Robust Fitting of Discontinuous Regression Functions*

Thomas C. M. Lee†
The University of Chicago, Department of Statistics
5734, S. University Avenue, Chicago, IL 60637, USA
Email: tlee@galton.uchicago.edu

Abstract

This paper proposes an automatic smoothing procedure that is designed for recovering discontinuous regression functions. The procedure models the target regression function with a series of disconnected cubic regression splines which partition the function's domain. In this way discontinuity points can be incorporated in a fitted curve simply as the boundary points between adjacent splines. The minimum description length principle is applied to choose the number and placement of these discontinuity points as well as the amount of smoothing. The procedure is further made robust by using a Gaussian mixture approach to model outliers, and a genetic algorithm is developed for numerically computing a "best" regression estimate. Simulation results show that the proposed method is superior to many existing smoothing methods when the target function is non–smooth.

1 Curve Model

Suppose n pairs of noisy measurements $\{x_i, y_i\}_{i=1}^n$ are observed that satisfy the following assumptions:

$$y_i = f(x_i) + e_i, \quad x_1 < \ldots < x_n, \quad e_i \sim \text{iid } N(0, \sigma^2).$$

The aim is to estimate f. It is anticipated that f may possess a few discontinuity points but otherwise is smooth. In Section 4 the assumption of Gaussian e_i's will be relaxed for allowing the presence of outliers.

Suppose it is known that there are $B-1$ discontinuity points in f and that these discontinuity points are located at $b_1, b_2, \ldots, b_{B-1}$. For convenience let $b_0 = x_1$ and $b_B = x_n$, and assume $b_0 < b_1 < \ldots < b_B$. Then one way to estimate f is to fit a separate cubic regression spline to each of the B disjoint segments $[b_{j-1}, b_j)$, $j = 1, \ldots, B$. If I_E is the indicator function for the event E, then such a disconnected regression spline model for f can be expressed as

$$f(x) = f_1(x) I_{\{b_0 \le x < b_1\}} + f_2(x) I_{\{b_1 \le x < b_2\}} + \ldots + f_B(x) I_{\{b_{B-1} \le x \le b_B\}}, \qquad (1)$$

where each of the f_j's is a cubic regression spline having m_j knots located at $k_{j1}, k_{j2}, \ldots, k_{jm_j}$. Furthermore,

$$f_j(x) = \alpha_{j0} + \alpha_{j1} x + \alpha_{j2} x^2 + \alpha_{j3} x^3 + \sum_{r=1}^{m_j} \beta_{jr} (x - k_{jr})_+^3, \qquad (2)$$

$b_{j-1} \le x < b_j$, $j = 1, \ldots, B$, where $\boldsymbol{\alpha}_j = \{\alpha_{j0}, \ldots, \alpha_{j3}\}$ and $\boldsymbol{\beta}_j = \{\beta_{j1}, \ldots, \beta_{jm_j}\}$ are model parameters and $j = 1, \ldots, B$, and the function $(a)_+ = \max(0, a)$. To simplify notation, let $\boldsymbol{b} = \{b_1, \ldots, b_{B-1}\}$, $\boldsymbol{m} = \{m_1, \ldots, m_B\}$ and $\boldsymbol{k}_j = \{k_{j1}, \ldots, k_{jm_j}\}$ for $j = 1, \ldots, B$. Although not necessary but for computational convenience, it is assumed that $\{b_j, k_{jr}; \text{for all } j, r\}$ is a subset of $\{x_1, \ldots, x_n\}$ and that

$$b_0 < k_{11} < \ldots < k_{1m_1} < b_1 < \ldots < b_{j-1} < k_{j1} < \ldots < k_{jm_j} < b_j < \ldots < b_B.$$

Of course, in most situations the number and locations of the discontinuity points and knots are not known and need to be estimated.

If f is modelled by (1) and (2), then an estimate $\hat{f}$ of f can be obtained by first estimating $\boldsymbol{\theta} = \{B, \boldsymbol{b}, \boldsymbol{m}, \{\boldsymbol{k}_j, \boldsymbol{\alpha}_j, \boldsymbol{\beta}_j\}_{j=1}^B\}$ and then plugging–in the resulting estimate $\hat{\boldsymbol{\theta}} = \left\{\hat{B}, \hat{\boldsymbol{b}}, \hat{\boldsymbol{m}}, \{\hat{\boldsymbol{k}}_j, \hat{\boldsymbol{\alpha}}_j, \hat{\boldsymbol{\beta}}_j\}_{j=1}^{\hat{B}}\right\}$ into (1) and (2). Hence using the disconnected regression spline approach, our original curve estimation problem can be posed as a model selection problem, with each candidate model specified by a $\hat{\boldsymbol{\theta}}$. The goal, then, is to select a "best" $\hat{\boldsymbol{\theta}}$. Notice that different $\hat{\boldsymbol{\theta}}$'s may have different dimensions, but once $\hat{B}$, $\hat{\boldsymbol{b}}$, $\hat{\boldsymbol{m}}$ and $\{\hat{\boldsymbol{k}}_j\}_{j=1}^{\hat{B}}$ are specified, natural maximum likelihood estimates of $\{\hat{\boldsymbol{\alpha}}_j, \hat{\boldsymbol{\beta}}_j\}_{j=1}^{\hat{B}}$ can be computed by least-squares regression.

*To appear in the *Proceedings of the Interface 31*.

†Present address: Department of Statistics, Colorado State University, Fort Collins, CO 80523-1877, USA. Email: tlee@stat.colostate.edu.

2 Model Selection Using MDL

Here we discuss the use of the minimum description length (MDL) principle to select a "best" $\hat{\boldsymbol{\theta}}$.

The MDL principle *defines* the best fitted model as the one that produces the shortest code length of the data; see Rissanen (1989) and references given therein. In this context the code length of an object is treated as the amount of memory space that is required to store the object.

When applying the MDL principle, it is common to split the code length for a set of data into two parts: (i) a fitted model plus (ii) the data "conditioned on" the fitted model; i.e., the residuals. For the present case, the data are $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ and a fitted model can be simply specified by $\hat{\boldsymbol{\theta}}$. We write the residual vector as $\hat{\boldsymbol{e}} = (\hat{e}_1, \ldots, \hat{e}_n)^T$, where $\hat{e}_i = y_i - \hat{f}(x_i)$ for $i = 1, \ldots, n$. Thus, in other words, one splits $\boldsymbol{y}$ into $\hat{\boldsymbol{\theta}}$ plus $\hat{\boldsymbol{e}}$.

If $L(z)$ denotes the code length of the object z, we have $L(\boldsymbol{y}) = L(\hat{\boldsymbol{\theta}}) + L(\hat{\boldsymbol{e}}|\hat{\boldsymbol{\theta}})$. Note that in this expression it is stressed that $\hat{\boldsymbol{e}}$ is conditional on $\hat{\boldsymbol{\theta}}$. Now the task is to find an expression for $L(\boldsymbol{y})$ so that the best MDL $\hat{\boldsymbol{\theta}}$ can be defined and obtained as its minimizer. It is shown in Lee (1999) that $L(\boldsymbol{y})$ can be well approximated by

$$\begin{aligned} \mathrm{MDL}(\hat{f}) &= L(\boldsymbol{y}) \\ &= \log \hat{B} + \sum_{j=1}^{\hat{B}} \log \hat{m}_j + \\ &\quad \sum_{j=1}^{\hat{B}} \left(3 + \frac{\hat{m}_j}{2}\right) \log \hat{l}_j + \\ &\quad \frac{n}{2} \log \left[\frac{1}{n} \sum_{i=1}^{n} \left\{ y_i - \hat{f}(x_i) \right\}^2 \right], \qquad (3) \end{aligned}$$

where $\hat{l}_j$ is the number of x_i's in the jth fitted disconnected spline $\hat{f}_j(x)$. We propose to select the minimizer of $\mathrm{MDL}(\hat{f})$ as our MDL–based curve estimate.

In Lee (1999) "best" choices of $\hat{\boldsymbol{\theta}}$ based on Akaike's information criterion (AIC) and generalized cross–validation (GCV) are also discussed.

3 Minimizing MDL($\hat{f}$) Using Genetic Algorithms

When the number of data points is large, finding the best estimate defined by $\mathrm{MDL}(\hat{f})$ would involve solving a large scale minimization problem. In the literature common techniques for dealing with these types of problems include knot addition, knot deletion, knot movement or combinations of them. However in this work we propose using genetic algorithms for solving our needs. Details are omitted but can be found in Lee (1999).

4 Robust Fitting

This section extends the proposed MDL–based smoothing procedure by using Gaussian mixtures to model outliers. It assumes that the noise e_i's are independent with density function

$$(1-w)p(e; 0, \sigma^2) + wp(e; 0, c^2\sigma^2), \qquad (4)$$

$0 \leq w < 0.5$, $c > 1$, where $p(\cdot; \mu, \sigma^2)$ is the Gaussian density with mean μ and variance σ^2. The percentage of outliers w is assumed to be small and will be estimated, while the variance inflation factor c is fixed *a priori*; see below for more comments on this. Model (4) is often known as inflated–variance model for outliers; see Titterington, Smith & Makov (1985, Ch. 4) and references given therein. Smith & Kohn (1996) also adopted a similar approach to make their Bayesian regression procedure robust.

Suppose that $\hat{B}$, $\hat{\boldsymbol{b}}$, $\hat{\boldsymbol{m}}$ and $\{\hat{\boldsymbol{k}}_j\}_{j=1}^{\hat{B}}$ are specified and that a legitimate guess of which y_i's are contaminated by outliers is made. If there are $\hat{n}_{\mathrm{OUT}}$ of these "suspected outliers", then w can be estimated by $\hat{w} = \hat{n}_{\mathrm{OUT}}/n$, and estimates of α_{ir}'s and β_{ir}'s (hence $\hat{f}$ and $\hat{\boldsymbol{e}}$) can be obtained by performing weighted regressions, with small weights (e.g., c^{-1}) attached to those "outlying y_i's". Recall that for the MDL principle the goal is to find an expression for $L(\boldsymbol{y})$ and use its minimizer as the final estimate. It is outlined in Lee (1999) that, when using (4) to model outliers, the corresponding $L(\boldsymbol{y})$ can be approximated by

$$\begin{aligned} \mathrm{RMDL}(\hat{f}) &= \log \hat{B} + \sum_{j=1}^{\hat{B}} \log \hat{m}_j + \sum_{j=1}^{\hat{B}} \left(3 + \frac{\hat{m}_j}{2}\right) \log \hat{l}_j + \\ &\quad \log \left[\left\{ \frac{1-\hat{w}}{\sqrt{2\pi\hat{\sigma}^2}} \exp \frac{-1}{2\hat{\sigma}^2} + \frac{\hat{w}}{\sqrt{2\pi c^2 \hat{\sigma}^2}} \exp \frac{-1}{2c^2\hat{\sigma}^2} \right\} \times \right. \\ &\quad \left. \sum_{i=1}^{n} \left\{ y_i - \hat{f}(x_i) \right\}^2 \right]. \qquad (5) \end{aligned}$$

Here $\hat{\sigma}^2 = \sum\{y_i - \hat{f}(x_i)\}^2/(n - \hat{n}_{\mathrm{OUT}})$, where the sum does not include those outlying y_i's. Our robust version of the MDL–based smoothing procedure is to define the best estimate as the minimizer of $\mathrm{RMDL}(\hat{f})$. Note that when $\hat{w} = 0$, $\mathrm{RMDL}(\hat{f})$ reduces to $\mathrm{MDL}(\hat{f})$ upto an additive constant.

In Lee (1999) a genetic algorithm for minimizing $\mathrm{RMDL}(\hat{f})$ is also developed.

We have the following remarks about the variance inflation factor c. In theory one does not have to fix it *a priori* as one can estimate $\sigma^2_{\mathrm{OUT}} = c^2\sigma^2$ by a normalized outlier residual sum of squares. However as the number of outliers is usually small, stable estimate of σ^2_{OUT} is hard to obtain. Such instability would almost certainly be carried over to $\mathrm{RMDL}(\hat{f})$, and hence it was decided to choose c *a priori*. As the main focus of this section is to perform robust fitting but not outlier detection, the penalty for misclassifying a y_i as an outlier is not usually severe. Since a small value of c would allow more y_i's to be classified as outliers, as long as the choice of c is not too small, one would not expect the quality of $\hat{f}$ to be severely affected. In our work we have chosen $c = 7$.

5 Numerical Experiment I

Here we report results of two numerical experiments. More simulation results can be found in Lee (1999).

In the first experiment the test functions were the four highly spatial inhomogeneous curves advocated by Donoho & Johnstone (1995). Altogether eight smoothing procedures were tested:

1. mdl: the proposed procedure with $\mathrm{MDL}(\hat{f})$ as the target,
2. rmdl: the proposed robust procedure with $\mathrm{RMDL}(\hat{f})$ as the target,
3. DMS: the Bayesian curve fitting procedure of Denison, Mallick & Smith (1998) with $\lambda = 5$,
4. sure: the SureShrink procedure of Donoho & Johnstone (1995),
5. Bayes: the BayesShrink procedure of Abramovich, Sapatinas & Silverman (1998),
6. SK: the Bayesian regression spline smoothing procedure of Smith & Kohn (1996),
7. RSW: local linear regression with the direct bandwidth plug–in choice of Ruppert, Sheather & Wand (1995), and
8. HST: nearest neighbour local polynomial estimator LOESS (Cleveland & Devlin 1988) with the $\mathrm{AIC_C}$ choice of smoothing parameter proposed by Hurvich, Simonoff & Tsai (1998).

Throughout all the experiments signal–to–noise ratios (snrs) are defined as: $\mathrm{snr} = \{\mathrm{var}(f)/\sigma^2\}^{\frac{1}{2}}$, the same as in Donoho & Johnstone (1995). Three snr levels were used: 5, 7 and 9; but we only report results corresponding to the case when snr=7, as results from the other two snrs are similar. For each combination of test function and snr, 50 sets of noisy observations with $n = 512$ regularly–spaced x_i's were simulated.

For each simulated data set, the eight smoothing procedures listed above were applied to estimate the test function. The numerical measure that was used to evaluate the quality of an estimated curve was $\mathrm{MSE}(\hat{f}) = \sum\{f(x_i) - \hat{f}(x_i)\}^2$. For each test function, the boxplots of the log of the $\mathrm{MSE}(\hat{f})$ values of all $\hat{f}$ are plotted in Fig. 1.

To visually evaluate and compare the performances of the above eight smoothing procedures, the following was done. For the test function *Blocks*, the 50 $\hat{f}$'s obtained by mdl were ranked according to their values of $\mathrm{MSE}(\hat{f})$. The 25th best $\hat{f}$, together with the corresponding simulated noisy data, are plotted in Fig. 2. Curve estimates obtained by applying the remaining seven smoothing procedures to this same simulated noisy data set are also plotted in Fig. 2. The same procedure was then repeated for the remaining three test functions, and the results are displayed in a similar manner in Fig. 2 and Fig. 3.

From these plots one may conclude empirically that mdl has the smallest MSE, and that the performance of rmdl depends heavily on the structures of the underlying curves. The latter is not surprising, as those highly oscillating structures in *Doppler* and *Bumps* could be easily mistaken as outliers by rmdl.

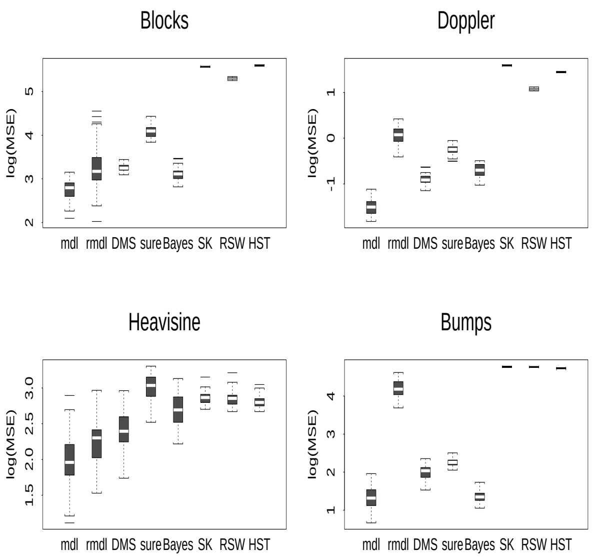

Figure 1: Boxplots of $\log\{\mathrm{MSE}(\hat{f})\}$ values of various smoothing procedures.

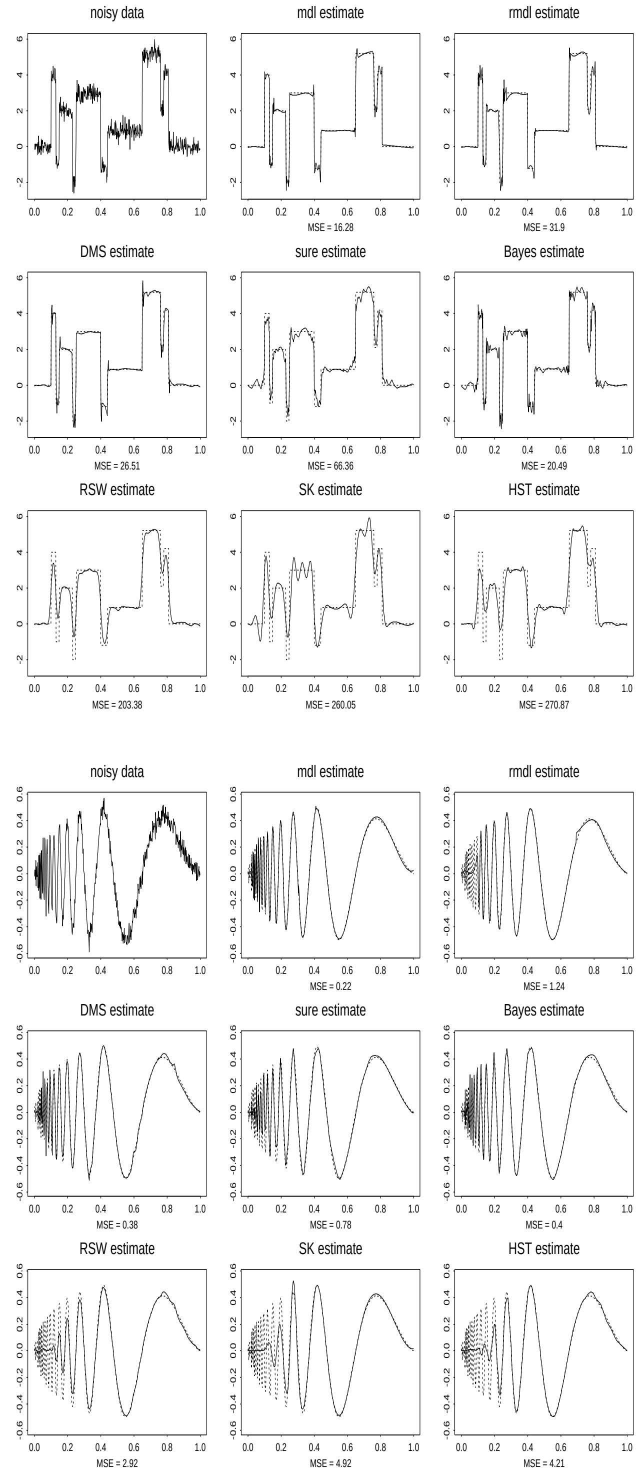

Figure 2: Test Functions *Blocks* (top three rows) and *Doppler* (bottom three rows). Solid lines: estimates; broken lines: true curves.

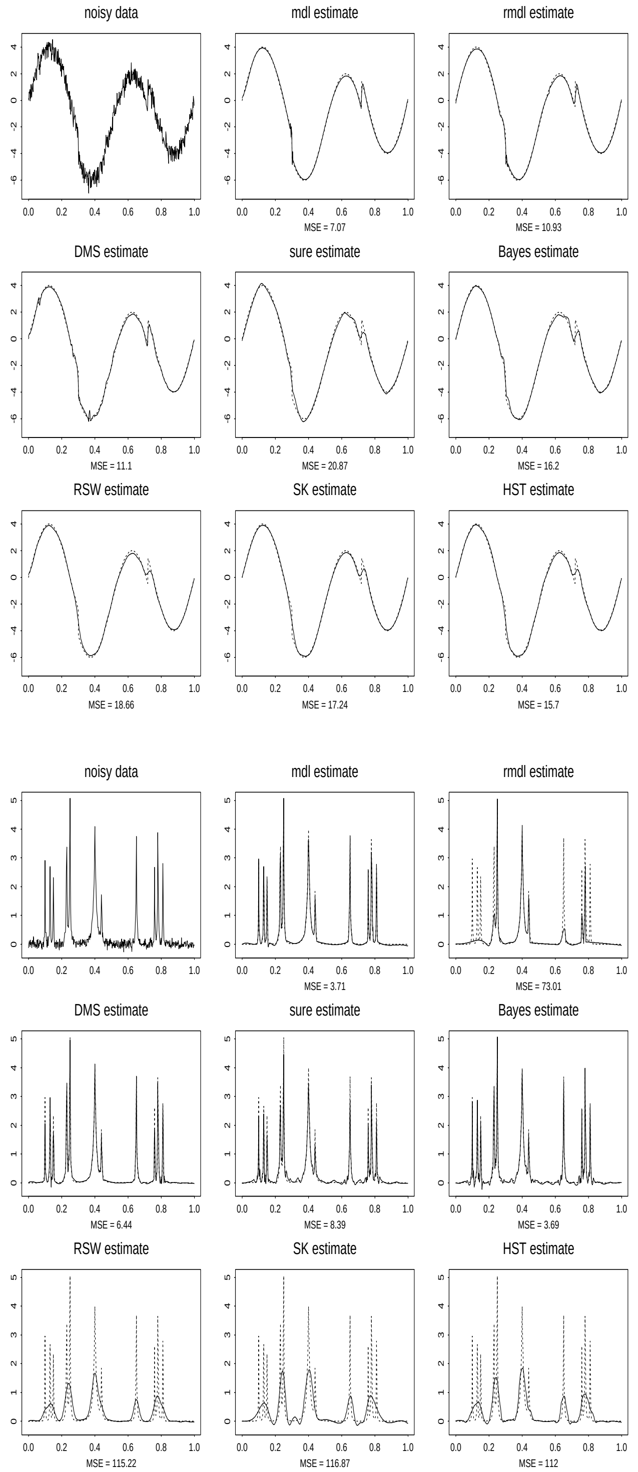

Figure 3: Test Functions *Heavisine* (top three rows) and *Bumps* (bottom three rows). Solid lines: estimates; broken lines: true curves.

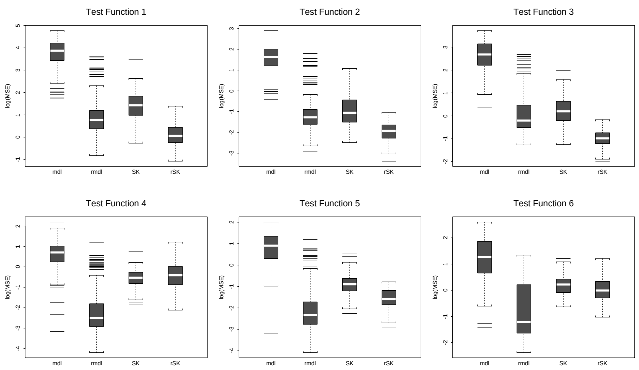

Figure 4: Boxplots of $\log\{\mathrm{MSE}(\hat{f})\}$ values of various smoothing procedures when data were contaminated by outliers.

6 Numerical Experiment II

The main focus of this second experiment was to assess the performance of rmdl when outliers are present. Six test functions were used; three of them are continuous and the remaining three are not. These test functions have been used by other authors and their formulae can be found in Lee (1999).

One hundred data sets were generated from Test Function 1, with snr=4, $n = 200$ and $U[0, 1]$ as the design density. Then for each simulated data set n_{OUT} outliers were introduced, where n_{OUT} is discrete uniform on $[0, \ldots, 5]$. The size of each outlier was generated from $(5+U[0, 15])\sigma$, with equal probability of being positive or negative. Finally four smoothing procedures, mdl, rmdl, SK and the robust version rSK of SK were applied to smooth the data sets. The whole process was repeated with Test Functions 2 to 6.

Boxplots for log(MSE) are given in Fig. 4, and noisy data and various curve estimates corresponding to the "50th smallest MSE rmdl estimates" are given in Fig. 5 and Fig. 6. From these plots one can conclude that, when the data are contaminated by outliers, rmdl outperforms mdl. Also, when the target curves are smooth, rSK is superior to rmdl, and the converse is true if the target curves are non–smooth.

Acknowledgement

The author would like to thank Dr. Mary Morrissey for many constructive and interesting comments, Dr. Kenneth Wilder for introducing him to genetic algorithms, and many authors for making their codes available.

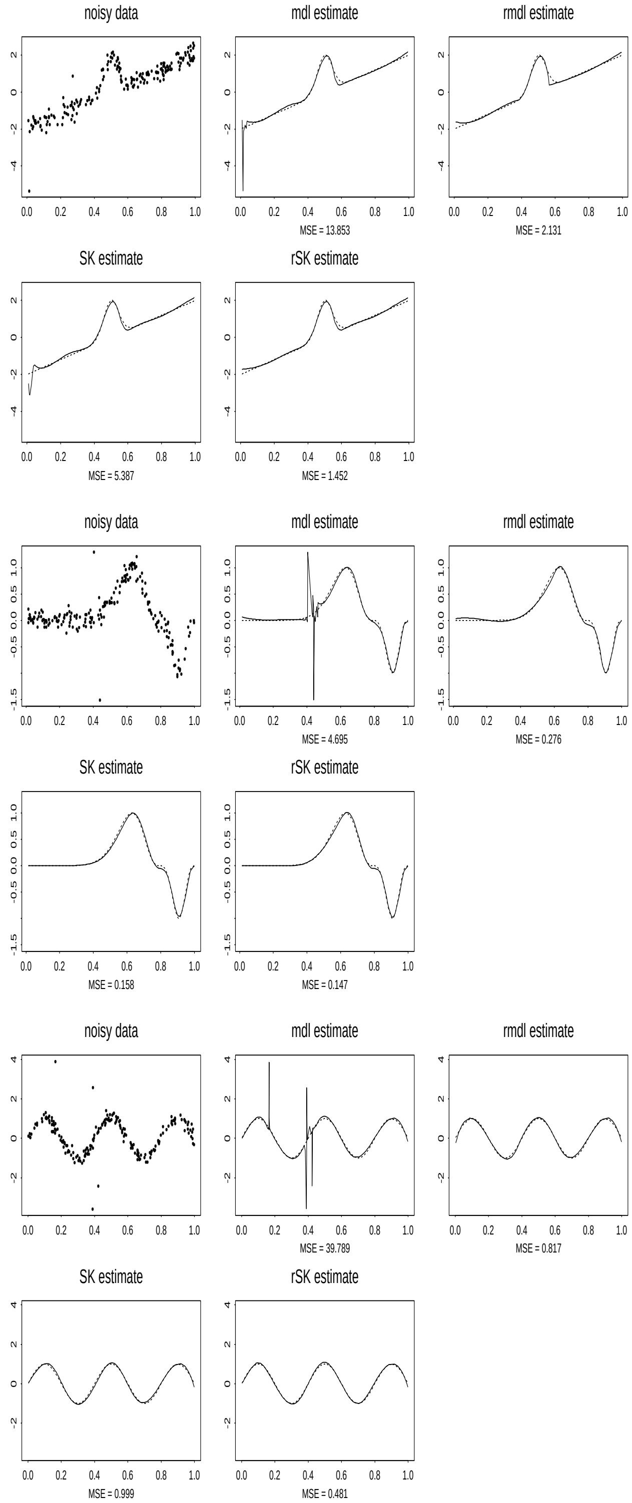

Figure 5: Various estimates of Test Functions 1 (top two rows), 2 (middle two rows) and 3 (bottom two rows) when data were contaminated by outliers. Solid lines: estimates; broken lines: true curves.

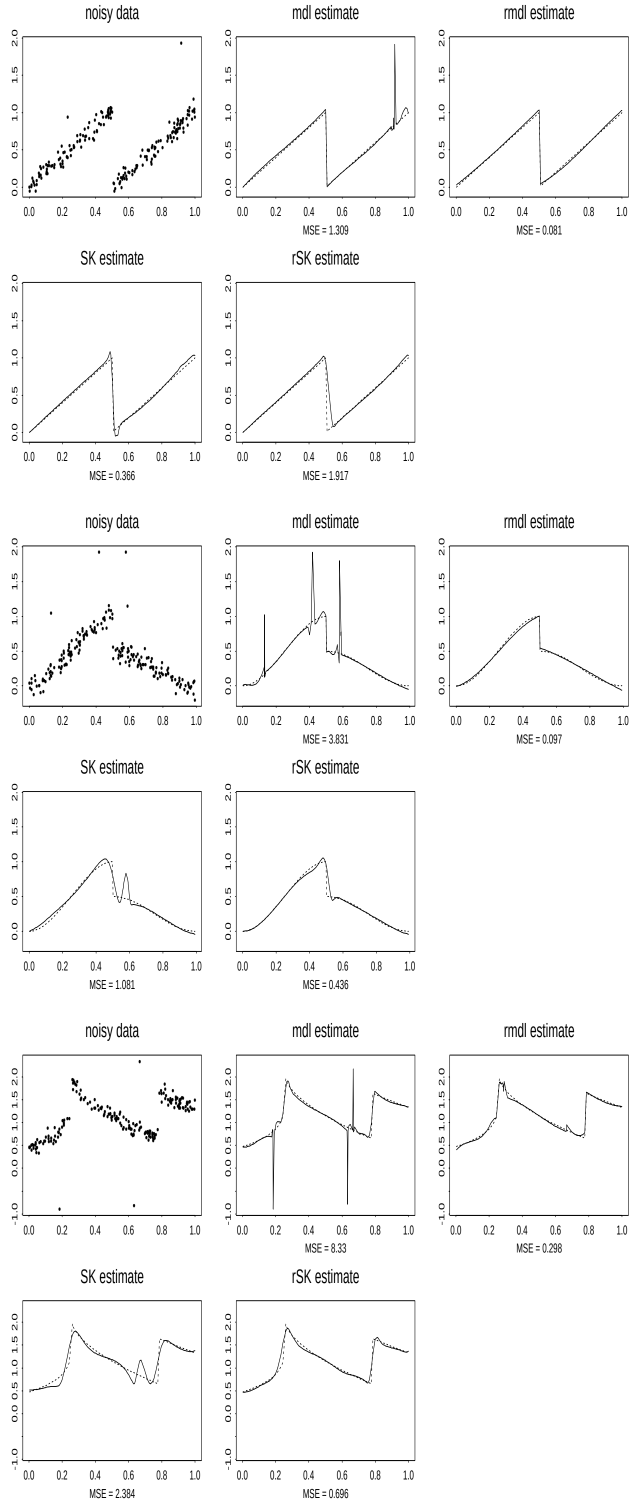

Figure 6: Various estimates of Test Functions 4 (top two rows), 5 (middle two rows) and 6 (bottom two rows) when data were contaminated by outliers. Solid lines: estimates; broken lines: true curves.

References

Abramovich, F., Sapatinas, T. & Silverman, B. W. (1998), 'Wavelet thresholding via a Bayesian approach', *J. Royal Statist. Soc. Ser. B* **60**, 725–749.

Cleveland, W. S. & Devlin, S. J. (1988), 'Locally weighted regression: an approach to regression analysis by local fitting', *J. Amer. Statist. Assoc.* **83**, 596–610.

Denison, D. G. T., Mallick, B. K. & Smith, A. F. M. (1998), 'Automatic Bayesian curve fitting', *J. Royal Statist. Soc. Ser. B* **60**, 333–350.

Donoho, D. L. & Johnstone, I. M. (1995), 'Adapting to unknown smoothness via wavelet shrinkage', *J. Amer. Statist. Assoc.* **90**, 1200–1224.

Hurvich, C. M., Simonoff, J. S. & Tsai, C.-L. (1998), 'Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion', *J. Royal Statist. Soc. Ser. B* **60**, 271–293.

Lee, T. C. M. (1999), 'Robust automatic smoothing for discontinuous regression functions'. Unpublished manuscript.

Rissanen, J. (1989), *Stochastic Complexity in Statistical Inquiry*, World Scientific, Singapore.

Ruppert, D., Sheather, S. J. & Wand, M. P. (1995), 'An effective bandwidth selector for local least squares regression', *J. Amer. Statist. Assoc.* **90**, 1257–1270.

Smith, M. & Kohn, R. (1996), 'Nonparametric regression using Bayesian variable selection', *J. Econometrics* **75**, 317–344.

Titterington, D. M., Smith, A. F. M. & Makov, U. E. (1985), *Statistical Analysis of Finite Mixture Distributions*, John Wiley & Sons Ltd. Chichester.