AStA 781 - Revision 2

Robust Nonparametric Estimation of the Intensity Function of Point Data

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

Abstract. Intensity functions describe the spatial distribution of the occurrences of point processes and are useful for risk assessment. This paper deals with robust nonparametric estimation of the intensity function of space-time data as earthquakes. The basic approach consists of smoothing the frequency histograms with the local polynomial regression (LPR) estimator. This method enables automatic boundary corrections and its jump-preserving ability can be improved with robustness. A robust local smoother is derived from the weighted-average approach to M-estimation and its bandwidths are selected with robust cross-validation (RCV). Further, a robust recursive algorithm is developed for sequential processing of the data binned in time. An extensive application to the Northern California earthquake catalog in the area of San Francisco illustrates the method and proves its validity.

Key Words. Cross-Validation, Earthquake Data, Kernel Regression, Local Polynomial, M-type Estimation, Recursive Algorithms, San Francisco Bay.

Acknowledgments: I am grateful to the Referees for their helpful remarks.

1. Introduction

A space-time point process is a sequence of random variables, whose realizations provide the spatial location and the occurrence time of the events. Typical examples are represented by earthquakes, but also urban crimes, birth-death of firms and epidemic phenomena belongs to this category (see Daley and Vere-Jones, 2003). Unlike continuous processes, which are representable on regular lattices, the main features of a point process are the coordinates of the events. Forecasting *where* and *when* a future event will occur is the final target in many fields of research. The statistical analysis of a point process is mainly concerned with the estimation of its *intensity function*. This is related to the conditional density and provides the frequency with which events are expected to occur in the neighbor of any point. This is useful for building maps of diffusion, contamination and risk.

Nonparametric estimation of intensity functions is commonly used when the mathematical structure of the process is unknown. In particular, the entire methodology of kernel density estimation (KDE) can be applied in multivariate form. This approach has been pursued in seismology by several authors: the seminal paper by Vere-Jones (1992) compared parametric and nonparametric approaches; the books by Bailey and Gatrell (1995) and Simonoff (1996) contains many numerical applications. More recently, Choi and Hall (1999, 2000) have included time in kernel smoothers, Stock and Smith (2002 a,b) have applied adaptive estimation, and Grillenzoni (2005) has developed sequential methods.

In the real world, intensity functions may not be completely smooth and are characterized by the presence of *discontinuities* in the form of edges and jumps. These features usually occur at the borders, as a consequence of physical and institutional barriers, but may also be produced by the dynamics of the process itself. As regards seismology and the distribution of earthquakes, typical examples are represented by tectonic faults and shocks after quiescent times. For social processes, they are represented by urban morphology and political changes. In any event, when edges are present in the intensity function, the simple kernel method tends to blur them and provides oversmoothed estimates.

An alternative approach to kernel density estimation consists of fitting the empirical histograms with a nonparametric smoother and then normalizing the area under the resulting surface (e.g. Fan and Gijbels, 1996 p.50). The local polynomial regression (LPR) has automatic boundary correction properties (see Cheng *et al.*, 1997) and can alleviate the problem at the borders. This approach has been extensively applied to univariate density functions in survival analysis, randomized experiments, effect treatment studies where censoring and grouping data are frequent (e.g. Bouezmarni and Scaillet, 2005).

However, to solve the problem of edge effects in a systematic manner, one should use *robust* (M-type) smoothers. These estimators have been successfully applied as edge-preserving filters to denoise digital images (e.g. Chu *et al.* 1998, Rue *et al.* 2002, Hillebrand and Müller, 2006). Common smoothers reduce the noise by local averaging the pixel luminance; however, this also blur the edges which separate homogeneous zones. Robust filters reduce this drawback since the score components which control outliers behave like threshold functions on the edges. As an extention, they could be applicable to smooth point data and their histograms, although important algorithmic adjustments are necessary.

In this paper we derive a robust LPR smoother based on the *weighted-average* form of M-estimates (see Hampel *et al.*, 1986 p.115). Using this approach to robustness seems natural for nonparametric smoothers because they are usually expressed as weighted means. The resulting algorithm utilizes kernel functions in place of the usual score functions, and this makes its structure totally nonparametric. Subsequently, the problem of bandwidth design is faced from the point of view of robust cross-validation (RCV, see Wang and Scott, 1994). This approach enables optimal selection of the coefficients which tune local adaptation (e.g. Leung, 2005); instead, for those which tune robustness, heuristic solutions are necessary. Finally, a recursive version of the M-smoother is applied to the estimation of the space-time intensity function of the earthquake data of San Francisco. Empirical results on real and simulated data supports the validity of the proposed methods.

2. Robust Regression and Density Estimation

Seismic data can be seen as a realization of a *marked* space-time point process. This is defined as a sequence of multivariate random variables $\{(x_k, y_k, z_k), t_k; m_k\}$ ordered by time (e.g. Daley and Vere-Jones, 2003). In particular, $k = 1, 2 \ldots N$ is the index of the sequence, (x, y, z) are spatial coordinates (longitude, latitude, depth), t is time and m is the magnitude (mark) of the events. The probabilistic properties of the point process are entirely described by its joint distribution $F[(x, y, z), t; m]$; the intensity function is defined in *conditional* form as $f[(x, y, z), \tau | \Im_t]$, where $\tau \leq t$ and $\Im_t$ is the set of information (history) up to time t. More specifically, if $\#(\cdot)$ counts the number of events in a neighbor of the point $\boldsymbol{p} = [(x, y, z), \tau]'$, then the conditional intensity $f(\cdot)$ is defined from the equation

$$\mathrm{P}\Big[\#(\,\boldsymbol{p}, \boldsymbol{p} + \mathrm{d}\boldsymbol{p}) > 0 \,\Big|\, \Im_t\Big] = f\Big(\boldsymbol{p}\,\Big|\Im_t\Big)\,\mathrm{D}\boldsymbol{p} + o\Big(\mathrm{D}\boldsymbol{p}\Big)$$

where $\Im_t = \{\,\boldsymbol{p}_k : \tau_k \leq t\}$ and $\mathrm{D}\boldsymbol{p} = \mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z\,\mathrm{d}\tau$ (see Daley and Vere-Jones, 2003 Chap. 13). The intensity function $f(\cdot)$ provides the rate of occurrence at the point $\boldsymbol{p}$ given the information up to time t; the process is stationary if it is invariant under translations in time and space.

Apart from theoretical definitions and parametric modelings (e.g. Zhuang *et al.*, 2002), the estimation of the conditional intensity on real data is usually performed with nonparametric methods. Omitting the depth coordinate z and assuming multiplicative kernels, Vere-Jones (1992), Bailey and Gatrell (1995), Choi and Hall (1999), Stock and Smith (2002) have focused on the kernel density

$$\hat{f}_N(x, y, t) = \frac{1}{N\kappa_1\kappa_2\kappa_3} \sum_{k=1}^{N} K_1\Big(\frac{\mathrm{x}_k - x}{\kappa_1}\Big)\, K_2\Big(\frac{\mathrm{y}_k - y}{\kappa_2}\Big)\, K_3\Big(\frac{\mathrm{t}_k - t}{\kappa_3}\Big)$$

where $K_i\,, i = 1, 2, 3$ are kernel functions; $\kappa_i > 0$ are their bandwidths; $(\mathrm{x}_k, \mathrm{y}_k, \mathrm{t}_k)$ are observations; (x, y, t) are variables and N is the sample size.

The above estimator treats the time dimension as a spatial axis, in the sense that it moves on it in any direction, and the estimates at any instant t also include future events. However, this contrasts with the *unreversible* nature of time and/or with the conditional nature of the intensity. To respect these features, one can

consider a sequential implementation and weighting observations with a one-sided exponential sequence tuned by a discounting factor $\mu \in (0,1)$

$$\hat{f}_{\mathrm{K}}(x,y|t) = \Big(\kappa_1\kappa_2 \sum_{\{\,k\,:\,\mathrm{t}_k \le t\,\}} \mu^{t-\mathrm{t}_k}\Big)^{-1} \sum_{\{\,k\,:\,\mathrm{t}_k \le t\,\}} \mu^{t-\mathrm{t}_k}\, K_1\Big(\frac{\mathrm{x}_k - x}{\kappa_1}\Big)\, K_2\Big(\frac{\mathrm{y}_k - y}{\kappa_2}\Big) \quad (1)$$

An interesting feature is that when the ages t_k are binned in a discrete sequence, the recursive version of (1) becomes $\hat{f}(x,y|t) = \mu\, \hat{f}(x,y|t-1) + (1-\mu)\, \hat{d}_t(x,y)$, where $\hat{d}_t(\cdot)$ is the *instantaneous* kernel density (see Grillenzoni, 2005 p.74).

Robust Estimation. In this paper, we develop robust versions of the adaptive estimator (1). The reason is that robustness has the advantage both to resist outlying observations and to preserve jumps and edges. For parametric densities, jumps usually arise at the borders (as in the exponential and uniform cases), but in a nonparametric framework discontinuities may be present everywhere. In particular, for point processes related to urban crimes (e.g. Levine, 2007), breaks usually occur both in space and time as a consequence of physical barriers or social changes. In any case, developing robust solutions for kernel densities is not simple since the definition of residuals and prediction errors is not direct.

Following Fan and Gijbels (1996, p.50), an alternative approach to kernel density estimation consists of fitting the empirical frequencies with a nonparametric smoother. For a space-time point process, this requires building a multivariate histogram by binning data in a suitable 3D grid. To simplify the exposition, we first consider sequential 2D histograms; given values of $t \ge \mathrm{t}_1$ and a regular grid of points $x_i,\ i = 1,2\ldots n_1$ and $y_j,\ j = 1,2\ldots n_2$, we define the frequencies

$$\mathrm{f}_{ij|t} = \sum\nolimits_{\{\,k\,:\,(\,x_{i-1} < \mathrm{x}_k \le x_i\,) \cap (\,y_{j-1} < \mathrm{y}_k \le y_j\,)\,|\,(\,\mathrm{t}_k \le t\,)\,\}} \frac{\mathrm{m}_k}{\bar{m}_t}$$

where $\{\mathrm{x}_k, \mathrm{y}_k, \mathrm{t}_k, \mathrm{m}_k\}$ are the data values. Notice that each event is weighted by its relative mark, where m_k is the observed magnitude and $\bar{m}_t$ is the mean computed on the data available up to time t. Adaptation in time can also be achieved by discarding oldest observations with a moving window $(t-T < \mathrm{t}_k \le t\,]$. On the basis of the frequencies, we can define the nonparametric intensity model

$$\mathrm{f}_{ij|t} = f(\dot{x}_i, \dot{y}_j|t) + \varepsilon_{ij}\,, \quad \varepsilon_{ij} \sim \mathrm{IN}(0,\sigma^2_\varepsilon)\,, \qquad \begin{matrix} i = 1,2\ldots n_1 \\ j = 1,2\ldots n_2 \end{matrix} \quad (2)$$

where $(\dot{x}_i, \dot{y}_j)$ are the central points of the square bins of the grid. In this context, robust density estimators can just be obtained by applying kernel M-type smoothers to (2). Such methods were introduced by Härdle and Gasser (1984) and were extended to local linear regression (LLR) by Fan *et al.* (1994). Their main field of application was regression models contaminated by outliers.

Bivariate M-smoothers have been proved effective to preserve edges in digital images denoising (see Chu *et al.*; 1998, Rue *et al.*, 2002; Hillebrand and Müller, 2006); this approach can potentially be applied to smooth point data and their histograms. The robust (M-type) local polynomial regression (LPR) estimator of the intensity model (2) can be defined as

$$\begin{aligned} \hat{f}_{\mathrm{M}}(x, y|t) &= \arg\min_{\beta_0} \left\{ \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} K_1\Big(\frac{\dot{x}_i - x}{\kappa_1}\Big) K_2\Big(\frac{\dot{y}_j - y}{\kappa_2}\Big) \times \right. \\ &\times \left. \rho\Big[\mathrm{f}_{ij|t} - \beta_0 - \sum_{k=0}^{p} \sum_{h=0}^{p} \beta_{kh} \, (\dot{x}_i - x)^k \, (\dot{y}_j - y)^h \Big] \right\} \end{aligned} \tag{3}$$

where the indexes k, h satisfy the constraint $(k+h) = p$. The components of (3) are defined as follows: $\rho[\,\cdot\,]$ is a convex function that controls the influence of anomalous data; β_0 corresponds to the function $f(\cdot)$; β_{kh} are auxiliary coefficients that improve stability at the borders; finally, $p \geq 0$ is the degree of the polynomial expansion of (2). For $p = 0, 1$ we have robust versions of the simple kernel smoother and of the local linear regression, respectively; for $p = 1$ the score component of (3) becomes $\rho[\, \mathrm{f}_{ij|t} - \beta_0 - \beta_{10}(\dot{x}_i - x) - \beta_{01}(\dot{y}_j - y)\,]$.

Edge-preserving is related to outlier resistance by the fact that observations near or on jump-points typically yield anomalous residuals. It follows that these data tend to be censored by the score components of robust smoothers. However, such scores behave as threshold functions, so that discontinuities in the estimated regression surface are finally generated. More generally, since M-smoothers weight observations also in the direction of the dependent variable, their local and adaptive properties are better than those of nonrobust estimators. On the other hand, outlier-resistance and jump-preserving need different types of loss functions.

The choice of the ρ-function in (3) concerns issues of robust (parametric) statistics. Huber (1981) states that $\rho(\cdot)$ should be *unbounded* and must achieve its

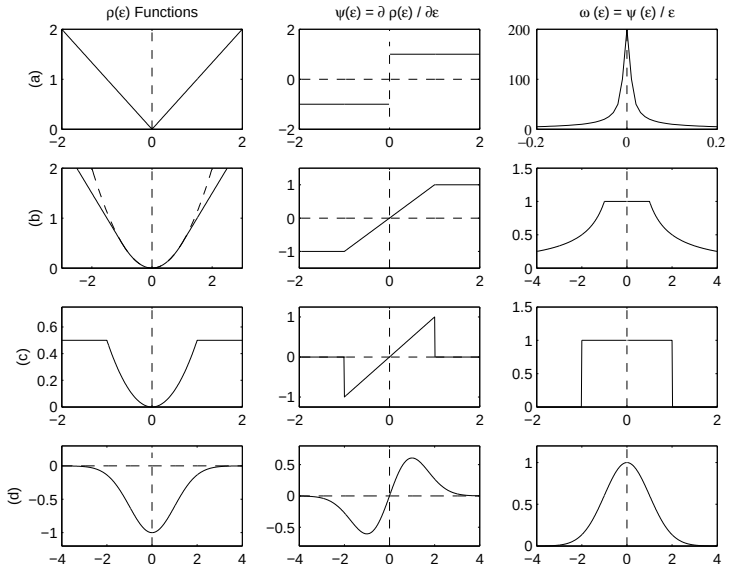
maximum value asymptotically, because outlying observations may contain useful information. On the contrary, Hampel *et al.* (1986) claim that it should be strictly bounded, because outliers are usually extraneous to the models. These two approaches have opposite consequences on the properties of convergence and adaptivity of the estimates, and, in nonparametric smoothing, they were applied to outlier removing and edge-preserving respectively.

Following Huber and Hampel philosophies, the most common unbounded (a,b) and bounded (c,d) loss functions are

$$\begin{aligned}
\text{a)} \quad \rho_a(\varepsilon) &= \Big|\,\varepsilon\,\Big| \\
\text{b)} \quad \rho_b(\varepsilon) &= \begin{cases} \varepsilon^2/2\,, & |\varepsilon| \le \lambda \\ \lambda\,|\varepsilon| - \lambda^2/2\,, & |\varepsilon| > \lambda \end{cases} \\
\text{c)} \quad \rho_c(\varepsilon) &= \begin{cases} \varepsilon^2/2\,, & |\varepsilon| \le \lambda \\ \lambda^2/2\,, & |\varepsilon| > \lambda \end{cases} \\
\text{d)} \quad \rho_d(\varepsilon) &= -L\Big(\varepsilon/\lambda\Big)/\lambda
\end{aligned} \tag{4}$$

where $L(\cdot)$ is a unimodal kernel function and $\lambda > 0$ is a tuning constant. This coefficient is usually designed according to the rate of outlier contamination and, under the assumption of Gaussian disturbances, one can set $\lambda = 2\,\sigma_\varepsilon$. The function (4,a) is the most simple and was stressed by Wang and Scott (1994); (4,b) was defi-

Figure 1. Display of the score functions in (4), with $L(\cdot)$ Gaussian and $\lambda = 1$.



ned by Huber and has monotone derivative: $\psi(\varepsilon) = \partial\,\rho(\varepsilon)/\partial\,\varepsilon$. Finally, (4,c) corresponds to the trimmed method and (4,d) is a smoothed solution which provides *redescending* ψ-functions (e.g. Hampel *et al.*, 1986). Graphical behavior of these functions, and of their transformations, is shown in Figure 1.

Letting $\boldsymbol{\beta} = [\,\beta_0,\, \beta_{10},\, \beta_{01} \ldots \beta_{0p}\,]'$, the vector of parameters to be estimated in (3) at every point (x, y), the minimization typically proceeds by nonlinear methods, such as Steepest-Descent. At the k-th iteration one has the M-algorithm

$$\hat{\boldsymbol{\beta}}_{\mathrm{M}}^{(k+1)}(x, y|t) = \hat{\boldsymbol{\beta}}_{\mathrm{M}}^{(k)}(x, y|t) + \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} K_{ij}(x, y)\, \boldsymbol{w}_{ij}\, \psi\Big[\,\mathrm{f}_{ij|t} - \boldsymbol{w}'_{ij}\, \hat{\boldsymbol{\beta}}_{\mathrm{M}}^{(k)}(x, y|t)\Big] \qquad (5)$$

where $K_{ij}(x, y)$ are the kernel weights (say), $\psi(\cdot) = \rho'(\cdot)$ and the vector $\boldsymbol{w}_{ij} = [\,1, (\dot{x}_i - x),\, (\dot{y}_j - y), (\dot{x}_i - x)^2 \ldots (\dot{y}_j - y)^p\,]'$. The initial value of (5) can be obtained from linear smoothers, as $\hat{\boldsymbol{\beta}}_{\mathrm{M}}^{(0)} = \hat{\boldsymbol{\beta}}_{\mathrm{LPR}}$. In general, the convergence of the algorithm (5) to the global optimal solution is guaranteed only if $\psi(\cdot)$ is monotone (namely, if the underlying loss functions are (4;a,b)); in the other cases, multiple local solutions are possible. On the other hand, redescending ψ-functions have better adaptive properties in the presence of jumps (see Rue *et al.*, 2002).

A gradient-free solution for (3), which may avoid the problems of (5), can be derived from the *weighted-average* form of M-estimates (see Hampel *et al.*, 1986 p.115). This arises by applying the Tukey transformation $\omega(\varepsilon) = \psi(\varepsilon)/\varepsilon$ to the rescaled normal equations of (3), given by

$$\begin{aligned}
&\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} K_{ij}(x, y)\, \psi\Big[(\,\mathrm{f}_{ij|t} - \boldsymbol{w}'_{ij}\, \boldsymbol{\beta})/\sigma_\varepsilon\Big]\, \boldsymbol{w}_{ij} = \mathbf{0} \\
&\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} K_{ij}(x, y)\, \omega\Big[(\,\mathrm{f}_{ij|t} - \boldsymbol{w}'_{ij}\, \boldsymbol{\beta})/\sigma_\varepsilon\Big]\, (\,\mathrm{f}_{ij|t} - \boldsymbol{w}'_{ij}\, \boldsymbol{\beta})\, \boldsymbol{w}_{ij} = \mathbf{0} \\
&\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} K_{ij}(x, y)\, \omega(\varepsilon_{ij}/\sigma_\varepsilon)\, \boldsymbol{w}_{ij}\, \mathrm{f}_{ij|t} = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} K_{ij}(x, y)\, \omega(\varepsilon_{ij}/\sigma_\varepsilon)\, \boldsymbol{w}_{ij}\, \boldsymbol{w}'_{ij}\, \boldsymbol{\beta}
\end{aligned} \qquad (6)$$

Notice from Figure 1, that the ω-functions have the same nature as kernels; in particular, in the case (4,d) one has $\omega(\cdot) = +L(\cdot)$. Thus, defining

$$W_{ij}(x, y; \varepsilon|t) = \frac{1}{\kappa_1 \kappa_2\, \lambda}\, K_1\!\left(\frac{\dot{x}_i - x}{\kappa_1}\right) K_2\!\left(\frac{\dot{y}_j - y}{\kappa_2}\right) L\!\left(\frac{\mathrm{f}_{ij|t} - \boldsymbol{w}'_{ij}\, \boldsymbol{\beta}}{\lambda}\right) \qquad (7)$$

and solving the system (6) for $\boldsymbol{\beta}$, in iterative form, one can obtain the estimator

$$\hat{\boldsymbol{\beta}}_{\mathrm{M}}^{(k+1)}(x,y|t) = \left[\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} W_{ij}\Big(x,y;\hat{\varepsilon}_{ij}^{(k)}|t\Big)\, \boldsymbol{w}_{ij}\, \boldsymbol{w}_{ij}' \right]^{-1} \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} W_{ij}\Big(x,y;\hat{\varepsilon}_{ij}^{(k)}|t\Big)\, \boldsymbol{w}_{ij}\, \mathrm{f}_{ij|t} \tag{8}$$

where $\hat{\varepsilon}_{ij}^{(k)}(x,y|t) = \Big[\, \mathrm{f}_{ij|t} - \boldsymbol{w}_{ij}'\, \hat{\boldsymbol{\beta}}_{\mathrm{M}}^{(k)}(x,y|t)\Big]$ is the ij-th residual evaluated at (x,y). Since (8) resembles the LPR smoother, it can be termed *quasi-linear*.

The estimate of the intensity function $f(x,y|t)$ is provided by the first element of the vector (8). In the Appendix we directly derive its expression for the simpler case $p=0$; by using the notation $L_\lambda(\cdot) = L(\cdot/\lambda)/\lambda$, it is given by

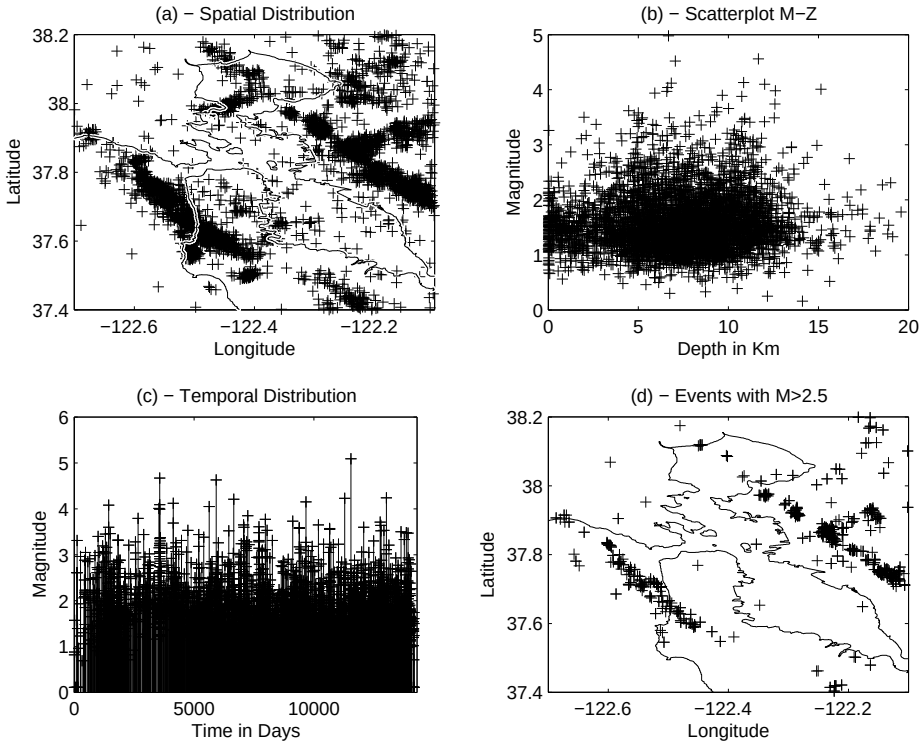
$$\hat{f}_{\mathrm{M},0}^{(k+1)}(x,y|t) = \frac{\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} K_{\kappa_1}(\dot{x}_i - x)\, K_{\kappa_2}(\dot{y}_j - y)\, L_\lambda\Big[\, \mathrm{f}_{ij|t} - \hat{f}_{\mathrm{M},0}^{(k)}(x,y|t)\,\Big]\, \mathrm{f}_{ij|t}}{\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} K_{\kappa_1}(\dot{x}_i - x)\, K_{\kappa_2}(\dot{y}_j - y)\, L_\lambda\Big[\, \mathrm{f}_{ij|t} - \hat{f}_{\mathrm{M},0}^{(k)}(x,y|t)\,\Big]} \tag{9}$$

This looks like a Nadaraya-Watson smoother, but is iterative and also weights observations in the direction of the dependent variable. With respect to common nonlinear algorithms, the quasi-linear methods (8)-(9) are much faster and avoid the use of ψ-functions; hence, they can reduce the risk of convergence to local minima. Another important feature of (8)-(9) is that they are easily implementable in *recursive* form; this point will be developed in the next section.

The robust mechanism underlying (8) has some connection with those developed by Cleveland (1979), Fan *et al.* (1994) and Assaid *et al.* (2000). These authors focused on the Huber's function $\psi_{\mathrm{H}}(\varepsilon) = \min[\lambda, \max(-\lambda, \varepsilon)]$, and distinguish themselves mainly for the weights $W_{ij}(\cdot)$. In Cleveland (1979) the first iteration is as follows: the LPR residuals $\hat{\varepsilon}_i$ are rescaled by means of a robust estimate of σ_ε; next they are transformed into weights by means of ψ_{H} and are multiplied by tricube weights K_i based on a nearest-neighbor bandwidth. Similarly, Fan *et al.* (1994) use the weights $K_i\, \psi_{\mathrm{H}}(\hat{\varepsilon}_i^*)/\hat{\varepsilon}_i^*$, where the star denotes rescaling, and Assaid *et al.* (2000) focused on K_i Gaussian designed with a modified cross-validation criterion. Instead, in our approach, the algorithms (8)-(9) have been obtained with the weighted-average form of M-estimates discussed in Hampel *et al.* (1986, p.115). As a result, their general feature is weighting residuals with kernel functions instead of the usual scores (see Figure 1 to appreciate the difference).

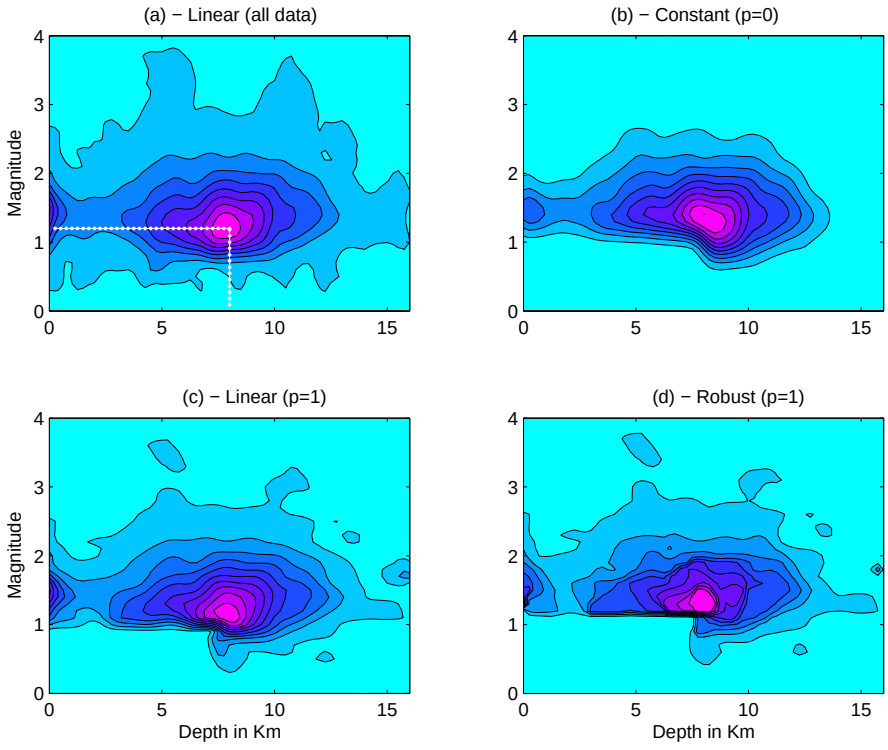
Applications. We now illustrate the methodology we have discussed on a real case-study concerning the earthquakes of the San Francisco bay. The data-set was downloaded from the Internet site of Northern California Earthquake Data Center, and covers a zone with longitude $-122.7 \leq x \leq -122.1$, latitude $37.4 \leq y \leq 38.2$, and time span Jan $1968 \leq t \leq$ Dec 2005 (the starting time was that available in the data-base). The total number of events is $N = 4369$ and their features are displayed in Figure 2. Panels (a,d) show that spatial pattern of events is not random and follows two nearly parallel stems of the St Andrea fault. Panel (c) shows that the temporal distribution is nearly uniform because the time interval is small and no outlying event occured; finally, Panel (b) shows that there is no linear relationship between magnitude of the events and their depth.

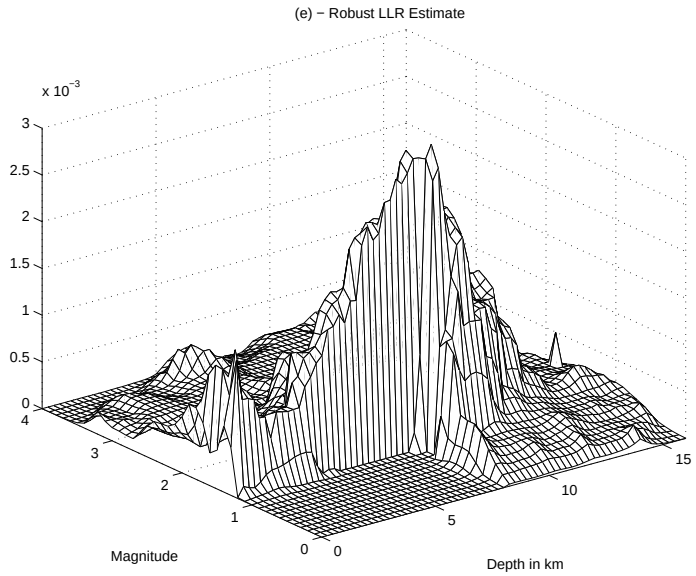
Figure 2. Main features of the Northern California earthquake catalog in the San Francisco bay in the period Jan 1968-Dec 2005: (a) Spatial distribution of epicenters; (b) Scatterplot between depth and magnitude; (c) Temporal distributions of events; (d) Spatial distribution of earthquakes with magnitude greater than 2.5. Data were downloaded from the Internet site `http://www.ncdec.org`.



To test the jump-preserving ability of the M-smoother (8), we have considered the density estimation of the data in Figure 2(b) (which regards magnitude and depth), by dropping the part of the histogram f_{ij} below the modal values of m, z (see Figure 3(a)). The size of the squared bins of the histogram was established on the basis of rules-of-thumb present in the literature (such as $2\,\hat{\sigma}_{x,y}/\sqrt{N}$), and led to a grid 40×60. The performance of smoothers (8)-(9) crucially depends on the values of the bandwidths $\kappa_{1,2}, \lambda$. A popular selection strategy consists of using quadratic cross-validation (CV), or its robust version based on absolute prediction errors (see Wang and Scott, 1994). We shall discuss in detail this problem in the next section; by now, we only state that constrained CV provided $(\kappa_1{=}\kappa_2{=}\lambda){=}1.9$. Resulting density estimates are displayed in Figure 3; where Panel (a) shows the local linear regression (LLR, $p{=}1$) on all data, Panels (b,c,d) show kernel regression

Figure 3. Regression estimates of the density of data in Figure 2(b): (a) LLR estimate on complete data; (b) Kernel regression on partial data; (c) LLR estimate on partial data; (d,e) Robust LLR estimate on partial data. All smoothers were designed with Gaussian kernels and the common bandwidth 1.9.





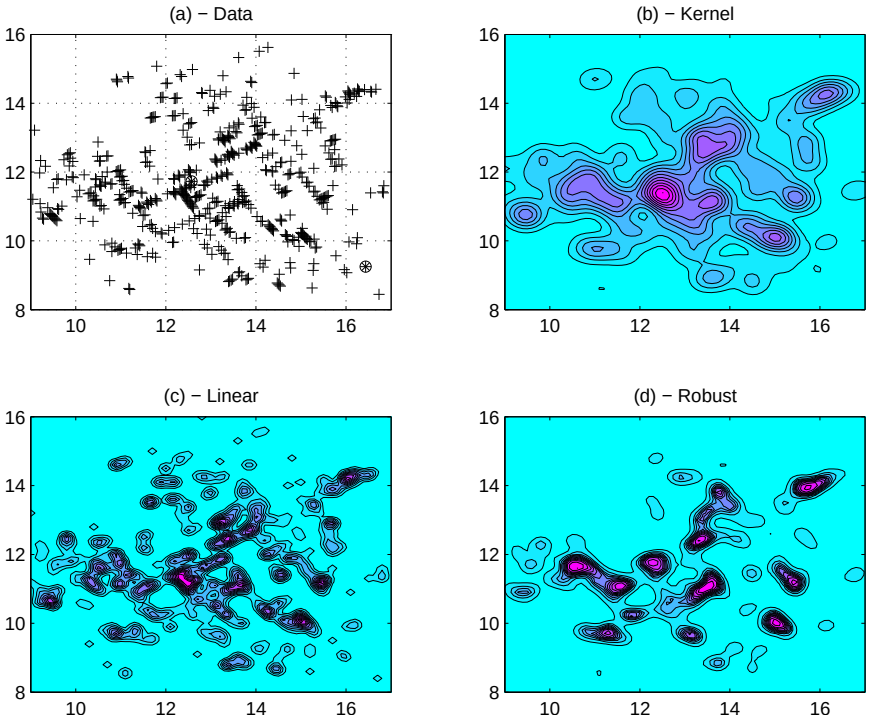
(i.e. p=0), LLR and robust LLR estimates on partial data. Clearly, the robust method is the best one to preserve the introduced jump (a 3D-view is provided in Panel (e)), although it has greater variability in smooth regions.

In addition to edge-preserving, we can also show the ability of robust smoothers to cluster spatial data. We consider the data-set of the epidemiologist John Snow, who studied the 1854 cholera outbreak in the Soho zone of London. In that time, causes of the infection were unknown; however, by mapping the death occurrences, he discovered that the epidemic originated from few public water-pumps. Point pattern is displayed in Figure 4(a), and the other panels show nonparametric estimates: kernel density, LLR, robust LLR. Smoothing coefficients were selected with mixed criteria and are described in the caption of Figure 4. The most relevant feature is the capability of the M-smoother to cluster data in few groups, whereas the other methods provide dispersive information.

3. Bandwidth Selection and Recursive Estimation

Bandwidth design is a fundamental problem for nonparametric estimators. The ideal approach is *plug-in*, where the theoretical MSE of the smoother is analyzed, the expression of the optimal bandwidth is derived and the unknown quantities (such

Figure 4. Smoothing epidemic data of John Snow (`www.ph.ucla.edu/epi/snow.html`): (a) Distribution of deaths, N=578; (b) Kernel density with $\hat{\kappa}_{1,2} = \hat{\sigma}_{x,y}/N^{1/5} = 0.52$; (c) Local linear regression with $\hat{\kappa}_{\mathrm{CV}} = 0.78$; (d) Robust LLR with $\hat{\lambda} = 2\,\hat{\sigma}_{\mathrm{M}} = 0.11$. All smoothers used Gaussian kernels and a grid 55×50 .



as the derivatives f'') are replaced by pilot estimates. However, this approach is analytically and computationally demanding and, in the case of discontinuous functions and complex coefficients (such as λ), it cannot be applied. Under regularity conditions, the CV method provides asymptotically optimal results (e.g. Härdle *et al.*, 2002 p.114), and is always implementable.

Cross-Validation. Following Wang and Scott (1994) and Leung *et al.* (1993, 2005) outliers and jumps produce significant bias in the selected bandwidths; to reduce their influence, one can use robust cross-validation (RCV). For the model (2) and the estimator (9), it consists of minimizing

$$P_n(\kappa_{1,2}, \lambda) = \frac{1}{n_1\, n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \varrho\Big[\, \mathrm{f}_{ij|t} - \hat{f}^{(k)}_{\mathrm{M}-ij}(\dot{x}_i, \dot{y}_j|t) \,\Big] \tag{10}$$

where $\varrho[\,\cdot\,]$ may be one of the ρ-functions in (4) and $\hat{f}^{(k)}_{\mathrm{M}-ij}(\cdot)$ are M-estimates obtained

by omitting the ij-th frequency. The criterion (10) leads to the optimal selection of the bandwidths which regard the local adaptation. In fact, Leung (2005) has shown that $\hat{\kappa}_{\mathrm{RCV}}$ converges in probability to κ_{opt} (which minimizes the integrated MSE), and is asymptotically independent of $\varrho[\cdot]$. This result was established for continuous models with outliers and for M-smoothers of Huber type, but it can be extended to more complex situations.

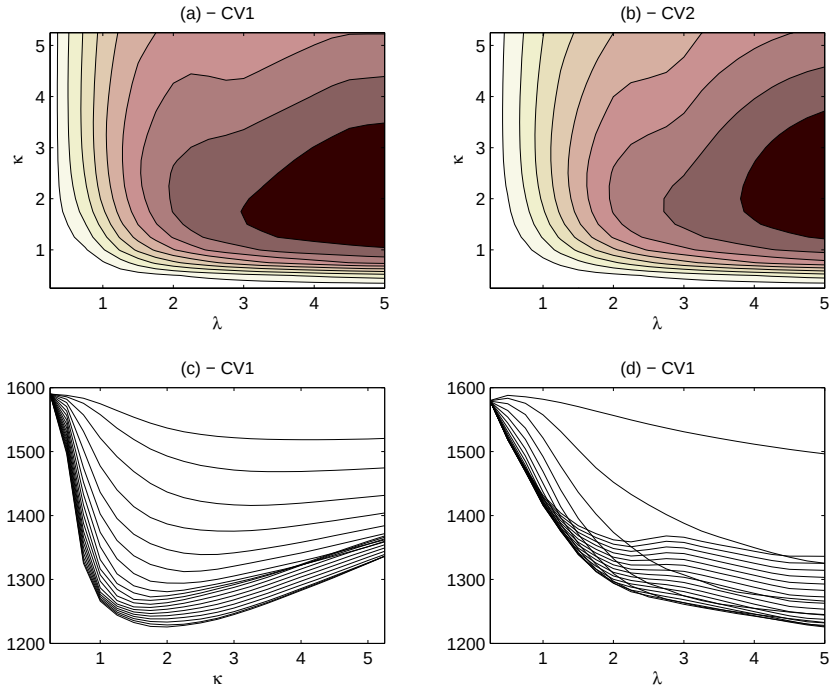
Two remarks are now necessary:

(i) The preferred ϱ-function in (10) is the absolute criterion (4,a), since it does not involve additional robustness coefficients, say λ_{ϱ}. In the other loss functions (4;b-d), this coefficient would create a circular problem with respect to the estimation of the bandwidth λ with (10). In any event, Wang and Scott (1994) and Leung (2005) have shown that the function $\varrho = |\cdot|$ has a sufficient degree of robustness for the selection of κ.

(ii) In the literature not much has been said about the selection of λ. In M-smoothers with Huber ψ-function, that coefficient is usually established a-priori (e.g. Leung, 2005), and the authors who tried to estimate it with quadratic CV obtained $\hat{\lambda}_{\mathrm{CV}} \to \infty$ (see Hall and Jones, 1990 p.1717). In this case, it is necessary to constrain its value to κ, or to adopt the heuristic design $\lambda = 2\,\sigma_{\varepsilon}$. Under Gaussianity, this solution enables 95% asymptotic relative efficiency (ARE) with respect to nonrobust smoothers.

Figure 5 shows the path of the absolute RCV-function applied to the partial data in Figure 3(a). With respect to the first bandwidth, it has a well-defined minimum at $\hat{\kappa} = 2$, whereas for the other it confirms $\hat{\lambda}_{\mathrm{RCV}} \to \infty$. This was explained by Hall and Jones (1990, p.1717) with the "preference of M-smoothers for the mean fit", but is due to the fact that $\lambda < \infty$ reduces the efficiency (increases the variability) of M-estimates in smooth regions. Constrained selection provided $(\hat{\kappa} = \hat{\lambda}) = 1.9$, and the median absolute deviation (MAD) of the residuals of Figure 3(c) gave $\hat{\sigma}_{\mathrm{M}}$=0.75 (this yields a slightly smaller value of $\lambda = 2\sigma_{\varepsilon}$). The results in Figure 3(d) were generated with these values.

Figure 5. Path of CV functions for the coefficients κ, λ of the smoother (8) (with p=1 and K, L Gaussian), applied to the partial data in Figure 3(a): (a,c,d) Absolute criterion (CV1); (b) Quadratic criterion (CV2).

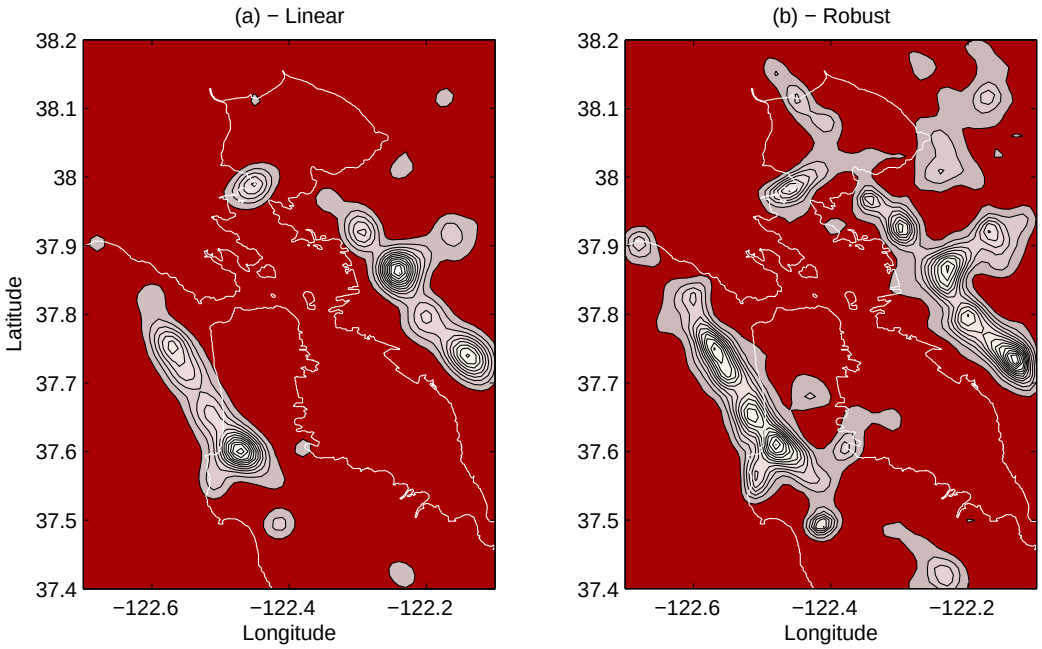


We now present the estimates of the intensity function (2) obtained on the whole data-set: $f(x, y|t_N)$. This exercise is common practice in seismology (e.g. Stock and Smith, 2002), where the time-dimension is treated separately. The area of study was partitioned in a grid of size 80×60 and the frequencies f_{ij} were computed by weighting the events with their relative magnitude. We compare the local linear regression with its robust version; quadratic and absolute CV provide

$$\begin{aligned} \text{CV2} \quad &: \quad \text{LLR}(\hat{\kappa}_1 = 1.43,\, \hat{\kappa}_2 = 1.21),\, \hat{\sigma}_\varepsilon = 7.3\,; \quad \text{M–LLR}(\hat{\kappa}_{1,2} = \hat{\lambda} = 1.32) \\ \text{CV1} \quad &: \quad \text{LLR}(\hat{\kappa}_1 = 1.19,\, \hat{\kappa}_2 = 0.92),\, \hat{\sigma}_\text{M} = 2.1; \quad \text{M–LLR}(\hat{\kappa}_{1,2} = \hat{\lambda} = 0.95) \end{aligned} \tag{11}$$

in the linear case, the bandwidths $\hat{\kappa}_{1,2}$ converge toward similar values, but in the robust one we still have $\hat{\lambda} \to \infty$. The constrained estimation ($\lambda = \kappa$) in equation (11) gives admissible values, even compared to the heuristic solution $\lambda = 2\sigma_\varepsilon$. With these bandwidths, we have generated the intensity functions in Figure 6; as for Figure 4, the robust method gives more detailed spatial information.

Figure 6. Intensity function of earthquakes on the entire period: (a) LLR estimate; (b) Robust LLR. All smoothers used Gaussian kernels and bandwidths in (11).



Recursive Estimation. So far we have considered M-smoothers which are not weighted in the time-dimension, or that just treat it sequentially, as (3) and (8). In these estimators, when the variable t changes, the entire past information must be reprocessed. If observations would be equally spaced, as in standard time-series, then a recursive implementation would be possible. The condition t-discrete can be achieved by binning data in weekly, monthly or yearly series. Thus, let $t = 1, 2, \ldots n_3$, and define the 3D-matrix $\boldsymbol{F} = \{\mathrm{f}_{ijt}\}$ of frequencies

$$\mathrm{f}_{ijt} = \sum\nolimits_{\{\,k\,:\,(t-1<\mathrm{t}_k\le t)\cap(x_{i-1}<\mathrm{x}_k\le x_i)\cap(y_{j-1}<\mathrm{y}_k\le y_j)\,\}} \frac{\mathrm{m}_k}{\bar{m}_N}$$

In this context, the estimator (3) can be easily weighted in time as (1)

$$\begin{aligned}\hat{f}_{\mathrm{M}}(x,y,t) \;&=\; \arg\min_{\beta_0}\left\{\sum_{h=1}^{t}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\mu^{t-h}\,K_1\!\left(\frac{\dot{x}_i-x}{\kappa_1}\right)K_2\!\left(\frac{\dot{y}_j-y}{\kappa_2}\right)\times\right.\\ &\times\;\left.\rho\Big[\mathrm{f}_{ijk}-\beta_0-\beta_1(\dot{x}_i-x)-\beta_2(\dot{y}_j-y)-\beta_3(t-h)\Big]\right\}\end{aligned}$$

where $\mu \in (0, 1]$ and we have assumed $p = 1$. Also, the corresponding quasi-linear estimator (8) can be obtained by multiplying the weights (7) by μ^{t-h}.

Now, the recursive version of the robust smoother can be derived by equating number of iterations and number of processed layers, i.e. $k = t$. Under this condition, the denominator ($\boldsymbol{R}$) and the numerator ($\boldsymbol{s}$) of (8) can be updated recursively, and the estimator is derived accordingly (see Grillenzoni, 2000)

$$\begin{aligned}
\hat{\varepsilon}_{ij}(x,y|t) &= \mathrm{f}_{ijt} - \boldsymbol{w}'_{ij}\,\hat{\boldsymbol{\beta}}_{\mathrm{M}}(x,y|t-1) \qquad (12)\\
\boldsymbol{R}(x,y|t) &= \mu\cdot\boldsymbol{R}(x,y|t-1) + \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} W_{ij}\Big(x,y;\hat{\varepsilon}_{ij}|t\Big)\,\boldsymbol{w}_{ij}\,\boldsymbol{w}'_{ij}\\
\boldsymbol{s}(x,y|t) &= \mu\cdot\boldsymbol{s}(x,y|t-1) + \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} W_{ij}\Big(x,y;\hat{\varepsilon}_{ij}|t\Big)\,\boldsymbol{w}_{ij}\,\mathrm{f}_{ijt}\\
\hat{\boldsymbol{\beta}}_{\mathrm{M}}(x,y|t) &= \boldsymbol{R}(x,y|t)^{-1}\boldsymbol{s}(x,y|t)\,, \qquad t=2,3\ldots n_3 \qquad (13)
\end{aligned}$$

where (12) is the one-step-ahead prediction error and the first element of (13) provides the intensity function. In this setting, it is clear the role of $\mu < 1$ to discount past information and to adapt the smoother to changing situations. Time-evolution may also be related to the intrinsic nonlinearity of the point process.

The smoothing coefficients of (13), can be designed on the basis of the prediction errors; in particular, a robust CV criterion is given by the sum of $|\,\hat{\varepsilon}_{ij}(x,y|t)|$. We have applied this method to yearly seismic data in the period 1968-2005, and distributed on a spatial grid of size 40×30. The predictive CV selection has provided $\hat{\mu} \to 0$, which means that the process has no memory, at least on yearly data. Increasing the temporal aggregation would increase the auto-correlation inside $\boldsymbol{F}$, but the resulting estimates would have limited predictive value. In any case, the path of the conditional CV function

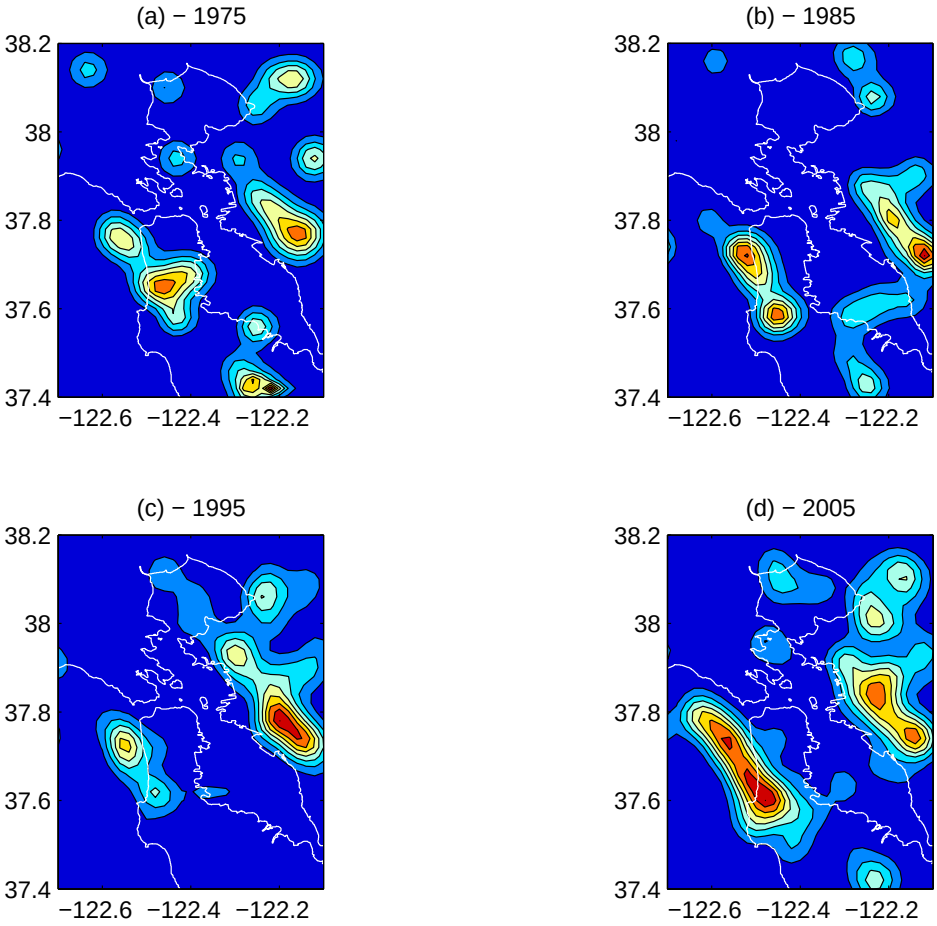
$$P_n(\mu|\hat{\lambda},\hat{\kappa}) = \sum_{t=2}^{38}\sum_{i=1}^{40}\sum_{j=1}^{30}\Big|\,\hat{\varepsilon}_{ij}(\dot{x}_i,\dot{y}_j|t)\Big|$$

given the bandwidths in (11), becomes nearly stable for $\mu \leq 0.5$.

Figure 7 displays recursive estimates (13) generated with $\mu = 0.5$ in the years t =1975, 1985, 1995, 2005. They show a significant variability and differences with respect to Figure 6, which provides the (static) total value in the period. These differences are yielded by the low value of μ, by the 10-year lag between the displayed frames and by the intrinsic randomness of seismic phenomena on the short period

(geologically speaking). However, the proposed method provides updated maps of risk which can be useful for point processes which have fast dynamics and must be monitored in real-time. Whatever the application is, the user must select the degree of time-aggregation and the value of the discounting factor with particular care, since they determine the goodness of the one-step-ahead forecast $\hat{f}(x, y|t)$.

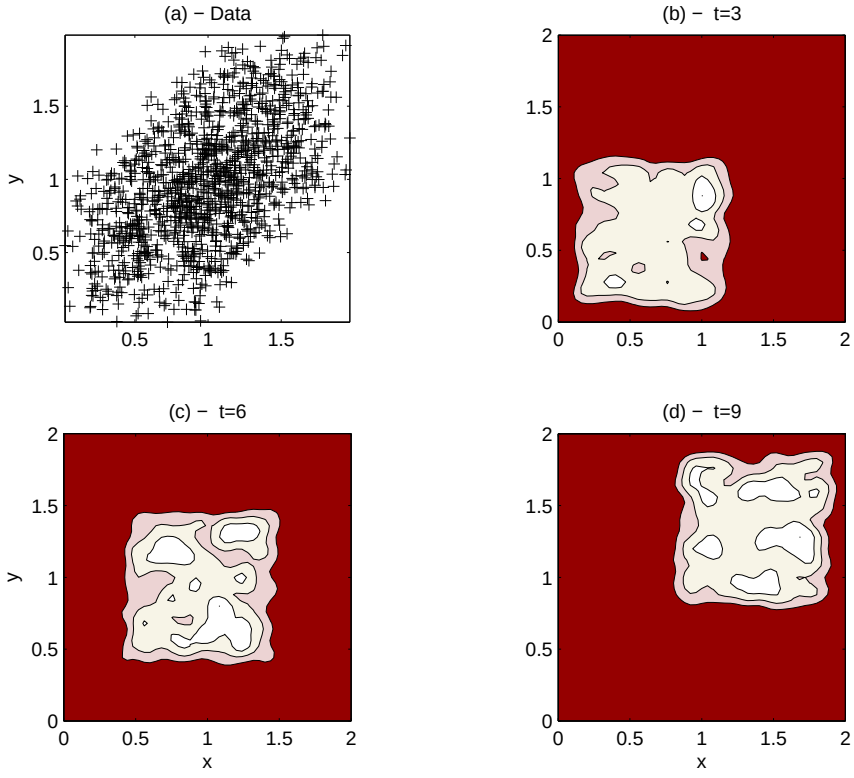
Figure 7. Robust recursive estimates of the conditional intensity of earthquakes in the years: (a) 1975; (b) 1985; (c) 1995; (d) 2005. The algorithm (13) was implemented with $p = 1$, Gaussian kernels, $\mu = 0.5$ and the bandwidths (11).



Simulation. Let us conclude with a small simulation experiment on mobile discontinuous functions. We generate a space-time point process whose basic densities are uniform on the unit interval. Specifically, given independent realizations

of $t_i \sim \mathrm{U}(0,1)$ and $(u_i, v_i) \sim \mathrm{U}_2(0,1)$ with $i = 1 \ldots 1500$; the time variable is sorted as $t_i \leq t_{i+1}$, and the spatial coordinates are generated as $(x_i, y_i) = (u_i, v_i) + t_i$. The intensity function of the process looks like a bivariate uniform density that moves in the North-East direction. It may represent a physical or a chemical process and a random realization is displayed in Figure 8(a). The experiment consists of estimating the bivariate uniform density at the points $t = 0.1, 0.2 \ldots 1$, with the recursive estimator (13). To this purpose, data were binned in a $50 \times 50 \times 10$ regular grid. Using cross-validation and rules of thumb, the smoothing coefficients were selected as $\kappa_{1,2} = \lambda = 0.05$ and $\mu = 0.5$. The results are displayed in Figure 8(b-c) for the instants $t = 0.3, 0.6, 0.9$. Since uniform densities are difficult to estimate even in the static case, the performance of (13) is sufficiently good.

Figure 8. Robust recursive estimation of a moving uniform density. (a) Data sample; Estimates at: (b) t=0.3; (c) t=0.6; (d) t=0.9. The algorithm (13) was implemented with $p = 1$, Gaussian kernels and coefficients $\mu = 0.5$, $\kappa = \lambda$=0.05.



4. Conclusions

This article has discussed nonparametric estimation of the conditional intensity of earthquake data. It provides the expected rate of occurrence at any point of the space, given the set of information available up to a given instant. The strategy of estimation consists of smoothing the observed frequency histogram with the method of local polynomial regression. This approach enables automatic boundary corrections and its jump-preserving ability can be improved with robustness. An iterative algorithm is derived from the weighted-average form of M-estimates and its smoothing coefficients are designed with mixed criteria. Finally, a recursive algorithm is proposed for sequential processing of the data binned in time.

A delicate aspect of the method is represented by the selection of the smoothing coefficients. For robust-space-time smoothers we have three kinds of coefficients: the bandwidths κ, the robustness parameter λ and the discounting factor μ. We have designed them with mixed strategies based on cross-validation, relative efficiency and prediction errors criteria, which are spread in the literature. In general, these require the optimization of loss functions, which are difficult to treat even when they are convex and differentiable. An alternative selection strategy for κ, λ, μ could be obtained from the *SiZer* approach of Chaudhuri and Marron (1999, 2000). This explores the behavior of kernel estimates in the space of smoothing coefficients and searches for significant features through curve derivatives. The method is appealing, but is difficult to apply to multivariate nonlinear smoothers.

Wide part of the paper has been devoted to an application to Northern California earthquake catalog in the area of San Francisco. We have shown that robust smoothers have good ability to estimate discontinuous density functions. Further, with respect to linear kernel methods they have greater sensitivity to capture spatial details and edges. Also, the proposed method has worked well at the recursive level, providing maps of seismic risk which evolve over time. This solution is useful in situations (such as monitoring air pollution) which evolve fastly and need real-time processing of data.

Appendix: derivation of (9)

For the robust smoother with p=0, the loss function corresponding to (3) is $R_n(f) = \sum_i \sum_j K_{ij}(x,y)\,\rho\Big(\mathrm{f}_{ij|t} - f\Big)$; and the normal equation corresponding to (6) becomes $R'_n(f) = \sum_i \sum_j K_{ij}(x,y)\,\psi\Big(\mathrm{f}_{ij|t} - f\Big) = 0$. Now, inserting the Tukey transformation $\omega(\varepsilon) = \psi(\varepsilon)/\varepsilon$ in the normal equation, we have

$$\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} K_{ij}(x,y)\,\omega(\mathrm{f}_{ij|t} - f)\,\mathrm{f}_{ij|t} = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} K_{ij}(x,y)\,\omega(\mathrm{f}_{ij|t} - f)\,f$$

and solving for f, in iterative form, provides the robust smoother

$$\begin{aligned} \hat{f}_{\mathrm{M}}^{(k+1)}(x,y) &= \left[\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} K_{ij}(x,y)\,\omega\Big(\mathrm{f}_{ij|t} - \hat{f}_{\mathrm{M}}^{(k)}(x,y)\Big)\right]^{-1} \\ &\times \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} K_{ij}(x,y)\,\omega\Big(\mathrm{f}_{ij|t} - \hat{f}_{\mathrm{M}}^{(k)}(x,y)\Big)\,\mathrm{f}_{ij|t} \end{aligned} \tag{14}$$

Now, in the case of the loss function (4,d), with $L(\cdot)$ Gaussian, one has

$$\psi\Big(\mathrm{f}_{ij|t} - f\Big) = \frac{-1}{\sqrt{2\pi}\lambda}\exp\left[-\frac{1}{2}\Big(\frac{\mathrm{f}_{ij|t} - f}{\lambda}\Big)^2\right]\frac{-1}{\lambda^2}\Big(\mathrm{f}_{ij|t} - f\Big)$$

that is $\omega(\cdot) \propto L(\cdot)$ (see Figure 1), and the smoother (9) directly follows from (14).

References

Assaid C.A. and Birch J.B. (2000), Automatic Bandwidth Selection in Robust Nonparametric Regression. *Journal of Statistical Computation and Simulations*, **66**, 259-272.

Bailey T.C. and Gatrell A.C. (1995), *Interactive Spatial Data Analysis*. Longman: Essex (UK).

Bouezmarni T. and Scaillet O. (2005), Consistency of Asymmetric Kernel Density Estimators and Smoothed Histograms with Application to Income Data. *Econometric Theory*, **21**, 390-412.

Chaudhuri P. and Marron J.S. (1999), SiZer for Exploration of Structures in Curves. *Jour. of Americ. Stat. Assoc.*, **94**, 807-823.

Chaudhuri P. and Marron J.S. (1999), Scale Space View of Curve Estimation. *Annals of Stat.*, **28**, 408-428.

Cleveland W.S. (1979), Robust Locally Weighted Regression and Smoothing Scatterplots. *Jour. of Americ. Stat. Assoc.*, **74**, 829-836.

Cheng M.-Y., Fan J. and Marron J.S. (1997), On Automatic Boundary Corrections. *The Annals of Statistics*, **25**, 1691-1708.

Choi E. and Hall P. (1999), Nonparametric Approach to the Analysis of Space-Time Data on Earthquake Occurrences. *Journal of Computational and Graphical Statistics*, **8**, 733-748.

Choi E. and Hall P. (2000), On the Estimation of Poles in Intensity Functions. *Biometrika*, **87**, 251-263.

Chu C.K., Glad I., Godtliebsen F. and Marron J.S. (1998), Edge-Preserving Smoothers for Image Processing. *Journal of the American Statistical Association*, **93**, 526-541.

Daley D.A., and Vere-Jones, D. (2003). *An Introduction to the Theory of Point Processes, Vol.1.* Springer: New York.

Fan J. and Gijbels I. (1996), *Local Polynomial Modelling and its Applications.* Chapman & Hall, London.

Fan J., Hu C.T. and Troung Y.K. (1994), Robust Nonparametric Function Estimation. *Scandinavian Journal of Statistics*, **21**, 433-446.

Grillenzoni, C. (2000), Nonparametric Regression for Nonstationary Processes. *Journal of Nonparametric Statistics*, **12**, 265-282.

Grillenzoni C. (2005), Nonparametric Smoothing of Spatio-Temporal Point Processes. *Journal of Statistical Planning and Inference*, **128**, 61-78.

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. *The 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`

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.

Simonoff J.S. (1996), *Smoothing Methods in Statistics*. Springer-Verlag: New York.

Stock C. and Smith E. (2002a), Adaptive Kernel Estimation and Continuous Probability Representation of Historical Earthquake Catalogs. *Bulletin of Seismological Society of America*, **92**, 901-912.

Stock C. and Smith E. (2002b), Comparison between Seismicity Models Generated by Different Kernel Estimations. *Bulletin of Seismological Society of America*, **92**, 913-922.

Vere-Jones D. (1992), Statistical Methods for the Description and Display of Earthquake Catalogs. In: A. Walden and P. Guttorp (eds.), *Statistics in the Environmental and Earth Sciences*, pp. 220-246. Arnold: London, 1992.

Wang F. and Scott D. (1994), The L_1 Method for Robust Nonparametric Regression. *Journal of the American Statistical Association*, **89**, 65-76.

Zhuang J., Ogata Y. and Vere-Jones D. (2002), Stochastic declustering of spacetime earthquake occurrences. *Jour. Americ. Stat. Assoc.*, **97**, 369-380.