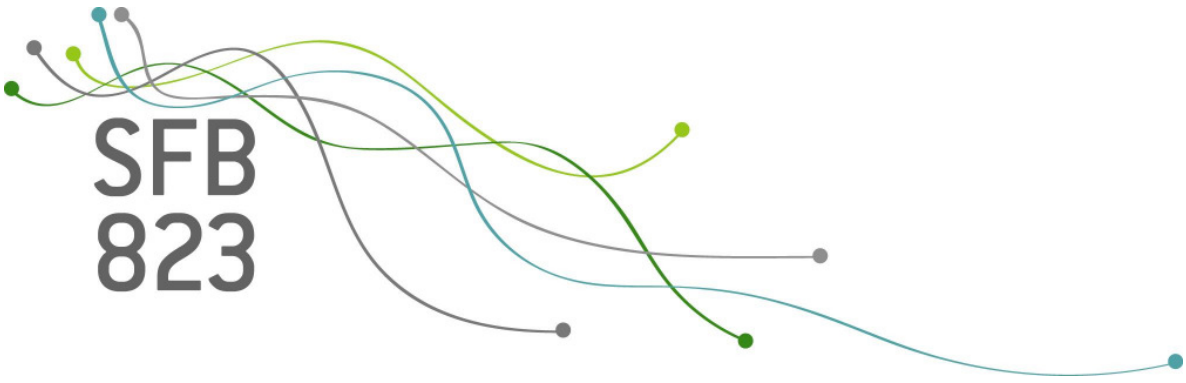
SFB 823

On robust cross-validation for nonparametric smoothing

Oliver Morell, Dennis Otto, Roland Fried

Nr. 17/2010

Discussion Paper



On robust cross-validation for nonparametric smoothing

Oliver Morell, Dennis Otto and Roland Fried

Department of Statistics
TU Dortmund University
44221 Dortmund, Germany

Abstract

Procedures for local-constant smoothing are investigated in a broad variety of data situations with outliers and jumps. Moving window and nearest neighbour versions of mean and median smoothers are considered, as well as double window and linear hybrid smoothers. For the choice of the window width or the number of neighbours the different estimators are combined with each of several cross-validation criteria like least squares, least absolute deviations, and median-cross-validation. It is identified, which method works best in which data scenarios. Although there is not a single overall best robust smoothing procedure, a robust cross-validation criterion, called least trimmed squares-cross-validation, gives good results for most smoothing methods and data situations, with cross-validation based on least absolute deviations being a strong competitor, particularly if there are jumps, but little problems with outliers in the data.

1 Introduction

We consider a regression model

$$Y_i = f(x_i) + E_i,\ i = 1, \ldots, n\ , \tag{1}$$

where f is an unknown piecewise continuous function, $x_1, \ldots, x_n$ are values of a covariate generated from a random design $X_1, \ldots, X_n$, $E_1, \ldots, E_n$ are i.i.d. errors possibly contaminated by some outliers, and $Y_1, \ldots, Y_n$ are observations of a response variable measured at $x_1, \ldots, x_n$, with realisation $y_1, \ldots, y_n$. For simplicity of the exposition we assume the data to be already ordered according to the size of the $x_i, x_1 \leq x_2 \leq \ldots \leq x_n$.

Local parametric fitting allows estimation of the unknown regression function f under weak assumptions on f, that is without the need of specifying a global functional form which is known except for some unknown parameters. We concentrate on local constant smoothing in the following. Several such

smoothers have been proposed, based on the idea to approximate f within suitably chosen local data windows by a constant.

The choice of the window width is crucial for the performance of any local fitting method. If it is chosen small, the variance of the estimate is large and the bias small. The estimation of f can become quite wiggly then. If the window width is chosen large, the variance of the estimate gets smaller, but the bias increases. Important details of f can be lost then. A data-based approach to select the window width adaptively is cross-validation (CV). Besides the commonly used L_2-CV, alternatives like L_1-CV (Yang and Zheng, 1992) and the robust median-CV (Yang and Zheng, 1998) have been proposed in the literature. We present a robust CV-criterion based on the idea of least trimmed squares (Rousseeuw, 1984), and compare it to the other criteria.

In the comparisons, we compare the CV-criteria with eight estimators based on local constant fits like moving averages, running medians, double window and linear hybrid smoothers in an extensive simulation study, considering data situations with jumps in the regression function f and outliers in the errors $E_1 \ldots, E_n$. As there is a lack of experimental comparisons, we look for recommendations, which smoother combined with which CV-criterion to use in different types of data situations.

Section 2 reviews local constant smoothing methods. Section 3 introduces different CV-criteria. Section 4 describes the data sets analyzed in the simulations and the results of the simulation study. Section 5 concludes.

2 Local constant smoothers

A multitude of methods is available for nonparametric fitting of an at least piecewise constant function f. A simple moving average smoother

$$\Phi_1(y_i) = \frac{1}{2k+1} \sum_{j=-k}^{k} y_{i-j}, \tag{2}$$

estimates f at each point x_i by the mean of the observations at the $2k+1$ design points $x_{i-k} \ldots, x_{i+k}$ centered at x_i. For the first and the last k design points, located at the boundary of the design space, we take the mean of the first and the last k observations, respectively, for the estimate. This kind of method we will call a moving window technique.

A modification of this procedure is the $(2k+1)$-nearest neighbour mean smoother. Let $x_i^{(j)} = \{|x_1 - x_i|, \ldots, |x_n - x_i|\}_{(j)}$ be the j-th nearest neighbour of x_i and $y_i^{(j)}$ the value observed at $x_i^{(j)}$, for $j = 1, \ldots, n$. Then the

$2k+1$-nearest neighbour mean is defined as

$$\Phi_2(y_i) = \frac{1}{2k+1} \sum_{j=1}^{2k+1} y_i^{(j)}. \tag{3}$$

Note that $\Phi_1(y_i)$ and $\Phi_2(y_i)$ are identical for $i = k+1, \ldots, n-k$ in case of an equidistant fixed design, but they are different in general, since the $2k+1$ nearest neighbours are not necessarily distributed as k points on the left and k points on the right hand side of x_i.

An advantage of the above smoothers based on averages is the high efficiency in case of normal errors. However, a single outlier affects the estimation and can make it completely meaningless locally. The robustness of an estimate against outliers can be measured by the finite sample breakdown point (Donoho and Huber, 1983). It corresponds to the minimal fraction of modifications in a sample of size $2k+1$ which can drive the estimate to the boundaries of the parameter space. In case of a sample of size $2k+1$ it is $1/(2k+1)$ for the sample mean, meaning that a single outlier can cause a spike of any size in the estimate of f. Moreover, mean smoothers smear jumps since the estimate averages observations before and after the location of the jump what clearly indicates its lack of robustness.

In an attempt to avoid these disadvantages of Φ_1 and Φ_2, we can use a robust measure of location instead of the non-robust average. Median-based smoothers improve upon both shortcomings since they are robust to outliers and preserve shifts much better. The moving window version is called running median smoother and is defined by

$$\Phi_3(y_i) = \mathrm{Med}(y_{i-k}, \ldots, y_{i+k}). \tag{4}$$

At points x_i close to the boundary the median of the first k and the last k observations, respectively, is taken as estimate of $f(x_i)$.

The $2k+1$ nearest neighbour median smoother is given by

$$\Phi_4(y_i) = \mathrm{Med}(y_i^{(1)}, \ldots, y_i^{(2k+1)}) \ . \tag{5}$$

Both variants offer a finite sample breakdown point of $(k+1)/(2k+1)$ within each window, which is optimal within the class of all location-equivariant estimators. Moreover, jumps between two constant parts of the function are preserved if there are at least $(k+1)$ observations for each part (Gather et al., 2006). Under Gaussian noise, the median offers an asymptotic efficiency relatively to the average of 63.7% in an increasingly large window.

Another approach from signal processing for local constant function fitting are linear median hybrid (LMH) filters (Heinonen, 1987). Linear subfilters $H_1, \ldots, H_m$ are calculated for each x_i and then the median of the outputs

of the m subfilters is taken for estimation of $f(x_i)$. As proposed by Heinonen, we use $m = 3$ and

$$\Phi_5(y_i) = \text{Med}(H_1(y_i), H_2(y_i), H_3(y_i)), \quad \text{with}$$
$$H_1(y_i) = \frac{1}{k}\sum_{j=1}^{k} y_{i-j},\ H_2(y_i) = y_i,\ H_3(y_i) = \frac{1}{k}\sum_{j=1}^{k} y_{i+j}, \tag{6}$$

where $H_1(y_i)$ and $H_3(y_i)$ take the average of the k observations left and right of the current design point x_i, whereas the filter output of $H_2(y_i)$ is y_i itself.

To increase the robustness of the procedure against outliers, we follow Fried et al. (2007) and use the median instead of the average in the subfilters to derive median median hybrid (MMH) filters, or better to say smoothers in our context,

$$\begin{aligned} \Phi_6(y_i) &= \text{Med}(M_1(y_i), M_2(y_i), M_3(y_i)), \quad \text{with} \quad M_2(y_i) = y_i, \\ M_1(y_i) &= \text{Med}(y_{i-k}, \ldots, y_{i-1}),\ M_3(y_i) = \text{Med}(y_{i+1}, \ldots, y_{i+k}). \end{aligned} \tag{7}$$

Including the central observation as a subfilter improves the preservation of level shifts in both cases. For design points close to the boundary the estimated values $\Phi_\ell(y_{k+1})$ and $\Phi_\ell(y_{n-k})$, $\ell = 5, 6$, are used for estimation of $f(x_i)$ at the first and the last k design points, respectively.

Another method from signal processing is the double window modified trimmed mean (DWMTM) (Lee and Kassam, 1985). Defining a trimming factor $\varpi \in (0, 0.5)$, a ϖ-trimmed mean is an average value of the observations, with the $100\varpi\%$ smallest and the $100\varpi\%$ largest values being disregarded. By taking ϖ about 20% a compromise between the sample median ($\varpi = 0.5$) and the sample mean ($\varpi = 0$) is obtained, what allows efficient estimations for a wide class of distributions with tails heavier than the Gaussian. However, trimmed means with $\varpi < 0.5$ do not preserve jumps exactly. Therefore a procedure with an adaptive, data-based choice of the trimming factor, like the DWMTM, is preferable. The DWMTM uses two windows of width k and l, respectively, with $l \leq k$. The median $\tilde{y}_i$ and the median absolute deviation from the median (MAD) $\widehat{s}_M$ as a robust measure of location and variability, respectively, are calculated from the possibly smaller inner window $y_{i-l}, \ldots, y_{i+l}$. Then all observations $z \in \{y_{i-k}, \ldots, y_{i+k}\}$ with

$$|z - \tilde{y}_i| > \delta \widehat{s}_M(y_{i-l}, \ldots, y_{i+l}) \tag{8}$$

are trimmed and the remaining values are averaged. Here, δ is a predefined constant regulating the amount of trimming, e.g. $\delta = 2$ (Lee and Kassam, 1985). Advantages of the DWMTM over the ordinary ϖ-trimmed means are the better noise suppression and the better preservation of fine details

of f. Noise can be suppressed more efficiently because we can use a larger outer window width of $2k+1$ data points for the smoothing and since the percentage of trimming is chosen adaptively. Relevant details of f can be preserved better because the median and MAD are calculated from a smaller window of $2l+1$ points. At the k smallest and largest design points, we again take the estimates of $f(x_{k+1})$ and $f(x_{n-k})$, respectively.

The robust version of locally weighted regression, or briefly Lowess (Cleveland, 1979) is also taken into account for comparison, since it is a commonly applied standard. Let W be the tricube function and B be the bisquare function, with

$$W(x) = \begin{cases} (1-|x|^3)^3 & , \text{ if } |x|<1 \\ 0 & , \text{ if } |x| \geq 1 \end{cases}, \tag{9}$$

$$B(x) = \begin{cases} (1-x^2)^2 & , \text{ if } |x|<1 \\ 0 & , \text{ if } |x| \geq 1 \end{cases}. \tag{10}$$

Let d_{ik} be the absolute distance between x_i and $x_i^{(2k+1)}$, which is again the $(2k+1)$-th nearest neighbour of x_i. By calculating $w_j(x_i) = W(\frac{x_j - x_i}{d_{ik}}), j = 1, \ldots, n$, weights for each design point are included in the estimation of $f(x_i)$. By construction these weights are only larger than zero for the $2k+1$ nearest neighbours. The residuals $\widehat{e_1}, \ldots, \widehat{e_n}$ of an initial locally weighted polynomial regression are calculated and used to derive new robustness weights $\delta_1, \ldots, \delta_n$, with

$$\delta_j = B\left(\frac{\widehat{e_j}}{6\widehat{e}_{Med}}\right), \tag{11}$$

where $\widehat{e}_{Med}$ is the median of the absolute residuals. Every residual which is too large compared to the median of the absolute residuals gets the weight $\delta_j = 0$. Then a new locally weighted regression is calculated with each data point (x_j, y_j) getting a new robust localized weight $\delta_j \; w_j(x_i)$: observations with large residuals in the step before get a small weight or are even trimmed completely. This step can be repeated, calculating another robust local polynomial regression with weights based on the previous step. Cleveland states that two iterations already lead to an adequate fit for many data sets. We use the R-function lowess (R Development Core Team, 2009) for computing these estimates.

3 Cross-validation

The performance of each of the procedures described in Section 2 depends on its smoothing parameter k, which delivers the window width $2k+1$. In

case of the DWMTM we have two smoothing parameters k and l, but we fix $l = \lfloor (k-1)/2 \rfloor$ here for the reason of simplicity and choose $\delta = 2$, as proposed by Lee and Kassam (1985).

A way to choose the smoothing parameter adaptively, i.e. based on the data, is cross-validation (CV). If $\widehat{f}(x_i)$ is the estimate of f at the point x_i derived by one of the smoothing methods from Section 2, then let $\widehat{f}_{-i}(x_i)$ be the estimate of f at the point x_i, when (x_i, y_i) itself is not used for the estimation. For the moving window techniques we base the estimation $\widehat{f}_{-i}(x_i)$ on the $2k+1$ points in $\{x_{i-1-k}, \ldots, x_{i-1}, x_{i+1}, \ldots, x_{i+k}\}$.

A common criterion for the choice of k is the traditional least squares CV, or briefly L_2-CV. It averages the squared distances between the observations y_i and the estimates $\widehat{f}_{-i}(x_i)$ and chooses k as

$$\arg\min_k \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \widehat{f}_{-i}(x_i) \right)^2 . \tag{12}$$

Outliers close to x_i affect the estimate $\widehat{f}_{-i}(x_i)$ and increase its distance to the observed value y_i. A somewhat robust alternative is to use absolute instead of squared distances. The least absolute deviations CV, or briefly L_1-CV-criterion (Yang and Zheng, 1992) for the choice of k is

$$\arg\min_k \frac{1}{n} \sum_{i=1}^{n} \mid y_i - \widehat{f}_{-i}(x_i) \mid . \tag{13}$$

Although absolute distances are less affected by outliers than squared ones, outliers still have an impact on the estimation of $\widehat{f}_{-i}(x_i)$ and thus on the choice of k. Zheng and Yang (1998) introduced median-CV

$$\arg\min_k \text{ Med} \left(\mid y_1 - \widehat{f}_{-1}(x_1) \mid, \ldots, \mid y_n - \widehat{f}_{-n}(x_n) \mid \right), \tag{14}$$

as a more robust alternative to L_1- and L_2-CV. A possible disadvantage of median-CV is that a lot of information gets lost in the determination of k since only the median residual is used. Therefore we use least trimmed squares-CV (LTS-CV) as another robust criterion. Let $r_i^2 = (y_i - \widehat{f}_{-i}(x_i))^2, i = 1, \ldots, n$ be the squared distance between the estimated $\widehat{f}_{-i}(x_i)$ and the observed value y_i, and $r_{(j)}^2, j = 1, \ldots, n$ its order statistics. Then LTS-CV is defined by

$$\arg\min_k \frac{1}{\lfloor hn \rfloor} \sum_{j=1}^{\lfloor hn \rfloor} r_{(j)}^2, \tag{15}$$

where $(1-h) \in (0, 0.5)$ is a trimming factor. All squared residuals r_i^2, are calculated and sorted, before the $n - \lfloor hn \rfloor$ largest squared residuals are trimmed and the others are averaged. We consider LTS-CV with $h \in \{0.5, 0.75\}$ and call it 50%-LTS-CV and 75%-LTS-CV, respectively. In the following we call median- and LTS-CV robust CV-methods, because they use a robust distance measure.

4 Comparative Study

We combine each of the eight smoothing methods from Section 2 with each of the five cross-validation criteria from Section 3. The resulting 40 smoothing procedures (also briefly called estimators in the following) are compared in a simulation study. The unknown regression function f is chosen to be piecewise constant since we assume the effects of jumps and outliers on the estimates to be more severe than a slight slope.

To generate a broad variety of different data situations we vary five factors for the unknown function f, the design and the noise of the model in (1). For each factor we consider three settings and generate $\nu = 1000$ data sets from each of the arising $\eta = 3^5 = 243$ combinations. The five factors and their three settings are:

(a) sample size: $n \in \{40, 100, 200\}$;

(b) percentage of outliers: $\pi \in \{0.01, 0.05, 0.15\}$

(c) absolute magnitude of the outliers: $\gamma \in \{3\sigma, 6\sigma, 12\sigma\}$

(d) number of level shifts: $m \in \{1, 2, 5\}$

(e) absolute magnitude of level shifts: $s \in \{1\sigma, 3\sigma, 6\sigma\}$

To generate n observations from model (1), we draw n values $X_1, \ldots, X_n$ from an uniform random design and error variables $E_1, \ldots, E_n$ which are i.i.d. $\mathcal{N}(\mu, \sigma^2)$ with $\sigma^2 = 1$. As usually, uniform design means that $X_1, \ldots, X_n$ are i.i.d. uniformly distributed on the interval $[0, 1]$.

Then $\max\{\lfloor n\pi \rfloor, 1\}$ of the noise values are modified to become outliers. The positions of the outliers are chosen at random by drawing $\max\{\lfloor n\pi \rfloor, 1\}$ of the n positions without replacement. At each position of an outlier, the value $\pm\gamma$ is added to the noise, with the same sign as the closest level shift, to produce a more challenging situation for the estimation procedures.

The positions $\xi_1, \ldots, \xi_m$ of the m jumps within the interval (0,1) are fixed for each m (for $m = 1$: $\xi_1 = 0.4$, for $m = 2$: $\xi_1 = 0.4$ and $\xi_2 = 0.6$, for $m = 5$:

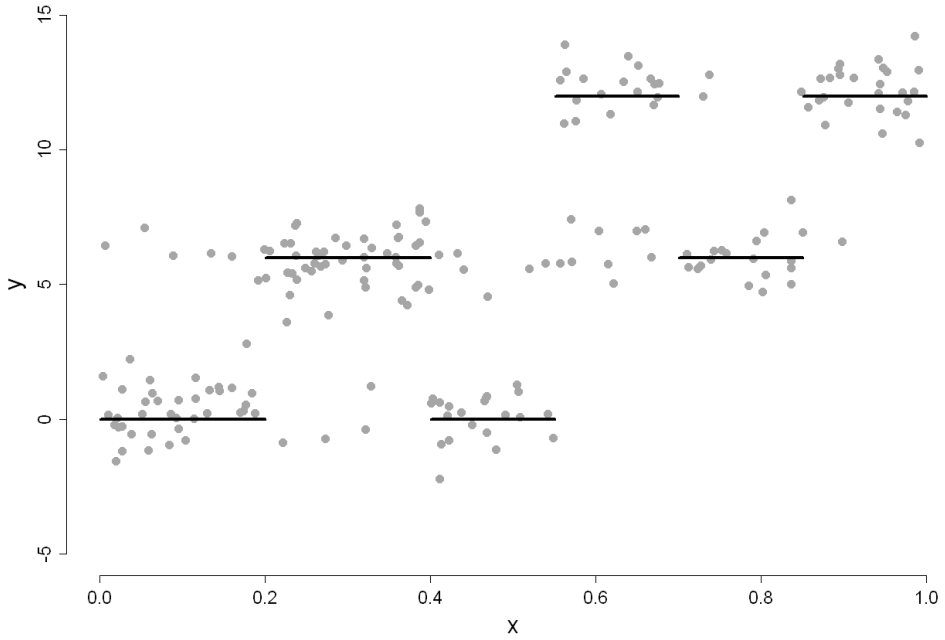


Figure 1: Data example with $n = 200$ observations, $\pi = 15\%$ outliers of size $\gamma = 6$ and $m = 5$ level shifts of height $s = 6$.

$\xi_1 = 0.2$, $\xi_2 = 0.4$, $\xi_3 = 0.55$, $\xi_4 = 0.7$ and $\xi_5 = 0.85$). In Figure 1 an example with $n = 200$, $\pi = 0.15$, $m = 5$ and $\gamma = s = 6\sigma$ is shown.

For every data set j, $j = 1 \ldots, \nu$, generated for each of the total $\eta = 243$ data situations, and each of 40 estimators we choose k by the corresponding CV method. Using this k we calculate the resulting estimates $\hat{f}_e(x_i^j)$ for positions $i = 1, \ldots, n$, data set j and estimator $e = 1, \ldots, 40$. The performance of the procedures for the j-th data set can be compared by the Averaged Squared Error (ASE; see Härdle, 2002, pp. 90)

$$d_j(\hat{f}_e, f) = \frac{1}{n} \sum_{i=1}^{n} \left(\hat{f}_e(x_i^j) - f(x_i^j) \right)^2 , \tag{16}$$

where x_i^j is the i-th design point for the j-th data set, respectively. This criterion measures the averaged squared distance between the true function f and the estimated function values for data set j. Finally, the different estimation procedures are compared for each data situation p, $p = 1, \ldots, \eta$ by the mean ASE-value (denoted by MASE), averaged across the $\nu = 1000$

data sets generated for this situation,

$$\bar{d}(\hat{f}_e, f) = \frac{1}{\nu}\sum_{j=1}^{\nu} \hat{d}_j(\hat{f}_e, f) = \frac{1}{\nu n}\sum_{j=1}^{\nu}\sum_{i=1}^{n}\left(\hat{f}_e(x_i^j) - f(x_i^j)\right)^2, \tag{17}$$

In order to simplify the evaluation of the 40 different estimation procedures for the 243 data situations, we define a summary measure to compare the relative performance of the methods. For each data situation, we consider the loss in the MASE-value due to not using the best estimator for this data situation, relatively to the minimal MASE-value for that situation. This gives us

$$\zeta_p^e = \frac{\bar{d}_p(\hat{f}_e, f) - \bar{d}_p^\star(\hat{f}, f)}{\bar{d}_p^\star(\hat{f}, f)}, \tag{18}$$

with $\bar{d}_p(\hat{f}_e, f)$ being the MASE-value of the estimator e for the p-th data situation as defined in (17) and

$$\bar{d}_p^\star(\hat{f}, f) = \min\left(\bar{d}_p(\hat{f}_1, f), \ldots, \bar{d}_p(\hat{f}_{40}, f)\right) \tag{19}$$

being the MASE-value of the estimator which performs best for the p-th data situation. Small values of ζ_p^e close to zero indicate that estimator e performs almost as good as the best estimator for data situation p.

In the following we will consider interesting subsets of data situations. Given such a subset with $\tilde{\eta}$ data situations, $p = 1 \ldots, \tilde{\eta}$, the average relative performance

$$\bar{\zeta}^e = \frac{1}{\tilde{\eta}}\sum_{p=1}^{\tilde{\eta}} \zeta_p^e. \tag{20}$$

is used as global performance measure in the comparison of the different estimation procedures.

4.1 Results for a single jump and a moderate number of outliers

In a first step we compare the estimators in data situations with problems caused mainly by outliers. Setting all other factors to their easiest values, we consider a uniform design with $n = 200$ data points and one jump with height 1σ. We calculate $\bar{\zeta}^e$ from equation (20) for the $\tilde{\eta} = 9$ data situations with different percentages ($\pi \in \{0.01, 0.05, 0.15\}$) and magnitudes

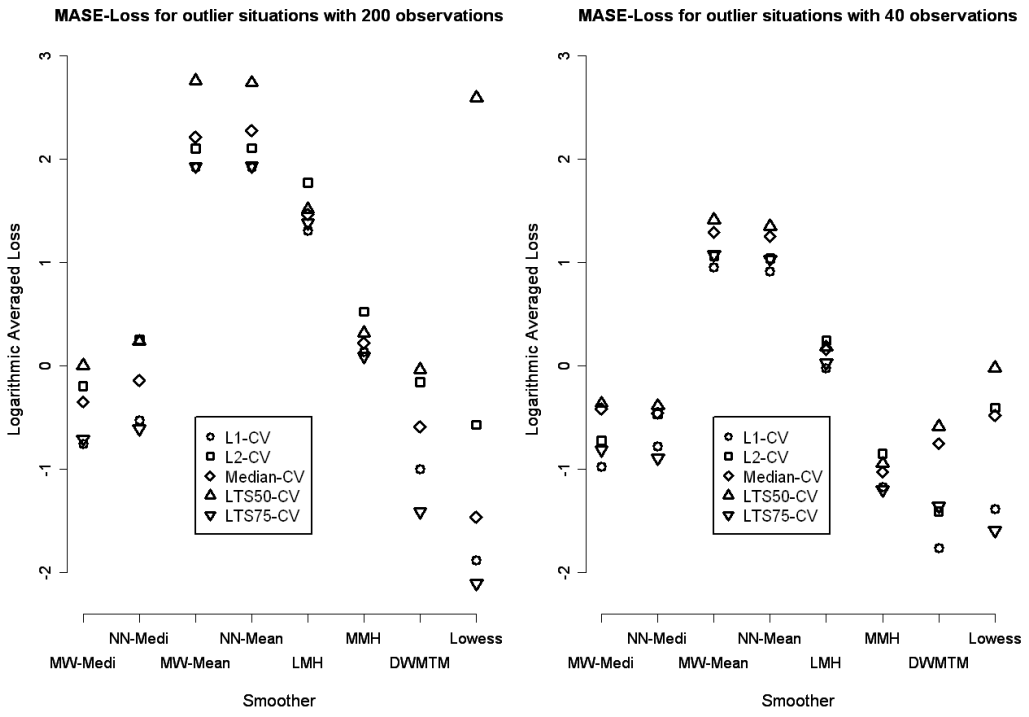


Figure 2: Average loss in MASE for different outlier situations and a sample size $n = 200$ (left) and $n = 40$ (right).

($\gamma \in \{3\sigma, 6\sigma, 12\sigma\}$) of outliers. The results are illustrated in Figure 2. We use a logarithmic scale for the ordinate since we want to visualise differences among good estimators. Lowess performs best, followed by DWMTM. The median smoothers and MMH also give acceptable results, while the mean smoothers and LMH are much worse. 75%-LTS- and L_1-CV give the best results for all methods.

Calculating $\bar{\zeta}^e$ for the same outlier situations, but with a smaller sample size of $n = 40$, DWMTM and Lowess with 75%-LTS- and L_1-CV again give the best results. For this sample size the MMH seems to be a better choice than the median smoothers. 75%-LTS-CV again leads to better results than Median-CV for all estimators, but for some of the smoothers it is outperformed by L_1-CV. For $n = 100$, the results are in between those for $n = 40$ and $n = 200$.

Summarizing, in data situations with one small jump and at most 15% outliers, Lowess and DWMTM with L_1- or 75%-LTS-CV are to be recommended, for all sample sizes considered here. For all smoothers, L_1- or 75%-LTS-CV seem to be better choices than their three competitors.

4.2 Results for a few outliers and several jumps

The same kind of comparison can be made for different situations with level shifts, fixing the percentage outliers at $\pi = 0.01$ and their magnitude at $\gamma = 3\sigma$. We calculate $\bar{\zeta}^e$ from the $\tilde{\eta} = 9$ factor combinations with different numbers ($m \in \{1, 2, 5\}$) and magnitudes ($s \in \{1\sigma, 3\sigma, 6\sigma\}$) of jumps, see Figure 3. For $n = 200$, generally, all smoothers show their best perfor-

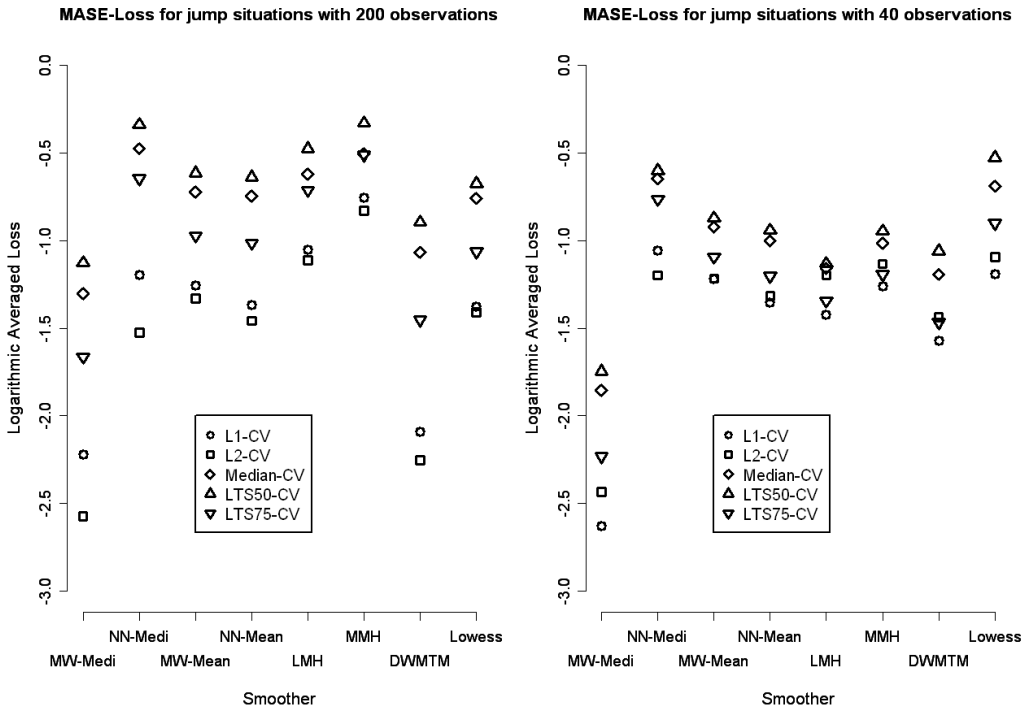


Figure 3: Average loss in MASE for different jump situations and a sample size $n = 200$ (left) and $n = 40$ (right).

mance here, if L_2-CV is used, followed by L_1-CV and 75%-LTS-CV, which again delivers better results than median-CV for all smoothers. The running median performs best, followed by DWMTM. The nearest neighbour median performs worse than the moving window median, irrespective of the CV-method used. This is due to the jump preserving property of the median, which needs the number of observations left and right of the jump used for the estimation to be equal. While for the moving window median this property is fulfilled, for the nearest neighbour version this is not necessarily the case in a random design. Lowess and the mean smoothers do not really perform well here, irrespective of the CV-criterion used.

For the smaller sample size $n = 40$ and the same jump situations the running median with L_1-CV performs best, but it delivers good results also with any of the other four CV-methods. For most of the smoothers L_1-CV is now better than L_2-CV. 75%-LTS-CV delivers better results than L_2-CV for LMH, MMH and DWMTM, and better results than median-CV for all smoothers. For $n = 100$, the results are again in between those for $n = 40$ and $n = 200$ and the running median with L_1- or L_2-CV performs best.

In data situations without large outliers, but jumps in f, the running median with L_1- or L_2-CV seems preferable to the other smoothers and L_1-CV and L_2-CV give better results than their robust competitors for most of the estimators. 75%-LTS-CV gives again better results than Median-CV.

4.3 Overall analysis

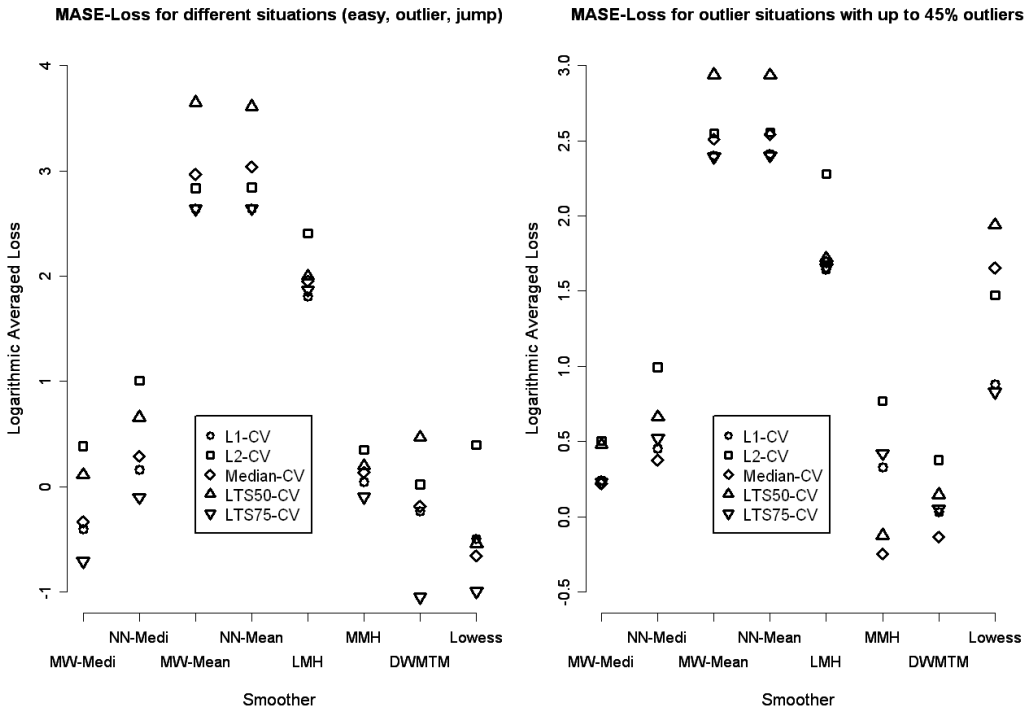


Figure 4: Loss in MASE averaged over very easy and very difficult situations (left) and averaged over all outlier situations up to 45% (right).

Since the results for outliers and level shifts differ, one is interested, which method performs best in general. We take the most difficult situations each with outliers (largest percentage rate and largest magnitude) and

jumps (largest number and largest height) from the first and the second comparison. As a third scenario we consider the data situation, where each factor stands on its easiest level, i.e. a large sample with $n = 200$ from an uniform design with one small jump and a small percentage outliers with a small magnitude.

W.r.t. $\bar{\zeta}^e$, the results from Figure 4 (left) are: The worst estimators are again all versions of mean smoothers, followed by the LMH. DWMTM, Lowess and the running median (in this ordering), all with 75%-LTS-CV, give the best overall performance. Again, for all CV-methods the running median performs better than the nearest neighbour median.

We also considered a discrete design as an alternative to the uniform one. This discrete design was generated by drawing n points from the discrete support $\{0, 0.01, \ldots, 0.99, 1\}$ with replacement. We did not find important differences to the results for the uniform design. All statements from the previous sections are confirmed for this alternative design type.

4.4 Situations with a large number of outliers

To examine the performance of the procedures in case of an increasing percentage of outliers more closely, we do some additional simulations. Only the case of $n = 200$ data points from a uniform design and one jump with height 1σ is considered. 1000 data sets are generated for each combination of percentage $\pi \in \{0, 0.05, \ldots, 0.4, 0.45\}$ and magnitude $\gamma \in \{3\sigma, 6\sigma, 12\sigma\}$ of outliers. If we include all thirty arising data situations in the calculation of $\bar{\zeta}^e$ from the realised MASE-values, the best values of $\bar{\zeta}^e$ correspond to estimators which deliver good estimations irrespective of the percentage and the magnitude of the outliers, see Figure 4 on the right. W.r.t. $\bar{\zeta}^e$, the MMH with median-CV performs best, followed by the DWMTM with median-CV and the MMH with 50%-LTS-CV. The DWMTM smoothers with any other CV-criterion except L_2-CV come next. The running median with median-, 75%-LTS- and L_1-CV also performs well. The mean smoothers, the LMH and Lowess do not perform well for any CV-criterion.

For a more detailed look at these results we divide the situations into three groups with different percentages of outliers $\pi \in \{0, 0.05, 0.1\}$, $\pi \in \{0.15, 0.2, 0.25\}$ and $\pi \in \{0.3, 0.35, 0.4, 0.45\}$ for each outlier magnitude $\gamma \in \{3\sigma, 6\sigma, 12\sigma\}$. The reason is that there is no CV-method, which performs well for both small and large percentages of outliers. We identify in each group for each smoother the CV-criterion which delivers the smallest value of $\bar{\zeta}^e$ and compare all smoothers for each magnitude γ, by plotting the logarithmic MASE-values for each percentage of outliers for $\gamma = 3\sigma$ and $\gamma = 12\sigma$ in Figure 5. The three informations in brackets give the CV-criteria, which

deliver the smallest relative loss in MASE for the three groups above for each smoother.

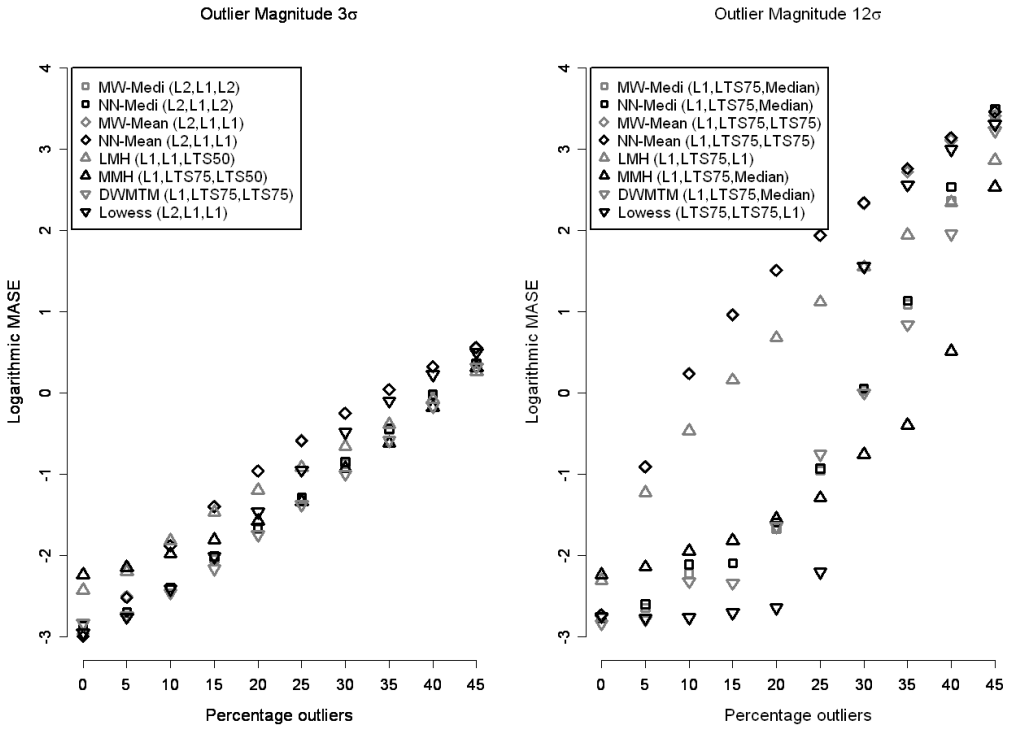


Figure 5: MASE-values for an increasing percentage of outliers.

For data sets with small outliers of size $\gamma = 3\sigma$, DWMTM with L_1-CV is to be recommended if the percentage π is small. For a larger percentage π DWMTM with 75%-LTS-CV should be used. The differences to other smoothers with their case by case best CV do not seem to be large here, but DWMTM is the only smoother which is among the best for all π. 75%-LTS-CV delivers better results than Median-CV for all smoothers and all π, but is mostly outperformed by L_1- or L_2-CV.

Considering the large magnitude $\gamma = 12\sigma$ Lowess with 75%-LTS-CV performs best if $\pi \leq 0.25$. For $\pi > 0.25$, Lowess becomes one of the worst smoothers and the MMH-smoother with Median-CV is to be recommended. 75%-LTS-CV loses its good behaviour for $\pi > 0.25$ and a large γ. Median-CV and 50%-LTS-CV deliver the best results for robust smoothers then. We only show the cases $\gamma \in \{3\sigma, 12\sigma\}$, as the case $\gamma = 6\sigma$ looks similar to $\gamma = 12\sigma$. The only difference is the performance of DWMTM, which outperforms the other smoothers for $\pi \in \{0.20, 0.25, 0.30, 0.35\}$ then.

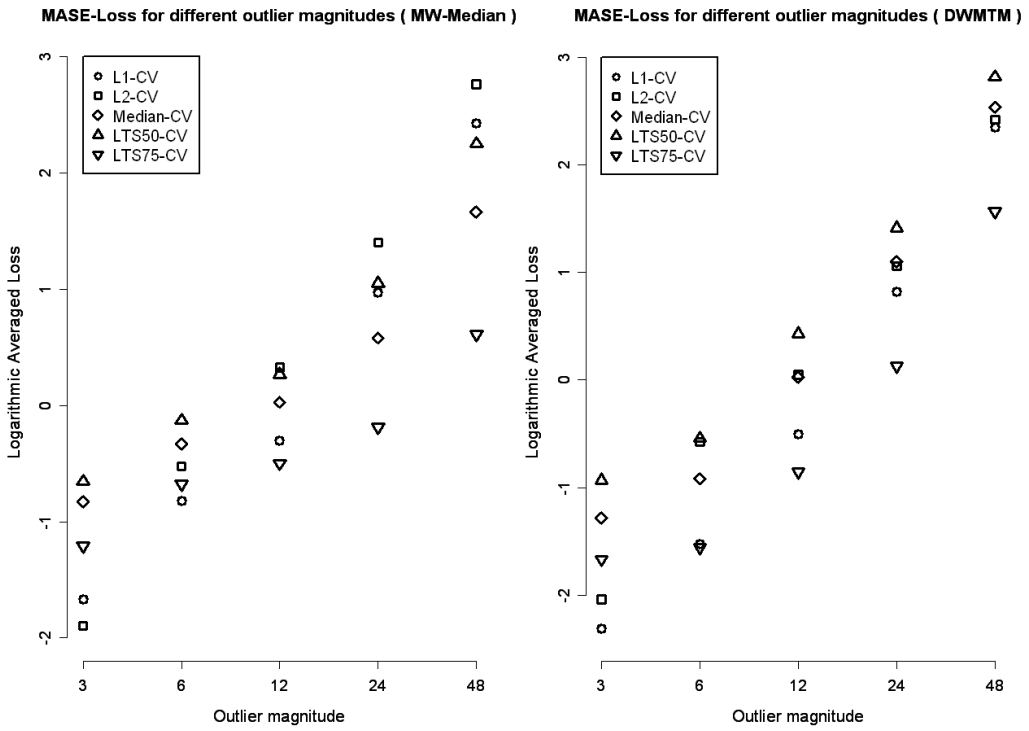


Figure 6: MASE-values for an increasing magnitude of outliers, averaged over different sample sizes and percentages of outliers.

We also consider the effect of huge outliers. We take again only one small jump and average over the different sample sizes $n \in \{40, 100, 200\}$ and percentages of outliers $\pi \in \{0.01, 0.05, 0.15\}$. Figure 6 illustrates the results for the running median (left) and the DWMTM (right). It shows that the advantage of 75% LTS-CV increases for very large outlier sizes $\gamma \in \{24\sigma, 48\sigma\}$ over all four competitors. The results can be transferred to the nearest neighbour Median, the MMH and Lowess, while the mean smoothers and the LMH are not robust and so even one large outlier affects the estimation irrespectively of the CV used.

5 Conclusions

Jumps and outliers are challenges for smoothers. From the methods considered here, the running median with L_2- or L_1-cross-validation (CV) is to be recommended, if situations with jumps are considered and outliers are not really relevant.

In situations where outliers instead of jumps play a dominant role, we get the following findings: For small outliers, DWMTM with L_1- or 75%-LTS-CV gives the best results. For large outliers with a magnitude of 12σ Lowess with 75%-LTS CV is best, if at most 25% outliers occur, but it loses its goodness for higher percentages of outliers. The MMH with Median-CV is to be recommended in case of more than 25% large outliers.

Since the percentage and the size of of outliers is often unknown in practice, a tentative recommendation is to use DWMTM in combination with L_1-CV or 75%-LTS-CV as a generally reliable default smoothing method in the presence of outliers and jumps. Given the good performance of L_2-CV in the presence of only a few outliers, of 75% LTS-CV in the presence of a moderate percentage of outliers and of 50% LTS-CV in the presence of a large percentage of outliers, it seems worthwhile to investigate LTS-criteria with an adaptive percentage of trimming to be constructed along similar lines as the adaptive LTS estimator of Hofmann et al. (2010).

An explanation of the good performance of the LTS-CV is that it fits the squared distance measure used for the evaluation of the methods. If we use absolute distances in form of an Averaged Absolute Error (AAE) instead of the ASE, 75%-LTS-CV looses its superiority in some cases. For the median smoothers and Lowess it is outperformed by the L_1-CV in the two outlier situations considered in Figure 2. But it is still second best and so a better robust criterion than Median-CV. It can be speculated whether a least trimmed absolute deviation criterion would perform even better then.

Acknowledgements. This work has been supported in part by the Collaborative Research Center "Statistical modeling of nonlinear dynamic processes" (SFB 823) of the German Research Foundation (DFG).

References

[1] Cleveland, W.S., 1979. Robust locally weighted regression and smoothing scatterplots. Journal of the American Statistical Association 74, 829–836.

[2] Donoho, D.L., Huber, P.J., 1983. The notion of breakdown point, in: Bickel, P.J., Doksum, K., Hodges, J.L. (Eds.), A Festschrift for Erich Lehmann. Wadsworth, Belmont, CA, pp. 157–184.

[3] Fried, R., Bernholt, T., Gather, U., 2007. Repeated median and hybrid filters. Computational Statistics and Data Analysis 50, 2313–2338.

[4] Gather, U., Fried, R., Lanius, V., 2006. Robust detail-preserving signal extraction, in: Schelter, B., Winterhalder, M., Timmer, J. (Eds.), Handbook of Time Series Analysis. Wiley, Weinheim, pp. 131–157.

[5] Härdle, W. (2002): Applied nonparametric regression. Cambridge University Press, Edinburgh.

[6] Heinonen, P, Neuvo, Y., 1987. FIR-median hybrid filters. IEEE Transactions of Acoustics, Speech and Signal Processing 35, 832–838.

[7] Hofmann, M., Gatu, C., Kontoghiorghes, E.J., 2010. An exact least trimmed squares algorithm for a range of coverage values. Journal of Computational and Graphical Statistics 19, 191-204.

[8] Lee, Y.H., Kassam, S.A., 1985. Generalized median filters and related nonlinear filtering techniques. IEEE Transactions of Acoustics, Speech and Signal Processing 33, 672–683.

[9] R Development Core Team, 2009. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

[10] Rousseeuw, P.J., 1984. Least median of squares regression. Journal of the American Statistical Association 79, 871–880.

[11] Yang, Y., Zheng, Z., 1992. Asymptotic properties for cross-validated nearest neighbor median estimators in nonparametric regression: the L_1-view, in: Jiang, Z., Yan, S., Cheng, P., Wu, R. (Eds.), Probability and Statistics. World Scientific, Singapore, pp. 242–257.

[12] Zheng, Z., Yang, Y., 1998. Cross-validation and median criterion. Statistica Sinica 8, 907–921.