span Smoother.

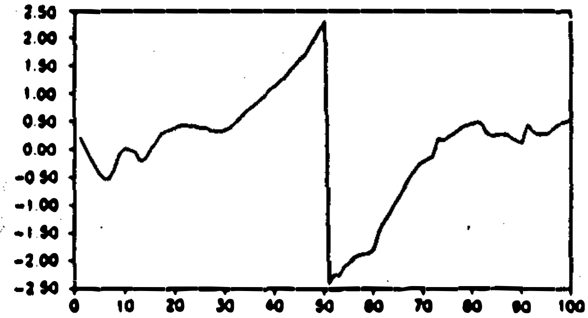


Figure 4. Differences of two smooth curves by Robust Centered Span smoother and Non-centered Span smoother

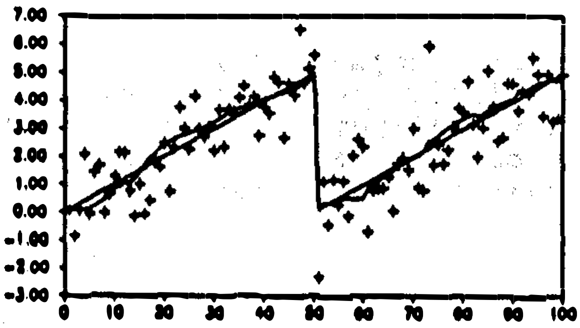


Figure 2. Smooth by Non-centered Span smoother.

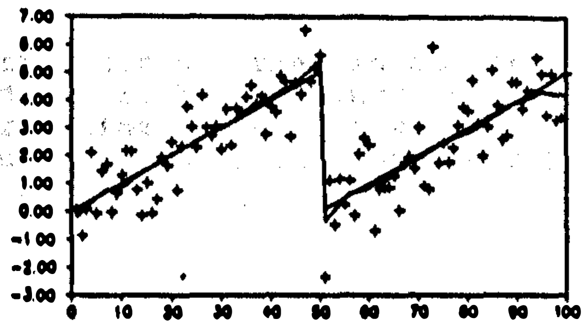


Figure 5. Smooth by Two-stage smoother

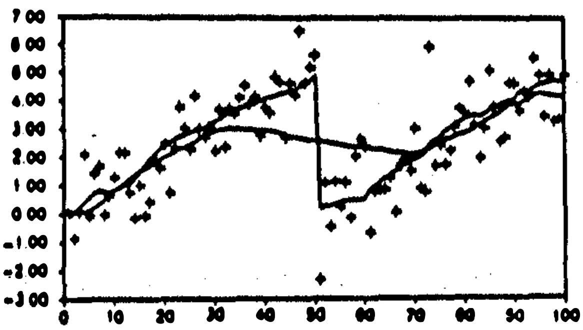


Figure 3. Comparison of two smooth curves by Robust Centered span smoother and Non-centered span smoother.