Eidgenössische
Technische Hochschule
Zürich

Ecole polytechnique fédérale de Zurich
Politecnico federale di Zurigo
Swiss Federal Institute of Technology Zurich

Seminar für Statistik

Research Note:

Robustifying a (Local, Nonparametric) Regression Estimator

Martin Maechler
May 1989,

April 1992

This technical note is basically section 5.5 from my Ph.D. thesis, Mächler (1989). The goal is to make the following findings more widely available. I assume here that the reader is familiar with Cleveland (1979), the first journal paper to introduce LOWESS. Note that since, Cleveland and several coworkers have continued to research and improve the local regression methodology, notably allowing for multiple carriers and also working out methods for inference, see Cleveland and Devlin (1988), or chapter 8 of Chambers and Hastie (1991). Also, the name of LOWESS has been changed to *"loess"*. However, nothing has been changed in the algorithm used to robustify LOWESS/loess, and here, I am only considering these robustness properties. Therefore, all the subsequent references to "LOWESS" equally apply to "loess".

My goal here — besides analyzing the breakdown behavior of LOWESS — is to give a general procedure for robustification of regression estimators of a very general class. Therefore, I treat the local regression smoother, e.g., loess, as a given "blackbox" algorithm. For the given data x_i, y_i and weights w_i, for $i = 1, \ldots, n$, I only assume that we dispose of a "smoother" $\mathbf{S}$, i.e., a (nonparametric or parametric) regression algorithm which computes fitted values,

$$\hat{\mathbf{y}} = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n)^\top = \mathbf{S}(\mathbf{x}, \mathbf{y}, \mathbf{w}), \tag{1}$$

where boldface (lower case) letters denote vectors (of length n), e.g., $\mathbf{y} = (y_1, y_2, \ldots, y_n)^\top$.

The basic idea is to do reweighting iterations on this estimate, i.e., re-setting the weights w_i, by multiplying the original weights by robustness weights that are computed from the residuals of the estimate of the previous iteration. If we let $\mathbf{r} := \mathbf{y} - \hat{\mathbf{y}}$ denote the vector of residuals, the robustness weights are defined as $w_i^R = B(r_i/s(\mathbf{r}))$, where $s()$ is a robust estimate of scale. The downweighting function B is related to the ψ function which defines a (robust) M-estimator (e.g., in linear models, see Hampel et al. (1986)), by $B(x) = \psi(x)/x$.

Note that this reweighting procedure is *the* classical robustness algorithm of Huber (1964) where it was applied for the special case of location ("constant regression"). Further note that here, it is much simpler and (computationally) more efficient than to use separate robust regression estimates at each location x_i (using the data in a "neighborhood" of x_i).

1. LOWESS is not robust

This procedure of iterated reweighting (for smoothing) was proposed already by Cleveland (1979) for his "LOWESS" smoother which does local *linear* regression. He uses Tukey's "bi-weight" function for down-weighting, which corresponds to a 're-descending' ψ-function of the corresponding M-estimator. It is known (see, e.g., Hampel et al. (1986)) that in this case the M-estimators are *not* uniquely defined, but depend algorithmically on the starting value of a kind of Newton-algorithm to find the estimate. Here, Cleveland chose to start with simple least-squares which considerably spoils the intended robustness:

> *The breakdown point of LOWESS is very low.*
> *It is practically zero when its smoothing parameter f is less than* 0.5.

Consider the following example. Given are 225 data points $x_i = i$ and $y_i = \delta_{i,113}$. This means constant data with a single outlier in the middle. This is not a special configuration, since the same phenomenon occurs if, e.g., $y_i \sim U[0,1]$, for $i \neq 113$, and $y_{113} = 10^7$. For $f \leq .506$ where $f = h/n$ and h is the bandwidth, LOWESS *interpolates* all the data including the outlier in the constant case and it still interpolates the outlier when the y_i's are jittered. What happens ? The initial non-robust least-squares estimator downweights 110 of the 'good' neighbors and the outlier to zero because the 'MAD' (median of the absolute residuals) scale estimate is much smaller than the least- squares residuals, there. This gives (nearly) zero weights for the whole center batch of the data which entails interpolation for the outlier, there. In figure 1, I give the table of LOWESS parameter

`f`	N
0.5	1500
0.506	225
0.51	149
0.52	73
0.55	29
0.6	13
0.65	9
2/3	7

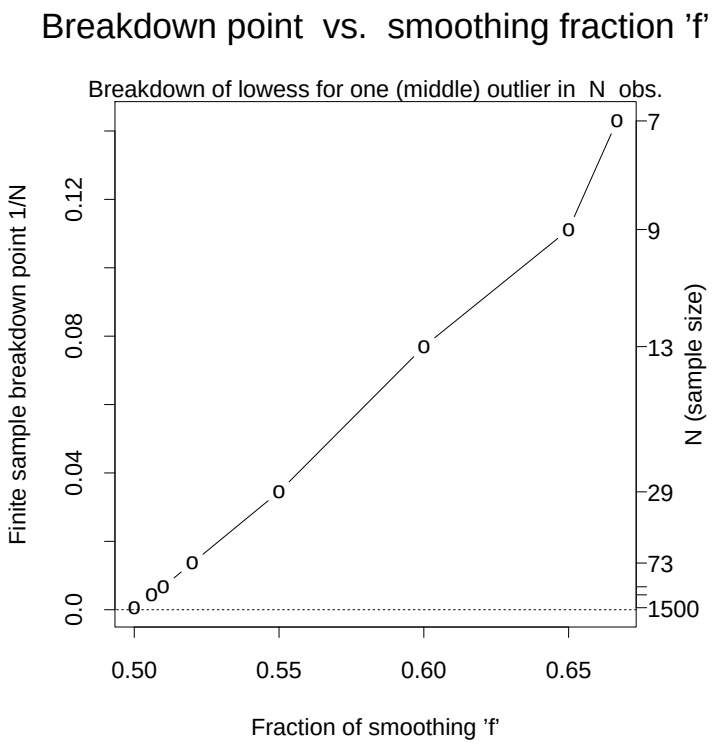

Figure 1: Breakdown point vs. smoothing fraction `f`

values `f` with corresponding sample size N for which LOWESS breaks down.

In more detail: Given the sample size N, the data is defined as follows. For $i = 1, \ldots, N$, let $x_i := i$ and y_i i.i.d.,$y_i \sim U[0,1]$ (uniformly distributed in $[0,1]$). Let $i_{\text{mid}} := 1 + \lfloor N/2 \rfloor$, then we put $y_{i_{\text{mid}}} := 10^{10}$, i.e., a single outlier. Now, we say that LOWESS breaks down, if after 3 re-weighting iterations, $max_i \hat{y}_i$ is still larger than 100.

In practice one would encounter this fact (accidental breakdown of LOWESS) quite rarely, since LOWESS is often used to 'enhance scatterplots' where extreme outliers are mostly rejected already before. Indeed, the breakdown point of any such local robust smoother will depend on the smoothing parameter in some way, but for LOWESS, e.g., one could (and should) require that the breakdown point is about $f/2$, since the local estimate uses $f \times 100\%$ of the data points.

2. Proposal for proper robustification

It is easy to achieve this desired breakdown point, 'mending' LOWESS by at least one of the following methods:

- Use another function for down-weighting, namely one which corresponds to a *monotone* ψ-function.
- Do not start with a least-squares sweep, but with a sweep by a simple but highly robust smoother (with the appropriate equivalent smoothing parameter) to get 'conservative' robustness weights.

If only the first remedy is used, it may be necessary to do many iterations until convergence, whereas applying only the second one keeps the problem of possibly non-unique solutions.

We propose and implemented the following procedure. We want to have the possibility to reject outliers completely. This can be accomplished by redescending M-estimators (see Hampel et al. (1986), Chapter 6 sect. 2.6). We use a 'Hampel' ψ which is simple and behaves nearly as the 'optimal V-robust' tanh-ψ. The latter is not very practical because its parameters are given only implicitly. The 'Hampel' ψ is given as

$$\psi_{a,b,r}(x) \stackrel{\text{def}}{=} \text{sign}(x) \ \min\left(|x|, a, \max\left(\frac{r-x}{r-b}, 0\right)\right) \tag{2}$$

with $0 < a \le b < r < \infty$

Being piecewise linear, it is a 'huber' ψ_a ($\psi_c(x) := \min(c, \max(-c, x))$) in the middle, i.e., for $x \in [-b, b]$, and then 'redescends' to become zero for $x \ge r$.

The cut-off parameter c of these ψ functions is a multiple of a robust scale measure of the residuals, in practice. For Hampel's redescenders, we propose to use the one parameter sub-class $\psi_c^{FH} := \psi_{c,2\min(2,c),4\min(2,c)}$, see below.

The following parameters for our robust (pre-)smoother proposal have to be specified:

- Bandwidth "h"
- Polynomial degree "deg"
- Number of "Huber" reweighting iter. "Roblt1"
- Number of "Hampel" reweighting iter. "Roblt2"
- Constant(s) for Huber's and Hampel's ψ (2) "k", or "a, b, r", resp.

1. Start with a 'most robust' smoother: Do a sweep by Tukey's 'running medians of M', where $M \approx \frac{2h}{x_n - x_1} \cdot n$. Call this estimate $\hat{f}_M$.

In order to diminish erosion apply "twicing", a special case of Tukey's "re-roughing". This is smoothing the residuals of $\hat{f}_M$ with the same smoother and add these smoothed residuals to $\hat{f}_M$ to give the new $\hat{f}^*_M$.

Apply "hanning", a simple running weighted mean with weights $\frac{1}{4}[1, 2, 1]$. This leads to smoother residuals which is useful for the scale estimate below if many residuals were zero, before. This is not too rare with running medians!

2. Compute robustness weights from the residuals of 1., using Huber's ψ and the median of the absolute residuals ('MAD') as scale estimate. We chose $k = 2 \cdot MAD$. This corresponds to $k = 1.35$ for scale estimates which are consistent at normal errors.

3. Do Roblt1 (we set Robit1:=2) sweeps of local polynomial regression with polynomials of degree deg and using bandwidth h. In each sweep, use the robustness weights of the previous sweep, and compute new 'Huber weights'.

4. Do Roblt2 (we set Roblt2:=1) sweeps of local polynomial regression, as above, but now computing 'Hampel weights' which reject far outliers. As constants for Hampel's three-part redescending ψ, we take $(a, 4, 8)$ if $0 \le a \le 2$, and $(a, 2a, 4a)$ if $a > 2$. The default, $a = 2$, leads to the "2-4-8-estimator", the simplest recommended default choice of a three-part redescender (see Hampel et al. (1986), section 1.1d).

Bibliography

Chambers, J. M. and Hastie, T. J. (1991). *Statistical Models in S.* Computer Science. Wadsworth.

Cleveland, W. S. (1979). Robust Locally Weighted Regression and Smoothing Scatterplots. *Journal of the American Statistical Association*, 74:828–836.

Cleveland, W. S. and Devlin, S. J. (1988). Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting. *Journal of the American Statistical Association*, 83:596–610.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions.* Wiley.

Huber, P. J. (1964). Robust Estimation of a Location Parameter. *Annals of Statistics*, 35:73–101.

Mächler, M. B. (1989). *'Parametric' Smoothing Quality in Nonparametric Regression: Shape Control by Penalizing Inflection Points.* Ph.d. thesis, no 8920, ETH Zurich, Statistik, ETH-Zentrum, CH-8092 Zurich, Switzerland.