STAMET-D-06-73 − Revision 2

Design of Kernel M-Smoothers for Spatial Data

Carlo Grillenzoni
University IUAV of Venice
30135 Venezia, Italy
(carlog@iuav.it)

Abstract. Robust nonparametric smoothers have been proved effective in preserving edges in image denoising. As an extention, they are capable to estimate multivariate surfaces containing discontinuities on the basis of a random spatial sampling. A crucial problem is the design of their coefficients, in particular those of the functions which concern robustness. In this paper it is shown that bandwidths which regard smoothness can consistently be estimated, whereas those which concern robustness cannot be estimated with plug-in and cross-validation criteria. Heuristic and graphical methods are proposed for their selection and their efficacy is proved in simulation experiments.

Key Words. Bandwidths selection, Iterative smoothers, Jump preserving, Parametric identifiability, Robust nonparametrics, Spatial data.

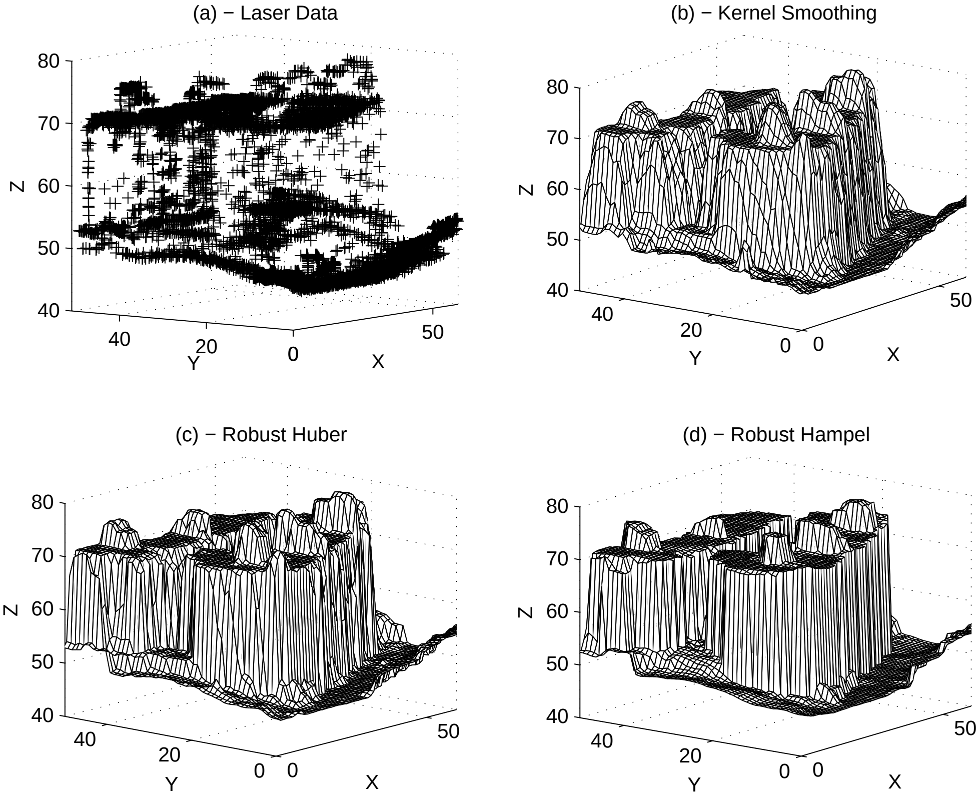
1. Introduction

Nonparametric smoothers are often used to estimate spatial surfaces on the basis of point observations. This happens in many phenomena either natural, as mapping the concentration of a pollutant on the ground (e.g. Francisco-Fernandez and Opsomer, 2005), or social, as monitoring the distribution of crimes in a city (see Levine, 2007). The estimated surfaces are used for creating maps of diffusion and risk and for planning the necessary reclaim actions. Problems arise when data contain jumps and discontinuities, due to the presence of physical and institutional barriers, since classical smoothers tend to blur them.

As an example, Figure 1(a) shows a sample of the laser data used by Wang and Tseng (2004) to measure the elevation relief in a urban area. Each point consists of the planar coordinates (x, y), provided by a GPS receiver, and the terrain height (Z) measured by a laser scanner installed on an aircraft. Data present significant jumps in correspondence of the building, and measurement errors due flight conditions and possible mismatch of the two instruments. Figure 1(b) displays the surface generated by the kernel regression (e.g. Härdle *et al.*, 2002) and shows its inadequacy to track the building walls. Because in correspondence of the jumps the kernel estimates also exhibit the largest residuals, it seems natural to solve the over-smoothing problem with the approach of robust statistics (e.g. Huber, 1981). In practice, data near the building walls can be considered as outliers and robust smoothers tend to ignore them in the local fitting. Figures 1(c,d) show the improvements obtained with M-type smoothers with Huber and Hampel design respectively.

Performance of kernel estimators strongly depends on the bandwidths which tune the degree of smoothing; in the case of discontinuous surfaces, they must be designed as small as possible. However, this is not sufficient to delete the bias at the jump-points and surely increases the variance in smooth regions. On the other hand, robust smoothing is nonlinear and involves additional coefficients which are related to the scale parameters. All of these coefficients can be designed with data-driven criteria based on the cross-validation (CV). In Francisco-Fernandez and Opsomer (2005), it is shown that the presence of spatial correlation in the residuals yields bias in the bandwidths selected with generalized CV. As a solution, they propose modeling the variogram of residuals and adjusting the CV criteria accordingly. This original solution can be directly applied to robust smoothing.

Figure 1. Fitting airborne laser data: (a) Raw point data; (b) Kernel smoothing; (c) Robust smoothing with Huber loss; (d) Robust with Hampel loss.



M-type estimation has provided an important contribution to robust statistics. His philosophy consists of replacing the objective functions of classical estimators with functionals $\rho(\cdot)$ which are less sensitive to extreme values. However, two alternative approaches are present in the literature: Huber (1981) states that $\rho(\cdot)$ must achieve its maximum value asymptotically, because outlying observations may contain useful information. On the contrary, Hampel *et al.* (1986) claim that it should be strictly *bounded*, because outliers are usually extraneous to the models. The two solutions have opposite consequences on the properties of consistency and adaptivity of the estimates, and they were applied to different problems in nonparametric smoothing.

Härdle and Gasser (1984) and Hall and Jones (1990) developed kernel M-smoothers following the first approach, in particular they referred to the Huber's ρ-function. Their estimators were mainly used to resist outliers in univariate models with fixed and random design respectively. Analogously, Wang and Scott (1994) considered the L_1 approach $\rho = |\cdot|$ and proved its robustness for data with heavy-tailed densities. Subsequently, Leung *et al.* (1993, 2005) developed robust strategies to design the bandwidths; they considered (non-quadratic) CV criteria based on the Huber's ρ-function, demonstrating better unbiasedness and consistency. However, they only focused on the coefficients of the smoothing components of the estimators, whereas those which tune robustness were selected a-priori.

M-smoothers which follow the Hampel's approach, were developed by Chu *et al.* (1998). They proved their efficacy as edge-preserving denoising filters for digital images. This property is different from outlier resistance and exploits the adaptive nature of bounded $\rho(\cdot)$ functions. Subsequently, Rue *et al.* (2002), Hwang (2004) extended the method to other kinds of nonparametric estimators, such as local polynomial regression. Boente *et al.* (1997) and Burt and Coakley (2000) discussed bandwidth selection based on robust plug-in strategies. Also in this case, however, only the coefficients of smoothing components were investigated. The difficulty to extend this approach to discontinuous surfaces renders CV preferable.

This paper investigates M-smoothers with bounded ρ-functions in the interpolation of spatial data containing discontinuities. Such data arises in many geostatistical and pattern recognition problems, in particular those concerned with laser scanning (e.g. Figure 1). With respect to image denoising, there is a situation of random regressors and data with an explicit 3D structure. Using the weighted average form of M-estimates (e.g. Hampel *et al.*, 1986) we develop robust smoothers with pseudo-linear structure, which are suitable for large data-sets. Special attention is devoted to bandwidth selection of both smoothing and robustness components; it is shown that problems of parametric identifiability arise in the CV selection of the coefficients which tune robustness, and heuristic solutions are proposed. Throughout, simulation experiments are carried out to illustrate the analyses.

2. Iterative Smoothers

Let $\{(\mathrm{x}_i, \mathrm{y}_i), \mathrm{Z}_i\}_{i=1}^{n}$ be a random sample of spatial measurements as those displayed in Figure 1(a). Assume that they can be represented by the model

$$\mathrm{Z}_i = g(\mathrm{x}_i, \mathrm{y}_i) + \varepsilon_i\,, \qquad \varepsilon_i \sim \mathrm{IID}(0, \sigma_\varepsilon^2)\,; \qquad i = 1, 2 \ldots n \tag{1}$$

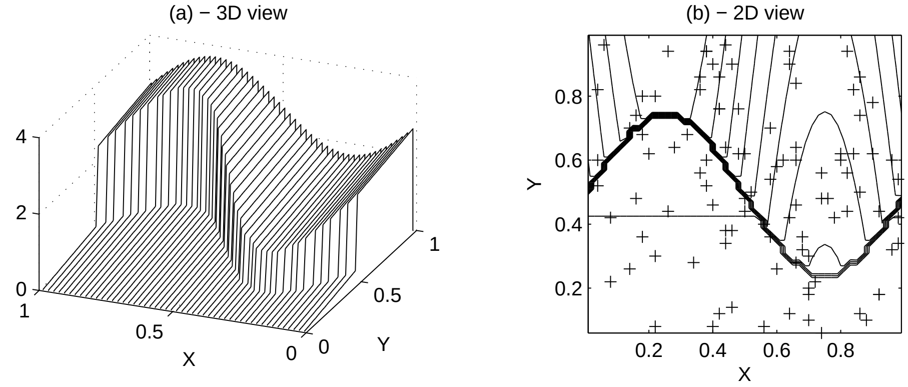
where the planar coordinates have a uniform density $f(x, y) = \alpha$, the response function $g(x, y) = \mathrm{E}(Z \,|\, x, y)$ has jumps located at unknown points, and $\{\varepsilon_i\}$ are independent and identically distributed (IID), with symmetric density $f(\varepsilon)$. Francisco-Fernandez and Opsomer (2005) have considered the problem of autocorrelation of the errors; this may arise either from the spatial clustering of the observations or from the lack of fit of the model (1), which only captures the spatial trend. In the case of uniform sampling and edge-preserving smoothing, the risk of autocorrelated errors is low, whereas in image processing it is a concrete possibility.

As regards the response function of (1), we assume the discontinuous model

$$g(x, y) = \gamma(x, y) + \delta_1 \cdot I\Big\{(x, y) \,:\, y \geq [\,\varphi(x) + \delta_2 \cdot I(x \geq \mathrm{x}_0)\,]\Big\}$$

where $\gamma(\cdot)$ is a continuous function, δ_1, δ_2 are jumps and $I(\cdot)$ is the indicator function. In the above scheme, the discontinuity edge follows the relationship $y = \varphi(x)$, which also has a jump at the point $x = \mathrm{x}_0$; whereas, the continuous part γ is bounded and differentiable. As an example, we consider $\gamma(x, y) = [\,0.75\,y - \sin(6.3\,x)\,]$ and $\varphi(x) = [\,0.5 - 0.125\sin(6.3\,x)\,]$, with $\delta_1 = 2$, $\delta_2 = 0$ and $0 \leq x, y \leq 1$. The surface is

Figure 2. Simulated surface and random sample of the experiment.



displayed in Figure 2(a) on a grid with resolution 50^2. A set of n=100 points is sampled from $g(x, y)$ by assuming $(x, y) \sim \mathrm{U}_2(0, 1)$, a bivariate uniform density. A realization of this design is displayed in Figure 2(b) and the corresponding heights $\mathrm{Z}_i = g(\mathrm{x}_i, \mathrm{y}_i)$ are blurred with a noise.

The experiment consists of reconstructing the surface in Figure 2(a) by fitting the data in Figure 2(b) with various smoothers. The basic scheme is the bivariate kernel (K) regression with *product* kernels (e.g. Härdle *et al.*, 2002 p.89)

$$\begin{aligned} \hat{g}_{\mathrm{K}}(x, y) &= \sum_{i=1}^{n} v_i(x, y)\, \mathrm{Z}_i \\ v_i(x, y) &= \frac{K_1\big[(\mathrm{x}_i - x)/\kappa_1\big]\, K_2\big[(\mathrm{y}_i - y)/\kappa_2\big]}{\sum_{j=1}^{n} K_1\big[(\mathrm{x}_j - x)/\kappa_1\big]\, K_2\big[(\mathrm{y}_j - y)/\kappa_2\big]} \end{aligned} \qquad (2)$$

where $(x, y) \in \Re^2$ are the points where the surface is estimated; K_1, K_2 are symmetric densities, and κ_1, κ_2 are smoothing coefficients. More complex estimators could be considered as in Francisco-Fernandez and Opsomer (2005).

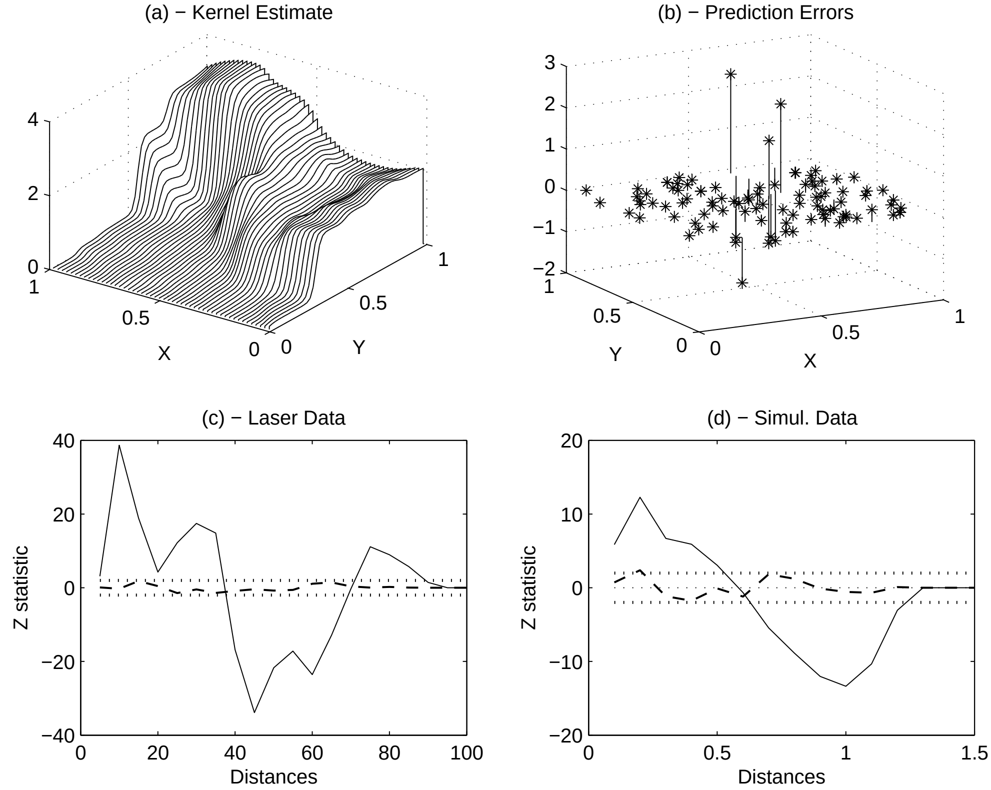
The design of κ_1, κ_2 is usually carried out with the cross-validation (CV) method. It consists of minimizing the sum of squared prediction errors $\hat{\varepsilon}_i$

$$Q_n(\kappa_1, \kappa_2) = \frac{1}{n} \sum_{i=1}^{n} \Big[\mathrm{Z}_i - \hat{g}_{\mathrm{K}-i}(\mathrm{x}_i, \mathrm{y}_i)\Big]^2 \qquad (3)$$

where $\hat{g}_{\mathrm{K}-i}(\cdot)$ are the estimates in (2) obtained by omitting the i-th observation. By replacing Z_i with $g(\mathrm{x}_i, \mathrm{y}_i) + \varepsilon_i$ in (3), it can be seen that minimization of Q_n is asymptotically equivalent to the minimization of the average squared error (ASE): $n^{-1} \sum_i [\, g(\mathrm{x}_i, \mathrm{y}_i) - \hat{g}_{\mathrm{K}-i}(\mathrm{x}_i, \mathrm{y}_i)]^2$ (e.g. Leung *et al.*, 1993 or Härdle *et al.*, 2002 p.114). This fact establishes the asymptotic optimality of the CV-approach.

Applying this framework to the data in Figure 2(b), under the choice of Gaussian kernels and the constraint $\kappa_1 = \kappa_2$, provided $\hat{\kappa}_{\mathrm{CV}} = 0.065$. With this value and the filter (2), we generated the surface in Figure 3(a); panel (b) displays the prediction errors $\hat{\varepsilon}_i$. It can be seen that kernel estimation does not preserve discontinuities and large errors occur in correspondence of the jumps. To have a measure of the non-normality of $\{\hat{\varepsilon}_i\}$ it can be noted that the sample variance was $\hat{\sigma}_\varepsilon^2 = 0.473^2$, whereas the (robust) median absolute deviation (MAD) provided $\hat{\sigma}_{\mathrm{MAD}} = 0.081$.

Figure 3. (a,b) Kernel regression estimates of the data in Figure 2(b), obtained with the algorithm (2), with Gaussian kernels and $\kappa_1 = \kappa_2 = 0.065$. (c,d) Standardized Moran's autocorrelations of series Z_i (solid) and errors $\hat{\varepsilon}_i$ (dashed) of the data in Figures 1 and 2; the dotted bands are 0.95 confidence regions for the null.



We also test for the presence of spatial correlation in the prediction errors. Figures 3(c,d) show the standardized Moran's correlograms of the series Z_i in Figures 1(a), 2(b) and of their kernel residuals. The standardization is done under the null hypothesis and their statistical distribution is asymptotically Normal. Despite of the jumps, original series are strongly autocorrelated, whereas the estimated errors are practically white-noises. Francisco-Fernandez and Opsomer (2005) have shown that the presence of spatial correlation in the residuals increases the variance of $\hat{g}_{\mathrm{K}}$ and the bias of $\hat{\kappa}_{\mathrm{CV}}$ (with respect to the optimal MSE value). They have solved the problem by modeling the residual variogram on the basis of a *pilot* smoothing, and then weighting the generalized CV with the implied autocorrelations. This approach can be directly extended to robust smoothing.

Robust estimation tends to solve the problem of oversmoothing of (2) by reducing the influence of large errors. The formal connection between jumps and outliers can be shown by including the discontinuous component of $g(\cdot)$ in the noise component of (1). Relaxing the ID assumption, it follows that ε_i have a mixture density of the type $f_\varepsilon^* = f_0 \cdot I(y < \varphi(x)) + f_1 \cdot I(y = \varphi(x))$, where f_0 is centred on 0 and f_1 is centred on δ_1. It is the latter which is responsible for outliers.

The kernel M-smoother is the solution of the locally weighted maximum likelihood type problem (e.g. Härdle and Gasser, 1984)

$$\hat{g}_{\mathrm{M}}(x,y) = \arg\min_g \left[R_n(g) = \frac{1}{n} \sum_{i=1}^{n} v_i(x,y)\, \rho\Big(\mathrm{Z}_i - g\Big) \right] \tag{4}$$

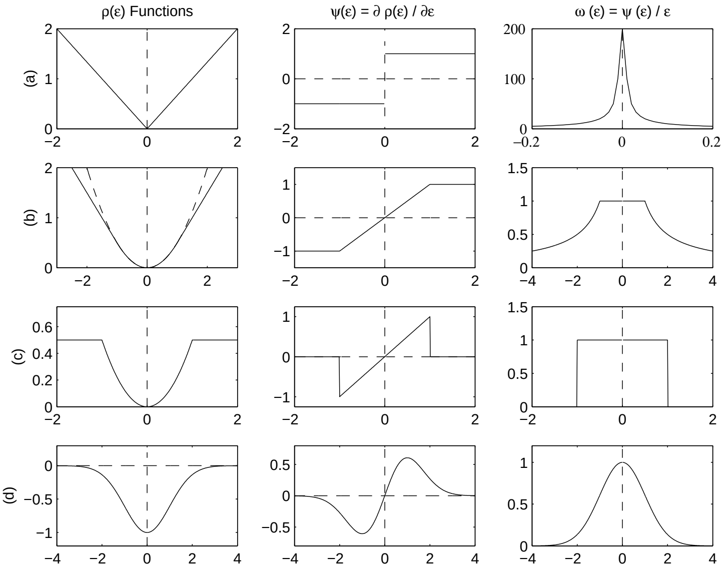
where the local weights $\{v_i\}$ are defined as in (2), and $\rho(\cdot)$ is a loss function which reduces the influence of outlying observations on the estimates. With respect to (1), data near the jump-points can be considered as outliers and robust smoothers tend to ignore them in the local fitting. However, the mechanism acts as a threshold, so that jumps in the estimated surface are finally generated.

Following the Huber and Hampel philosophies, the loss function can be designed as unbounded (a,b) or bounded (c,d), respectively:

$$\begin{array}{rrcl}
\text{a)} & \rho_a(\varepsilon) & = & \Big|\,\varepsilon\,\Big| \\
\text{b)} & \rho_b(\varepsilon) & = & \begin{cases} \varepsilon^2/2\,, & |\varepsilon| \le \lambda \\ \lambda\,|\varepsilon| - \lambda^2/2\,, & |\varepsilon| > \lambda \end{cases} \\
\text{c)} & \rho_c(\varepsilon) & = & \begin{cases} \varepsilon^2/2\,, & |\varepsilon| \le \lambda \\ \lambda^2/2\,, & |\varepsilon| > \lambda \end{cases} \\
\text{d)} & \rho_d(\varepsilon) & = & -L\Big(\varepsilon/\lambda\Big)/\lambda
\end{array} \tag{5}$$

where $L(\cdot)$ is a kernel function and $\lambda > 0$ is a tuning coefficient. The common feature of the above criteria is that the score function $\psi(\varepsilon) = \partial\,\rho(\varepsilon)/\partial\,\varepsilon$ is uniformly bounded; indeed, this is the true necessary condition of robustness. The loss function (5,a) was stressed by Wang and Scott (1994) and is independent of λ; (5,b) is the preferred one of Huber, and has a *monotone* derivative. (5,c) corresponds to the *trimmed* method and (5,d) is a smoothed solution which provides *redescending* ψ-functions. Graphical behavior of these functions is shown in Figure 4.

Figure 4. Display of the loss functions in (5) with $L(\cdot)$ Gaussian and $\lambda = 1$, together with the score function $\psi = \partial\rho/\partial\varepsilon$ and the weight function $\omega = \psi/\varepsilon$.



The minimization of (4), for every point (x, y), typically proceeds by nonlinear algorithms, such as the steepest descent one

$$\hat{g}_{\mathrm{M}}^{(k+1)}(x,y) = \hat{g}_{\mathrm{M}}^{(k)}(x,y) + \sum_{i=1}^{n} v_i(x,y)\,\psi\Big(\mathrm{Z}_i - \hat{g}_{\mathrm{M}}^{(k)}(x,y)\Big) \tag{6}$$

where $\psi(\cdot) = \rho'(\cdot)$ and the initial value is $\hat{g}_{\mathrm{M}}^{(0)}(\cdot) = \hat{g}_{\mathrm{K}}(\cdot)$. However, this solution is computationally demanding, and is suitable only if the grid of values for (x, y) and/or the sample size n are small. A simpler approach, which leads to a quasi-linear (or closed form) solution for (4), can be obtained from the *weighted average* form of M-estimates (see Hampel *et al.*, 1986 p.115). Introducing the constraint $K_1 = K_2$ and the notation $K_\kappa(\cdot) = K(\cdot/\kappa)/\kappa$, one has the following:

Proposition 1. *Assume that the nonlinear model (1) is estimated with the kernel M-smoother (4) with bounded ρ-function (5,d), with $L(\cdot)$ Gaussian. Then, the iterative algorithm (6) admits the "pseudolinear" representation*

$$\hat{g}_{\mathrm{M}}^{(k+1)}(x,y) = \frac{\sum_{i=1}^{n} K_\kappa(\mathrm{x}_i - x)\,K_\kappa(\mathrm{y}_i - y)\,L_\lambda\Big[\mathrm{Z}_i - \hat{g}_{\mathrm{M}}^{(k)}(x,y)\Big]\,\mathrm{Z}_i}{\sum_{i=1}^{n} K_\kappa(\mathrm{x}_i - x)\,K_\kappa(\mathrm{y}_i - y)\,L_\lambda\Big[\mathrm{Z}_i - \hat{g}_{\mathrm{M}}^{(k)}(x,y)\Big]} \tag{7}$$

Proof. See the Appendix 4.1.

Algorithm (7) resembles the linear estimator (2) but is iterative, hence the term quasi-linear or pseudo-linear. Its peculiar feature is the local weighting also in the direction of the dependent variable Z; actually, it is this weighting that allows robustness. It should also be noted the relationship of (7) with the linear σ-filter applied by Chu *et al.* (1998) to the denoising of digital images Z_{ij}

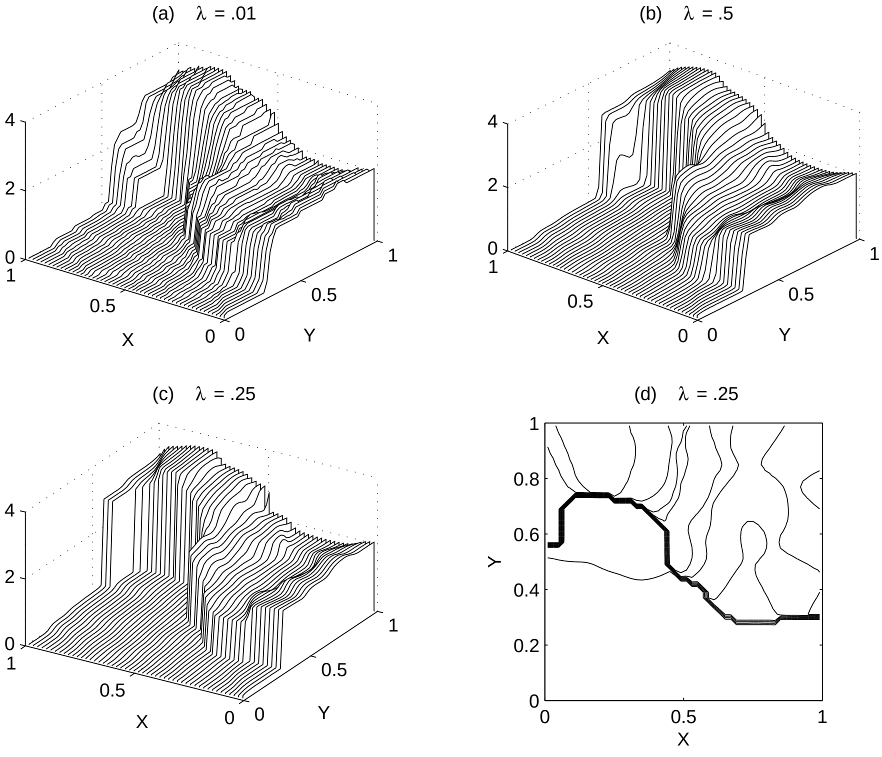
$$\hat{g}_{\mathrm{S}}(i,j) \propto \sum_{h=1}^{n_1}\sum_{k=1}^{n_2} K_\kappa\Big(i-h\Big)\, K_\kappa\Big(j-k\Big)\, L_\lambda\Big(Z_{ij}-Z_{hk}\Big)\, Z_{ij}$$

The analogy becomes evident if one replaces $\hat{g}_{\mathrm{M}}^{(k)}(x,y)$ within $L(\cdot)$, with the observation Z_j which is spatially closer to Z_i. However, this substitution significantly worsens the jump-preserving ability of the estimator.

The version of (7) corresponding to the trimmed ρ-function (5,c), can be obtained by replacing $L(\cdot)$ with the indicator function $I\Big(\Big|\,\mathrm{Z}_i - \hat{g}_{\mathrm{M}}^{(k)}(x,y)\,\Big| \le \lambda\Big)$ (see panel 9 of Figure 4). Instead, in the case of unbounded ρ-functions (5;a,b) the structure of $L(\cdot)$ is much more complex. The estimates in Figure 1(c) for laser data were obtained with the nonlinear (4) and the Huber loss, whereas those in Figure 1(d) were generated with the pseudolinear (7) and the Hampel loss. The jump-preserving ability of the latter is better, due to the fact that bounded ρ-functions are more robust and so have better adaptive properties on the edges. In the presence of large data-sets (as in laser scanning), the speed of (7) can be improved by splitting the sample in disjoint random sub-subsets. These sub-samples can then be processed *sequentially* along with the iterations of the algorithm.

Returning to the simulation experiment, Figure 5 shows the estimates (7) computed on the data of Figure 2(b). They were generated with Gaussian kernels, $\kappa = 0.065$ (which is the CV bandwidth of the kernel estimation) and tentative values of $\lambda \in (0.01, 0.5)$. The resulting surfaces range from near instability (as in panel (a)) to near oversmoothing (as in panel (b)); as a consequence, the best choice seems the intermediate solution in panels (c,d) with λ=0.25. We discuss in detail the problem of bandwidth selection in the next section. As in Figure 1(c), we have also checked that the jump-preserving ability of the smoother (4) with loss functions (5;a-c) was inferior to that of (7).

Figure 5. Surfaces obtained by fitting the data in Figure 2(b) with the smoother (7) under the designs K, L Gaussian, κ=0.065 and λ=0.01 (a), 0.5 (b), 0.25 (c,d).



The ability of robust smoothers to track discontinuities arises from the fact that they have better local properties with respect to the classical kernel regression. In particular, they weight observations also in the direction of the dependent variable Z. Anomalous observations generated by jumps are treated as outliers and therefore they are censored in the local estimation. As regards the relationships between the various algorithms, we have shown that the nonlinear problem (4) can be solved with the weighted smoother (7), which mimics the kernel regression (2). This proves the close connection between M-smoothers and the σ-filter used in image processing (e.g. Chu *et al.*, 1998). Finally, the extension of the method to *local polynomial* regression can be obtained by replacing the function g with a polynomial $g_i(x, y)$ (see Rue *et al.*, 2002 and Hwang, 2004). If this is linear, the quantity to be used in R_n-(4) is $\rho[\,\mathrm{Z}_i - g_0 - g_1\,(x - \mathrm{x}_i) - g_2\,(y - \mathrm{y}_i)\,]$, where g_0 provides the surface estimate. This approach enables smaller bias at boundary regions, however it has a greater computational complexity and numerical instability.

3. Bandwidth Selection

For kernel smoothers, the CV method is asymptotically optimal, in the sense that it provides bandwidths which minimize the asymptotic mean integrated squared error (AMISE) of the estimates $\hat{g}_{\rm K}$ (e.g. Härdle *et al.*, 2002 pp.110-114). Thus, the criterion Q_n-(3) could also be used for robust smoothers; the sole warning is that at each i-th point, the estimates must be properly iterated. For example, for the algorithm (7) the leave-one-out estimate of the regression function is

$$\hat{g}_{{\rm M}-j}^{(k+1)}({\rm x}_j,{\rm y}_j) \propto \sum_{i\neq j}^{n} K_\kappa({\rm x}_i-{\rm x}_j)\,K_\kappa({\rm y}_i-{\rm y}_j)\,L_\lambda\Big[{\rm Z}_i-\hat{g}_{{\rm M}-j}^{(k)}({\rm x}_j,{\rm y}_j)\Big]\,{\rm Z}_i$$

computational advantages of pseudolinear algorithms are now clear. Application to the data of Figure 2(b) provided $\hat{\kappa}_{\rm CV}=0.06$ and $\hat{\lambda}_{\rm CV}\to\infty$, which means that the robust smoother tends to the kernel estimator, namely $\hat{g}_{\rm M}\to\hat{g}_{\rm K}$. This disappointing situation was also observed by Hall and Jones (1990, p.1717) as regards M-smoothers with Huber ρ-function, applied to regression models contaminated by outliers. However, they did not investigated the underlying causes.

Because classical CV is based on squared prediction errors, it is very sensitive to outliers generated by jumps, even when a robust smoother is used to preserve them. The presence of just one outlier is sufficient to yield biased estimates of the bandwidths, either in the direction of oversmoothing or undersmoothing. Leung *et al.* (1993) and Wang and Scott (1994) have solved this problem by using *robust* cross-validation (RCV) criteria. They claim that a better approximation to the MSE optimality in finite samples can be obtained by minimizing

$$P_n(\kappa,\lambda)=\frac{1}{n}\sum_{j=1}^{n}\varrho\Big[\,{\rm Z}_j-\hat{g}_{{\rm M}-j}^{(k)}({\rm x}_j,{\rm y}_j)\,\Big] \qquad (8)$$

where $\varrho[\,\cdot\,]$ can be one of the ρ-functions in (5). The side-effect of (8) is that the ϱ-error optimality of the estimates is not demonstrated, just because the expression of κ which minimizes ${\rm E}\{\int\!\!\int\varrho[\,\hat{g}_{\rm M}(x,y)-g(x,y)]\,{\rm d}x\,{\rm d}y\}$ is unknown (e.g. Boente *et al.*, 1997). However, under regularity conditions, as those listed in the Appendix, the RCV approach is asymptotically optimal.

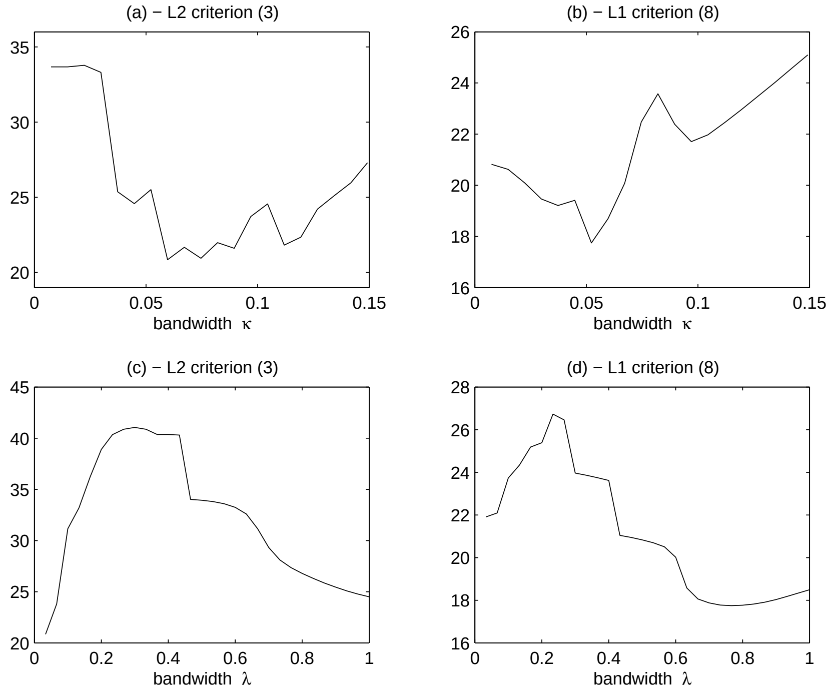
Proposition 2. *Assume that model (1) has continuous response function (i.e. $g = \gamma$), and is estimated by the M-smoother (4). Then, under the assumptions A1-A5 listed in the Appendix, the bandwidth $\hat{\kappa}_{\rm RCV}$ which minimizes the criterion (8), is asymptotically optimal in MSE sense and is independent of the function $\varrho(\cdot)$.*

Proof. See Leung (2005) and Appendix 4.2.

This result confirms the optimality of the CV approach (e.g. Hall and Jones, 1990 p.1754), even in its robust version (8), whose main constraint is that $\varrho(\cdot)$ has bounded first derivative. The practical consequence is that robust *plug-in* strategies of Boente *et al.* (1997) can be avoided in the estimation of $\kappa_{\rm opt}$. These methods are better than cross-validation, but are computationally demanding. On the other hand, many authors have not considered the estimation of λ and just selected it a-priori in the context of the Huber ρ-function (e.g. Leung, 2005).

The simplest choice of $\varrho(\cdot)$ in the functional P_n-(8) is the absolute criterion (5,a), because it does not need the specification of a tuning coefficient λ_ϱ (say). For the other ϱ-functions (5;b-d), the coefficient λ_ϱ must be designed according to the

Figure 6. Paths of quadratic and absolute CV functions for the coefficients (κ, λ) of the smoother (7), applied to the data of Figure 2(b).



experimental conditions (e.g. Leung, 2005). The graphs of the CV functions (3) and (8) applied to the data in Figure 2(b), is given in Figure 6. In agreement with Proposition 2, we found that robust criteria have a similar path; therefore, we only show that of the absolute ϱ-function. Here, we can see that quadratic and absolute criteria yield close estimates of κ (0.05, 0.06). Instead, for the second coefficient, only the robust criterion has a well-defined minimum at $\lambda = 0.75$. However, this value is too large and does not enable jump-preserving (see Figure 5(b)).

Figure 6 shows that CV methods are inadequate to design the robustness coefficient λ. The reason is that such coefficient should be sensitive to the jumps and discontinuity edges, but these points form a set which has area zero. In other words, the number of observations near to, or on the jump-points is too small to influence the CV functions. This situation can be formalized as follows:

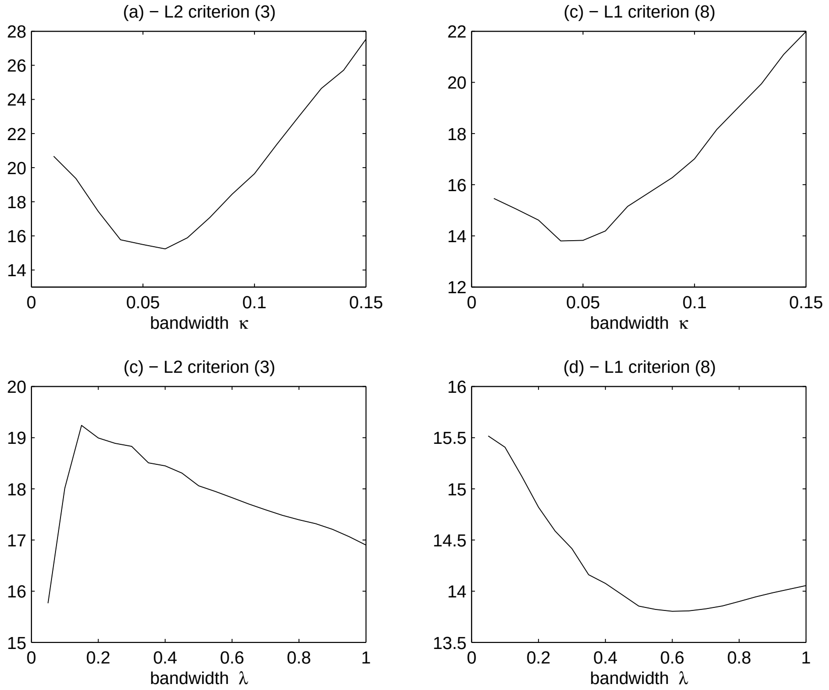
Proposition 3. *Assume that the model (1) is estimated with the M-smoothers (4),(7) and assume that the bandwidths* κ, λ *are selected with the CV criteria (3),(8). Then, the robustness coefficient* λ *is not parametrically identified, i.e.* $\lambda_{\text{opt}} \to \infty$.

Proof. See the Appendix 4.3.

This result explains the difficulties of Hall and Jones (1990), Chu *et al.* (1998) and Hwang (2004) in using quadratic CV to select the robustness coefficients of M-smoothers of Huber and Hampel type. On the other hand, Wang and Scott (1994) and Boente *et al.* (1997) just avoided the problem by using the absolute criterion, both in the equation (4) and (8). Finally, Leung *et al.* (1993, 2005) assigned a-priori values to λ on the basis of the experimental conditions.

Before proceeding, it is useful to investigate the behavior of the cross-validation criteria in Monte Carlo simulations. Figure 7 has the same information content as Figure 6, but is based on the mean values of 30 independent replications. The patterns are now much clearer, and confirm that Q_n-(3) and P_n-(8) have similar and well-defined minima with respect to κ (0.055, 0.045). Instead, the minimum point $\lambda = 0.6$ in Figure 7(d) is biased upward because yields oversmoothing. These results empirically confirm the conclusion of Proposition 3.

Figure 7. Mean values of quadratic and absolute CV functions, for the coefficients (κ, λ) of the smoother (7), on 30 replications of data of the model (1).



In robust statistics, λ has the role of scale parameter and is estimated or designed according to the distribution of outliers. However, when this is unknown, λ should just realize a trade-off between efficiency and robustness of the estimates (e.g. Maronna *et al.*, 2006 p.65). Indeed, the robustness is inversely proportional to λ, but the efficiency (in the absence of outliers), is directly proportional to it. Now, setting $\lambda = C\sigma_\varepsilon$, with $C > 0$, it can be shown that M-estimates of the mean of a Gaussian model maintain 95% relative efficiency with respect to least-squares only if $1 < C < 5$. Specifically, for the Huber loss one has $C_H = 1.345$, whereas for the Tukey bisquare function one has $C_T = 4.685$ (see Fox, 2002). This approach can be extended to M-smoothers with negative Gauss loss.

Proposition 4. *Assume that model (1) has continuous response function (i.e. $g \equiv \gamma$) and Gaussian disturbances ε_i. Then, under the assumptions A1-A5 listed in the Appendix, the M-smoother (7) with design $\lambda = 2.111\,\sigma_\varepsilon$ maintains 95% asymptotically relative efficiency (ARE) with respect to the Kernel estimator (2).*

Proof. See the Appendix 4.4.

Since the bandwidth κ and the scale σ_ε are consistently estimable (see Proposition 2), we can define effective strategies to design λ. With reference to the data in Figure 2(b) and the value $\hat{\kappa}_{\text{RCV}} = 0.05$ of Figure 6(b), we have 3 solutions:

1. *Heuristic.* Under the assumption of $f(\varepsilon)$ Gaussian, the ARE solution of size 94% suggests $\lambda = 2\,\sigma_\varepsilon$. In this case, the crucial point is the choice of the error variance. The estimate $\hat{\sigma}_{\text{MAD}} = 0.076$ yields the mild value $\hat{\lambda}_{\text{ARE}} = 0.152$, and the results in Figure 5 confirm its validity.

2. *Constrained.* In order to solve the non-identifiability problem of λ, one may impose the constraint $\lambda = D\,\kappa$ and then apply the cross-validation selection. The definition of the constant $D > 0$ strongly depends on the structure of the kernel $L(\cdot)$. However, a specific design could be obtained from the heuristic solution, such as $\hat{D} = \hat{\lambda}_{\text{ARE}}/\hat{\kappa}_{\text{RCV}} \approx 3$.

3. *Graphical.* Chu *et al.* (1998) suggested "visual evaluation" of the estimates $\hat{g}_{\text{M}}$ to tune the bandwidths. This approach can be made less subjective by defining the set of *admissible* values. Running the M-smoother (7) with $\hat{\kappa}_{\text{RCV}}$ and tentative values for λ, one can find the sets $S_1 = \{\lambda < \lambda_1^*\}$ for which it is unstable (namely $\hat{g}_{\text{M}} \to \infty$), and $S_2 = \{\lambda > \lambda_2^*\}$ for which it is oversmoothed (i.e. $\hat{g}_{\text{M}} \to \hat{g}_{\text{K}}$). The optimal design is then given by the midpoint $\lambda^* = (\lambda_1^* + \lambda_2^*)/2$, and in Figure 5 we obtained $\lambda^* = 0.25$.

Summarizing the results of these methods we have $\lambda \in (0.15, 0.25)$. It is interesting noting that this set includes the maximizing point of Figure 7(c), rather than the minimizing point of 7(d). Given the relationship between quadratic cross-validation and the average squared error, it can be shown that the path of $\text{ASE}(\lambda|\hat{\kappa}_{\text{RCV}}) = n^{-1}\sum_i[\,g(\text{x}_i, \text{y}_i) - \hat{g}_{\text{M}}(\text{x}_i, \text{y}_i)]^2$ is close to Figure 7(c). Hence, it seems that optimal value of λ is the one which maximizes the MSE. This *seeming* contradiction arises from the fact that discontinuity edges have area zero and, in smooth regions, the estimator $\hat{g}_{\text{M}}$ is less efficient than $\hat{g}_{\text{K}}$. In other words, the adaptivity of robust smoothers at the jump-points is largely paid for in the continuous regions.

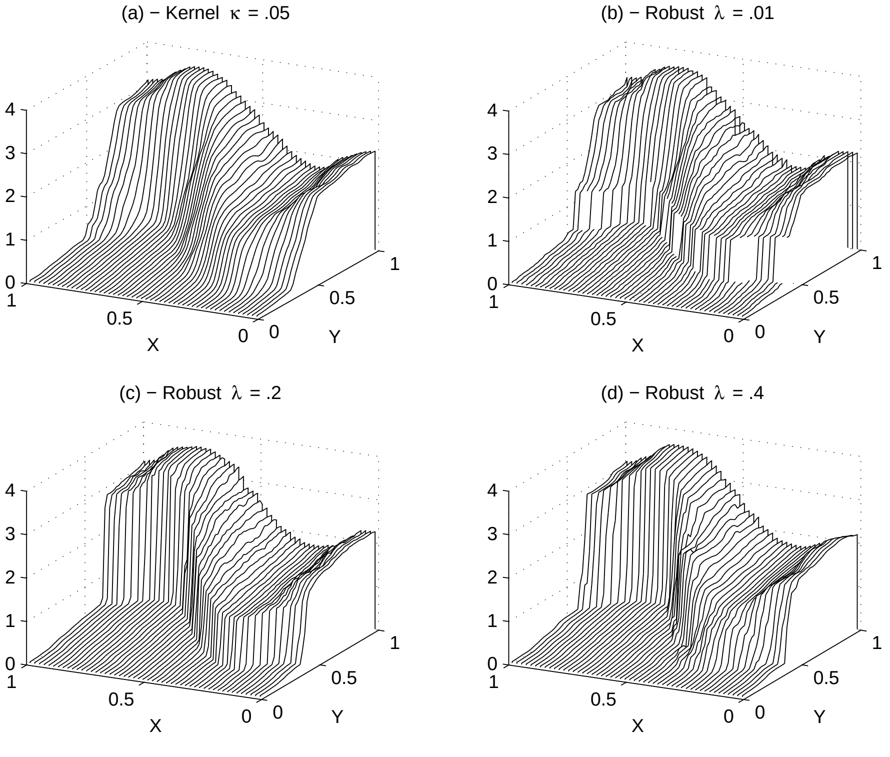
The simplest selection strategy we have outlined so far can be summarized as:

Algorithm. *The robust design of the M-smoothers (4),(7) is given by:*
Step 1. Select $\hat{\kappa}_{\rm RCV}$ by minimizing the robust criterion (8) with $\varrho = |\cdot|$,
Step 2. Select $\hat{\lambda}_{\rm ARE} = 2\,\hat{\sigma}_{\rm MAD}$ by using the prediction errors at Step 1.

At Step-1, given the non-estimability of λ, it may be preferable to use a simple Kernel smoother such as (2). In this case, the pattern of the RCV function $P_n(\kappa)$ would become much more smooth than those in Figures 6(b) and 7(b).

We conclude by iterating the experiment of Figure 5 on ten independent realizations $\{{\rm x}_i, {\rm y}_i, {\rm Z}_i\}$. Running kernel and robust smoothers with $\hat{\kappa} = 0.05$ (see Figure 7(a,b)) and $\lambda = 0.01, 0.2, 0.4$ and averaging the resulting estimates, we obtained the surfaces in Figure 8. Again, we see that $\lambda = 0.01$ generates instability (the blanks in the panel (b) mean ∞ values) and $\lambda = 0.4$ yields oversmoothing. Hence, the best choice is the midpoint $\lambda^* = 0.2$.

Figure 8. Mean values of kernel and robust estimates (2) and (7), generated with $\kappa = 0.05$ and $\lambda \in [\,0.01, 0.4\,]$, on 10 independent replications.



4. Conclusions

In this paper we have investigated the problem of fitting point data, sampled by discontinuous surfaces, with robust smoothers. Main fields of application are in geostatistics and environmetrics, especially in urban areas. We have check that kernel M-smoothers have a good jump-preserving ability if they are implemented with bounded loss functions (or redescending score functions). This performance can be attributed to the best adaptive capability of robust smoothers with respect to the conventional ones. Using the weighted average form of M-type estimates, we have developed iterative smoothers which retain a linear structure. These turns out particularly useful for computing cross validation functions and, therefore, for selecting optimal bandwidths. In this paper we have shown that coefficients which tune robustness are not parametrically identified (with respect to CV) and heuristic rules must be adopted for their design. We have proposed 3 strategies which provide similar values and have worked well in simulation experiments.

4. Appendix: heuristic proofs

Main Assumptions

We summarize the technical assumptions underlying the nonlinear model (1) and the nonparametric estimators.

A1. The density function $f(x,y)$ is uniform and has bounded support, $f(\varepsilon)$ is symmetric about zero and twice differentiable $f''(\varepsilon)$.

A2. The continuous component $\gamma(\cdot)$ of the regression function $g(\cdot)$ is twice differentiable and has $\iint \gamma''(x,y)^2 \, \mathrm{d}x \, \mathrm{d}y < \infty$.

A3. The kernel functions K_1, K_2, L of the M-estimators (4) and (7) are continuous density functions, symmetric about zero.

A4. The loss functions $\rho(\cdot)$, $\varrho(\cdot)$ of the M-estimator (4) and CV criterion (8), are symmetric about zero and have bounded first derivatives.

A5. Asymptotic analyses of the estimators are performed under the condition $n\kappa \to \infty$ as $n \to \infty$ and $\kappa \to 0$, where $\kappa = (\kappa_1 = \kappa_2)$.

Let us also define the following moments and integral notations

$$\mu_2(F) = \int u^2 F(u)\,\mathrm{d}u, \qquad S_2(F) = \int F(u)^2\,\mathrm{d}u \tag{9}$$

4.1 Proof of Proposition 1

From the first order condition of (4) one has $R'_n(g) = \sum_i v_i(x,y)\psi(\mathrm{Z_i} - g) = 0\,;$ letting $0 = g - g$ and iterating one obtain the steepest descent algorithm (6). In the parametric context, M-estimators in *weighted average* form were introduced by Tukey by defining the residual weight function $\omega(\varepsilon) = \psi(\varepsilon)/\varepsilon$ (see Hampel *et al.*, 1986 p.115). Now, inserting $\psi(\varepsilon) = \omega(\varepsilon)\,\varepsilon$ in the equation $R'_n(g) = 0$, we have

$$\sum_{i=1}^n v_i(x,y)\,\omega(\mathrm{Z_i} - g)\,\mathrm{Z_i} = \sum_{i=1}^n v_i(x,y)\,\omega(\mathrm{Z_i} - g)\,g$$

and solving for g, in iterative form, provides the weighted (W) smoother

$$\hat{g}_{\mathrm{W}}^{(k+1)}(x,y) = \left[\sum_{i=1}^n v_i(x,y)\,\omega\Big(\mathrm{Z}_i - \hat{g}_{\mathrm{W}}^{(k)}(x,y)\Big)\right]^{-1} \sum_{i=1}^n v_i(x,y)\,\omega\Big(\mathrm{Z}_i - \hat{g}_{\mathrm{W}}^{(k)}(x,y)\Big)\,\mathrm{Z}_i \tag{10}$$

Now, in the case of the loss function (5,d), with $L(\cdot)$ Gaussian, one has

$$\psi\Big(\mathrm{Z}_i - g\Big) = \frac{-1}{\sqrt{2\pi}\lambda}\exp\left[-\frac{1}{2}\Big(\frac{\mathrm{Z}_i - g}{\lambda}\Big)^2\right]\frac{-1}{\lambda^2}\Big(\mathrm{Z}_i - g\Big)$$

that is $\omega(\cdot) \propto L(\cdot)$ (see Figure 4), and the estimator (7) directly follows from (10). In parametric models, it can be shown that W-estimators have the same influence function and asymptotic variance as M-estimates (e.g. Hampel *et al.*, 1986 p. 116). If the weights $v_i(x,y)$ are non-negative, the same property can be extended to robust smoothers and one can conclude that (6) and (7) are statistically equivalent.

4.2 Proof of Proposition 2

Consider the simple model $Z = g(x) + \varepsilon$, with g continuous and x fixed (this will be relaxed later). For the M-smoother (4), Leung (2005) has shown that the RCV criterion is asymptotically equivalent to the average squared error (ASE) and its expectation. It follows that $\hat{\kappa}_{\mathrm{RCV}}$ which minimizes (8) converges to

$$\hat{\kappa}_{\mathrm{MASE}} = \arg\min_{\kappa}\ \mathrm{E}\Big\{\mathrm{ASE} = \frac{1}{n}\sum_{i=1}^n \Big[\hat{g}_{\mathrm{M}}(\mathrm{x}_i) - g(\mathrm{x}_i)\Big]^2\Big\}$$

Now, the mean ASE is asymptotically equivalent to the mean integrated squared error (MISE, see Härdle *et al.*, 2002 p.110), whose asymptotic expression for (4) can be obtained from Härdle and Gasser (1984) or Boente *et al.* (1997)

$$\mathrm{AMISE}[\,\hat{g}_{\mathrm{M}}(x)\,] = \frac{1}{4}\,\kappa^4\,\mu_2^2(K)\,S_2(g'') + \frac{1}{n\kappa}S_2(K)\,\mathrm{E}[\rho'(\varepsilon)^2]\,\mathrm{E}[\rho''(\varepsilon)]^{-2} \tag{11}$$

where μ_2, S_2 are operators defined in (9) and $\rho(\cdot)$ is the loss function of (4). Notice that the above combines elements of the simple kernel regression and the parametric M-estimation of a location parameter (e.g. Huber, 1981) and is minimized by

$$\kappa_{\mathrm{opt}} = \left\{ \frac{S_2(K)}{\mu_2^2(K)\,S_2(g'')\,n} \cdot \frac{\mathrm{E}[\rho'(\varepsilon)^2]}{\mathrm{E}[\rho''(\varepsilon)]^2} \right\}^{1/5} \tag{12}$$

Summarizing, we have $\hat{\kappa}_{\mathrm{RCV}} \to \hat{\kappa}_{\mathrm{MASE}} \to \kappa_{\mathrm{opt}}$, and the remarkable fact is that κ_{opt} is independent of the loss function $\varrho(\cdot)$ used in (8). It follows that also $\hat{\kappa}_{\mathrm{RCV}}$ is independent, the only constraint is that $\varrho'(\cdot)$ is bounded (see assumption A4).

Under the assumption A1, the extension of (12) to x stochastic is direct because the density $f(x)$ is uniform. However, the analysis must be performed with conditional expectations to the set $\boldsymbol{X}_n = [\mathrm{x}_1, \mathrm{x}_2 \ldots \mathrm{x}_n]$. The MASE becomes $\mathrm{E}\{\mathrm{ASE}\,|\boldsymbol{X}_n\}$, and the conditional MISE becomes $\mathrm{E}\{\int [\,\hat{g}_{\mathrm{M}}(x) - g(x)\,]^2 f(x)\,\mathrm{d}x\,|\boldsymbol{X}_n\}$, which is also weighted by $f(x)$ (see Wand and Jones, 1995 p.138). When x is stochastic, the conditional variance behaves like $\sqrt{n\kappa}\;\mathrm{V}[\,\hat{g}_{\mathrm{K}}(x)|\boldsymbol{X}_n] \to S_2(K)\,\sigma_\varepsilon^2/f(x)$; moreover, if $f'(x) = 0$ the conditional bias is equal to that of x fixed (e.g. Härdle *et al.* 2002 p.93). It follows that the asymptotic conditional MISE of $\hat{g}_{\mathrm{M}}$ is similar to (11) with just $S_2(g'')$ recomputed as $\int g''(x)^2 f(x)\,\mathrm{d}x$ (see also Hall and Jones, 1990 p.1715). This slight modification should also be inserted in (12).

4.3 Proof of Proposition 3

Consider the simple model $Z = g(x) + \varepsilon$, with a jump located at the point $x = \mathrm{x}_0$. Under the assumptions A1-A5, the asymptotic MSE of the M-estimator (4) has been investigated by Chu *et al.* (1998) and Rue *et al.* (2002). Its conditional expression can be summarized as follows

$$\mathrm{E}\Big\{[\,\hat{g}_{\mathrm{M}}(x)-g(x)]^2\Big|\boldsymbol{X}_n\Big\} \approx \begin{cases} (\pi\,\delta)^2+\pi(1-\pi)\,\delta^2, & x\in(\mathrm{x}_0\pm\kappa) \quad \text{(a} \\ C_1\,\kappa^4+C_2/(n\,\kappa\,\lambda^3), & x \ \text{elsewhere} \quad \text{(b} \end{cases} \tag{13}$$

where $\delta > 0$ is the size of the jump and $\pi = \int_{\delta/2}^{\infty} f(\varepsilon)\,\mathrm{d}\varepsilon$. The equation (13,a) shows that the MSE does not vanish asymptotically and, therefore, the M-smoother is not consistent at the jump point x_0. However, the formula of π shows that if $f(\varepsilon)$ has a bounded support, with range less than δ, then the consistency may exist (see also Hillebrand and Müller, 2006).

The formula (13,b) holds for $g(\cdot)$ continuous, and the constants are given by

$$C_1 = \frac{1}{4}\,\mu_2^2(K)\,g''(x)^2\,, \qquad C_2 = \sigma_\varepsilon^2\,S_2(K)\,S_2(L')\,f_\varepsilon(0)\,f_\varepsilon''(0)^{-2}$$

Burt and Coakley (2000) have provided an expression of the MSE which also includes a complex term $C_3\,\kappa^3/(n\,\lambda^5)$. However, this confirms that the robustness bandwidth is present in (13) only in the form $1/\lambda$. It follows that the minimization of the MSE only admits the solution $\lambda_{\mathrm{opt}} \to \infty$, which means that the coefficient is not parametrically identified. The practical consequence is that λ cannot be designed with CV techniques because these are asymptotically MSE optimal.

Expression (13,b) suggests other important remarks about λ. First, the consistency of $\hat{g}_{\mathrm{M}}$ in continuous regions holds even when $\lambda \gg 0$; the true necessary condition is that $n\kappa \to \infty$ (see assumption A5). This fact is natural in parametric M-estimation where λ has the role of scale parameter, but it is not well recognized in the analysis of M-smoothers, where it is considered a bandwidth, hence $\lambda \to 0$ (see Hwang, 2004 or Hillebrand and Müller, 2006). Finally, the expression of κ which minimizes the AMISE of (13) is similar to (12) with $\rho = -L$

$$\kappa_{\mathrm{opt}}^* = \left\{ \frac{\sigma_\varepsilon^2\,S_2(K)}{\mu_2^2(K)\,S_2(g'')\,n} \cdot \frac{S_2(L')\,f_\varepsilon(0)}{\lambda^3\,f_\varepsilon''(0)^2} \right\}^{1/5}$$

This bandwidth could be estimated with robust plug-in methods (see Boente *et al.*, 1997, or Burt *et al.*, 2000), but they require an a-priori choice for λ.

4.4 Proof of Proposition 4

Consider the simple model $Z = g(x) + \varepsilon$, with g continuous and $f(x)$ uniform; the weighted conditional AMISE of the kernel regression (2) is given by (11) with the term $\mathrm{E}(\cdot)/\mathrm{E}(\cdot)^2$ replaced by σ_ε^2. Therefore, from the variance components, the asymptotic relative efficiency (ARE) formula becomes

$$\mathrm{ARE}\Big(\hat{g}_{\mathrm{M}}, \hat{g}_{\mathrm{K}}\Big) = \frac{\sigma_\varepsilon^2}{\tau_\rho}, \qquad \tau_\rho = \frac{\mathrm{E}\,[\,\rho'(\varepsilon)^2\,]}{\mathrm{E}\,[\,\rho''(\varepsilon)\,]^2} \tag{14}$$

which is independent of κ. Now, considering the ρ-function (5,d), with $L(\cdot)$ Gaussian, one has $\rho'(\varepsilon) = -\mathrm{N}(\varepsilon; 0, \lambda)\varepsilon/\lambda^2$; therefore

$$\begin{aligned}
\mathrm{E}\left[\rho'(\varepsilon)^2\right] &= \int \left[\frac{u/\lambda^2}{\sqrt{2\pi}\,\lambda}\exp\Big(\frac{-\varepsilon^2}{2\lambda^2}\Big)\right]^2 \left[\frac{1}{\sqrt{2\pi}\,\sigma_\varepsilon}\exp\Big(\frac{-\varepsilon^2}{2\sigma_\varepsilon^2}\Big)\right] \mathrm{d}\varepsilon \\
&= \frac{1}{2\pi\lambda^6\sigma_\varepsilon}\int \frac{\varepsilon^2}{\sqrt{2\pi}}\exp\Big(\frac{-\varepsilon^2}{\lambda^2} + \frac{-\varepsilon^2}{2\sigma_\varepsilon^2}\Big)\,\mathrm{d}\varepsilon \\
&= \frac{1}{2\pi\lambda^6\sigma_\varepsilon}\sqrt{\alpha}\int \frac{\varepsilon^2}{\sqrt{2\pi\alpha}}\exp\Big(\frac{-\varepsilon^2}{2\alpha}\Big)\,\mathrm{d}\varepsilon\,, \qquad \alpha = \frac{\sigma_\varepsilon^2\lambda^2}{2\sigma_\varepsilon^2 + \lambda^2} \\
&= \frac{\sigma_\varepsilon^2}{2\pi\lambda^3(2\sigma_\varepsilon^2 + \lambda^2)^{3/2}}
\end{aligned} \tag{15}$$

Analogously, it can be shown that $\rho''(\varepsilon) = \mathrm{N}(\varepsilon; 0, \lambda)/\lambda^2 - \mathrm{N}(\varepsilon; 0, \lambda)\varepsilon^2/\lambda^4$, and using $\beta = (\sigma_\varepsilon^2\lambda^2)/(\sigma_\varepsilon^2 + \lambda^2)$ one can obtain

$$\begin{aligned}
\mathrm{E}\left[\rho''(\varepsilon)\right] &= \mathrm{E}\left[\frac{1}{\sqrt{2\pi}\,\lambda^3}\exp\Big(\frac{-\varepsilon^2}{2\lambda^2}\Big)\right] - \mathrm{E}\left[\frac{\varepsilon^2}{\sqrt{2\pi}\,\lambda^5}\exp\Big(\frac{-\varepsilon^2}{2\lambda^2}\Big)\right] \\
&= \frac{1}{\sqrt{2\pi(\lambda^2 + \sigma_\varepsilon^2)}\,\lambda^4}\int\left[\frac{\lambda^2}{\sqrt{2\pi\beta}}\exp\Big(\frac{-\varepsilon^2}{2\beta}\Big) - \frac{\varepsilon^2}{\sqrt{2\pi\beta}}\exp\Big(\frac{-\varepsilon^2}{2\beta}\Big)\right]\mathrm{d}\varepsilon \\
&= \frac{1}{\sqrt{2\pi(\lambda^2 + \sigma_\varepsilon^2)}\,\lambda^4}\Big(\lambda^2 - \beta\Big) \;=\; \frac{1}{\sqrt{2\pi(\lambda^2 + \sigma_\varepsilon^2)^3}}
\end{aligned} \tag{16}$$

Substituting (15) and (16) into (14), and letting $\lambda = C\sigma_\varepsilon$, it follows

$$\mathrm{ARE}\Big(\hat{g}_{\mathrm{M}}, \hat{g}_{\mathrm{K}}\Big) = \frac{\lambda^3(2\sigma_\varepsilon^2 + \lambda^2)^{3/2}}{(\lambda^2 + \sigma_\varepsilon^2)^3} = \frac{C^3(2 + C^2)^{3/2}}{(C^2 + 1)^3}$$

and for C=2 we have the level of ARE=0.94. Of course, for kernel smoothers this result holds only in continuous regions of a regression surface.

References

Boente G., Fraiman R. and Meloche J. (1997), Robust Plug-in Bandwidth Estimators in Nonparametric Regression. *Jour. of Stat. Plann. Inf.*, **57**, 109-142.

Burt D.A. and Coakley C.W. (2000), Automatic Bandwidth Selection for Modified M-Smoothers. *Journal of Statistical Computation and Simul.*, **66**, 295-319.

Chu C.K., Glad I., Godtliebsen F. and Marron J.S. (1998), Edge-Preserving Smoothers for Image Processing. *Jour. of Amer. Stat. Assoc.*, **93**, 526-541.

Francisco-Fernandez M. and Opsomer J.D. (2005), Smoothing parameter selection methods for nonparametric regression with spatially correlated errors. *Canad. Jour. of Statistic*, **33**, 279-295.

Fox J. (2002), *An R and S-PLUS Companion to Applied Regression.* Sage Publications, Thousand Oaks (CA).

Hall P. and Jones M.C. (1990), Adaptive M-Estimation in Nonparametric Regression. *The Annals of Statistics*, **18**, 1712-17-28.

Hampel F., Ronchetti E., Rousseeuw P. and Stahel W. (1986), *Robust Statistics: the Approach Based on Influence Functions.* Wiley, New York.

Härdle W. and Gasser T. (1984), Robust Non-parametric Function Fitting. *Journal of Royal Statistical Society, ser. B*, **46**, 42-51.

Härdle W., Müller M., Sperlich S. and Werwatz A. (2002), *Nonparametric and Semiparametric Models.* Springer, Berlin.

Hillebrand M. and Müller C.H. (2006). On Consistency of Redescending M-Kernel Smoothers. *Metrika*, **63**, 71-90.

Huber P.J. (1981), *Robust Statistics.* Wiley, New York.

Hwang R.-C. (2004), Local Polynomial M-smoothers in Nonparametric Regression. *Journal of Statistical Planning and Inference*, **126**, 55-72.

Leung D.H.-Y., Marriott F.H.C. and Wu E.K.H. (1993), Bandwidth Selection in Robust Smoothing. *Journal of Nonparametric Statistics*, **2**, 333-339.

Leung D.H.-Y. (2005), Cross-Validation in Nonparametric Regression with Outliers. *Annals of Statistics*, **33**, 2291-2310.

Levine N. (2007), *CrimeStat: A Spatial Statistics Program for the Analysis of Crime Incident Locations.* National Institute of Justice, Washington DC. Available at `http://www.icpsr.umich.edu/CRIMESTAT`

Maronna R.A., Martin R.D. and Yohai V.J. (2006), *Robust Statistics: Theory and Methods.* Wiley, New York.

Rue H., Chu C.-K., Godtliebsen F. and Marron J.S. (2002), M-Smoother with Local Linear Fit. *Journal of Nonparametric Statistics*, **14**, 155-168.

Wand M.P. and Jones M.C. (1995), *Kernel Smoothing.* Chapman & Hall, London.

Wang F. and Scott D. (1994), The L1 Method for Robust Nonparametric Regression. *Journal of the American Statistical Association*, **89**, 65-76.

Wang M. and Tseng Y.-H. (2004), LiDAR data segmentation and classification based on octree structure. Available at `http://www.isprs.org/istanbul2004/comm3/papers/286.pdf`