Interactive local bandwidth choice

J. S. Marron* and F. Udina†

27 February, 1995

Abstract

A tool for user choice of the local bandwidth function for a kernel density estimate is developed using KDE, a graphical object-oriented package for interactive kernel density estimation written in LISP-STAT. The bandwidth function is a cubic spline, whose knots are manipulated by the user in one window, while the resulting estimate appears in another window. A real data illustration of this method raises concerns, because an extremely large family of estimates is available.

1 Introduction

Smoothing methods are useful for gaining insights from data. See Eubank (1988) [4], Härdle (1990) [8], Müller (1988) [13], Scott (1992) [17], Silverman (1986) [19] and Wahba (1990) [22] for many interesting examples. In a number of cases, it is desirable to use different amounts of smoothing in different locations. Here we study location dependent smoothing in the context of kernel density estimation, although our ideas extend easily to other settings. The smoothing parameter of a kernel density estimator is often called the

*CSIRO Division of Mathematics and Statistics, Locked Bag 17 North Ryde, NSW 2113 Australia (on leave from the University of North Carolina). *Internet e-mail:* `marron@stat.unc.edu`

†Departament d'Economía, Universitat Pompeu Fabra, Balmes 132, 08008 Barcelona, Spain. *Internet e-mail:* `udina@upf.es`

bandwidth. Basic concepts of the kernel density estimator are discussed in Section 2.

A useful way to choose the bandwidth for a given data set is by an interactive trial and error process. In particular, substantial insight comes from being able to choose a bandwidth after looking at the estimate corresponding to a previous choice. In modern computer interfaces this kind of interaction is typically done using a mouse or similar mechanism by means of a graphical device called a *slider*. With this graphical metaphor, the user drags the thumb nail of the slider by moving the mouse to choose from a range of values. The user can also click on the arrows at each side of the slider to increment or decrement the bandwidth by a fixed factor. Then the corresponding estimate is shown and the user can react by choosing a new value.

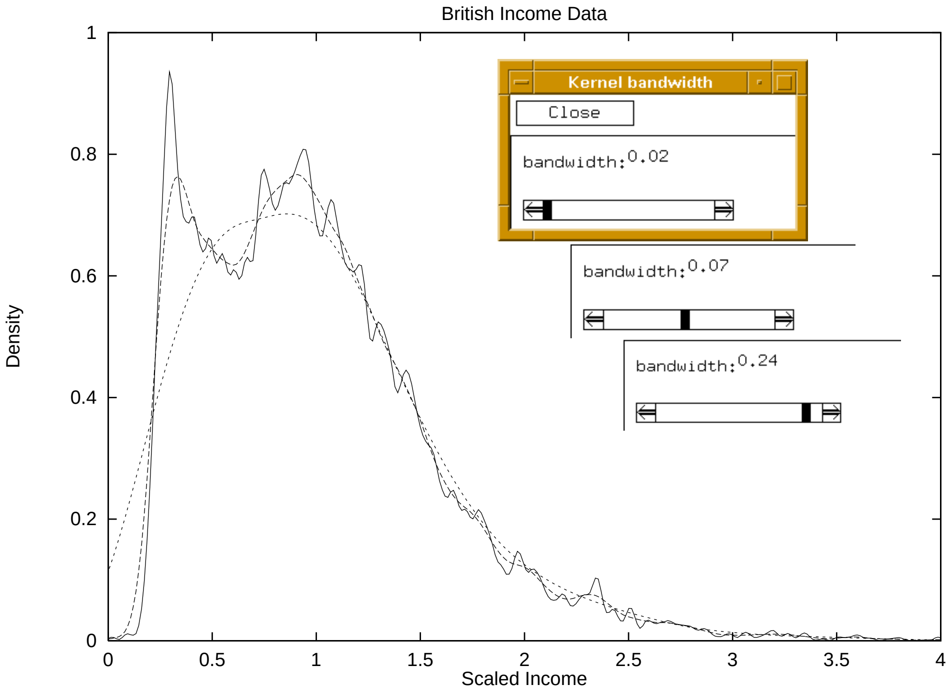

FIGURE 1: *Biweight kernel density estimators for the income data. Global bandwidth estimators, with bandwidths* $h = 0.02$, 0.07, 0.24. *The slider used to choose values is also shown.*

The programming environment LISP-STAT, Tierney (1990) [20], provides a convenient environment for use of such tools in a statistical application.

We have used LISP-STAT to develop KDE, a graphical package that allows convenient interactive density estimation by implementing sliders and other graphical mechanisms together with fast methods to compute and draw the estimates. The sliders shown in Figure 1 are real KDE sliders, and the curves shown in the figure are computed by KDE for each of the bandwidth values.

An interactive analysis of a data set often results in a desire to use different amounts of smoothing in different locations. A data set of this type is the income data, as shown in Figure 1, discussed in Schmitz and Marron (1992) [16]. This is a set of $n = 7201$ observations of family income, for the year 1975, in the United Kingdom. The data have been scaled by dividing by their mean. Figure 1 shows an overlay of three kernel density estimates, using the Gaussian kernel with different bandwidths. Note each of the three bandwidths reveals different interesting structure. The undersmoothed curve reveals a tight spike in the data at low incomes (shown to be due to "pensioners" in the population in Schmitz and Marron (1992) [16]). The medium amount of smoothing recovers the second wider peak, and the larger bandwidth gives a smoother tail. An estimator which uses different amounts of smoothing in different locations could mimic each of these estimates where they perform well.

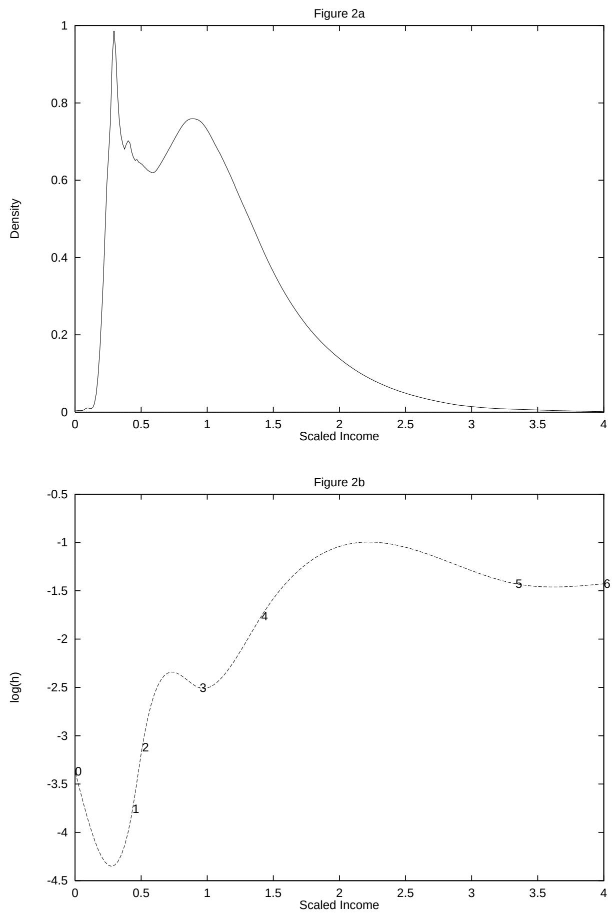

FIGURE 2: *Biweight kernel density estimators for the income data. Location adaptive estimator in (a). Bandwidth function for (a) shown in (b).*

Figure 2a shows a location varying bandwidth kernel estimator for the income data. The upper part of the figure shows a kernel density estimate which uses the bandwidth function shown in the lower part. In this paper we suggest making choice of this bandwidth function interactive. Our user interface to the smoothing parameter function is a cubic spline. The function is controlled by the user through manipulation of the knots of the spline, shown as integers in Figure 2b. Knot locations are changed through a mouse “click-and-drag” operation in the “bandwidth window” while the resulting density estimate is recomputed and redrawn in the upper window.

The implementation of this idea shown here is built with `kde objects`, using LISP-STAT. See, for example, the estimator in Figure 2a, with its local bandwidth function in Figure 2b. Details of the implementation are given in Section 5.

A detailed illustration of our interactive local bandwidth method, applied to a particular data set, is given in Section 3, where a number of different local bandwidth estimates are considered for a single data set. An important lesson is that the number of possible estimates available is extremely large. Also it is very easy to misrepresent the data. In particular important features, such as modes, can be moved around, as well as added or removed from the estimate, nearly at will. Hence interactive local smoothing parameter choice needs very careful use in real applications.

In Section 4 we investigate the set of all estimates that are available from consideration of arbitrary local bandwidth functions.

We do not study data based local bandwidth selection, but our results suggest care needs to be taken in that area. In particular, it seems quite easy to arrive at an uninformative or even misleading result. There is clear indication that precautions against this are needed. This could be accomplished by plotting bandwidth functions as we do here, although other graphical displays could be effective as well.

2 Kernel density estimation

Given a set of data $X_1, ..., X_n$, the kernel density estimate is

$$\widehat{f}_h(x) = n^{-1} \sum_{i=1}^{n} K_h(x - X_i), \tag{1}$$

where h is the bandwidth, and where $K_h(\cdot) = \left(\frac{1}{h}\right) K(\frac{\cdot}{h})$, for a "kernel function" K, which is often taken to be a symmetric probability density. Note that h controls the "width" of the kernel function (e.g. it is the standard deviation if K is the standard normal density). The curves in figure 1 show this estimate, for 3 different values of h. See Silverman (1986) [19] and Scott (1992) [17] for useful discussion of many aspects of this estimator.

Many approaches to location varying kernel density estimation have been suggested, including the nearest neighbor methods of Loftsgaarden and Quesenberry (1965) [12] and others, the variable kernel methods of Breiman, Meisel and Purcell (1977) [2] and Abramson (1982) [1], the shifted kernel method of Samiuddin and El-Sayyad (1990) [15] and the transformation method of Wand, Marron and Ruppert (1991) [23]. For discussion of the many possibilities, and overview of the field, see Chapter 5 of Silverman (1986) [19], Jones (1990) [9] and Jones, McKay and Hu (1994) [11].

Here two main types of bandwidth variation are studied. The first is "depending on location x", where h in (1) is replaced by a function $h(x)$. This includes the nearest neighbor estimators. Insight into this approach comes from focusing on each point x, and viewing the kernel estimate as based on the number of data points in a "nearby window". The width of that window changes with location x.

The second type of bandwidth variation studied here is "depending on data values X_i", where h in (1) becomes $h(X_i)$. This is usefully understood by thinking of the kernel estimate as being the "sum of small probability masses", as illustrated in figures 2.4 and 2.5, page 14, of Silverman (1986). This type of estimate achieves different amounts of smoothing in different locations by allowing these masses to have different widths.

An important difference between the two types of estimates is that a kernel density estimate with bandwidth depending on data values is constrained to have total area under the curve equal to one. No such restriction applies when the bandwidth depends on x, and as seen in Section 4, the area can be any positive value.

3 An illustration

The "Chrondite data", were made famous in the "bump hunting" literature by Good and Gaskins (1980) [7]. Here we analyze the same scaled version of the data with location varying bandwidth kernel density estimators. The biweight kernel is used in all examples in this section. Location varying smoothing, with bandwidth depending on location, i.e. $h(x)$, is done in all examples until Figure 7.

Figure 3a shows the result of a reasonable amount of global smoothing. Figure 3b shows that the same bandwidth was used everywhere. This amount of smoothing reveals 3 modes in the data, as found by Good and Gaskins, and several other "bump hunters" since.

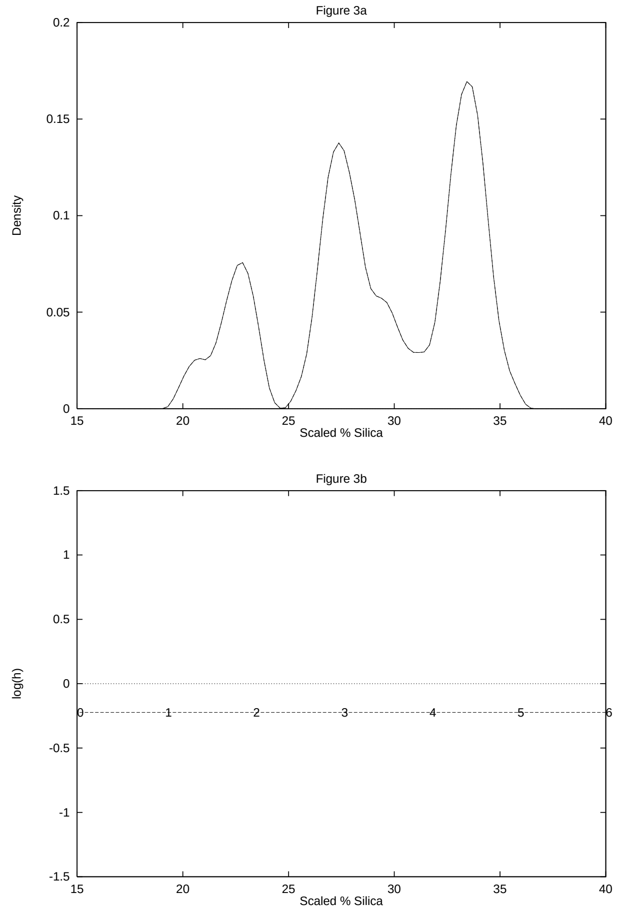

FIGURE 3: *Density estimate for the Chrondite data in (a), bandwidth function in (b). Same bandwidth used in all locations, shows trimodal structure.*

Figure 4 shows the result of some manipulation of the bandwidth function. Note that the first mode in Figure 3 has now been separated into two modes, by using a relatively small bandwidth in that region. The second mode in Figure 3 is now smaller with a different shape. The third mode is still about the same, but note that a spurious fifth mode has appeared near $x = 38$, where there is no data! This last spurious mode was generated by using a large bandwidth function there, which "reached out for mass" to the regions with data. Note that location varying bandwidth estimation has the potential for serious misrepresentation of the data.

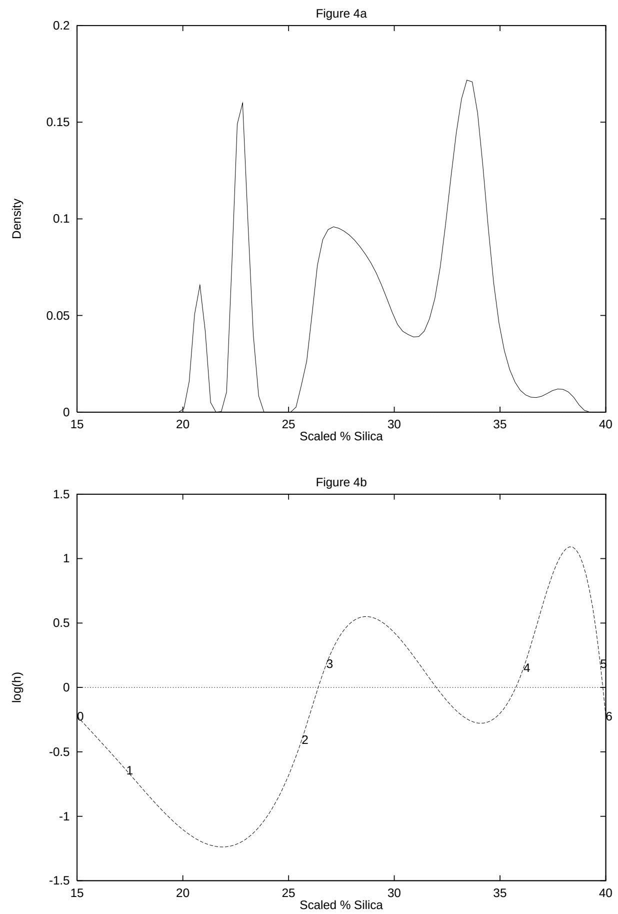

FIGURE 4: *Density estimate for the Chrondite data in (a), bandwidth function in (b). Now see 5 modes, with relative sizes and shapes different.*

Figure 5 shows the result of an attempt at further misrepresentation of the data. Note that the first mode in Figure 3 has been completely obliterated by a very large bandwidth for $17 < x < 24$, although two spurious modes created by the same effect as the fifth mode in Figure 4 have appeared at $x = 17$ and $x = 24$. The second mode in Figure 3 has become huge, with one spike separated on the right side, by a very small bandwidth in that area. The third mode of Figure 3 has also been exaggerated.

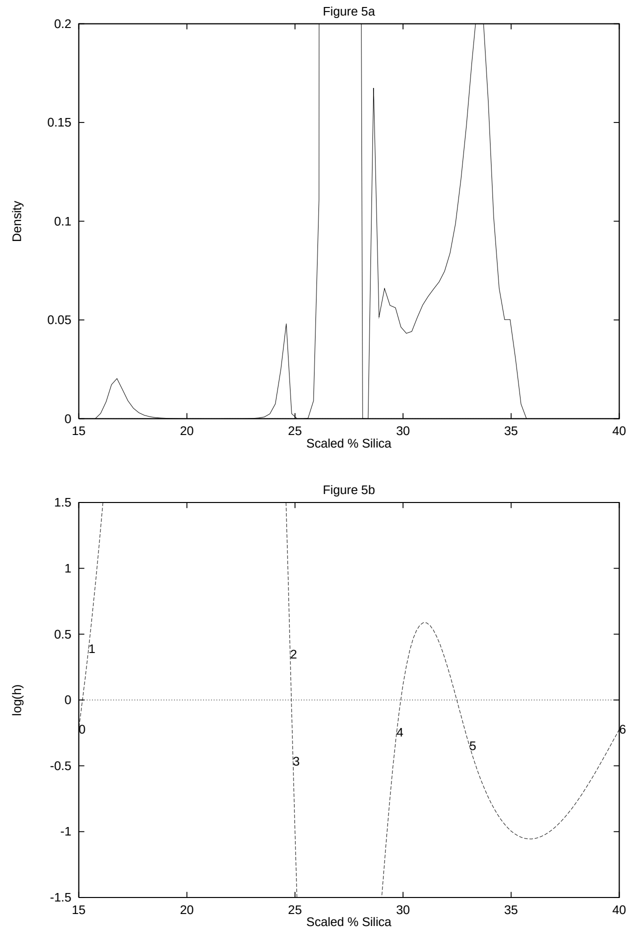

FIGURE 5: *Density estimate for the Chrondite data in (a), bandwidth function in (b). Density shape very different, and not representative of the data.*

An extreme example is shown in Figure 6. By taking the bandwidth function very large in the range of the data, the estimate is essentially 0 there (note that for any x, $\lim_{h\to\infty} \widehat{f}_h(x) = 0$). The estimator does show two modes, but these are outside the range of the data. These are caused by the same "reaching out" effect that generated the fifth mode in Figure 4.

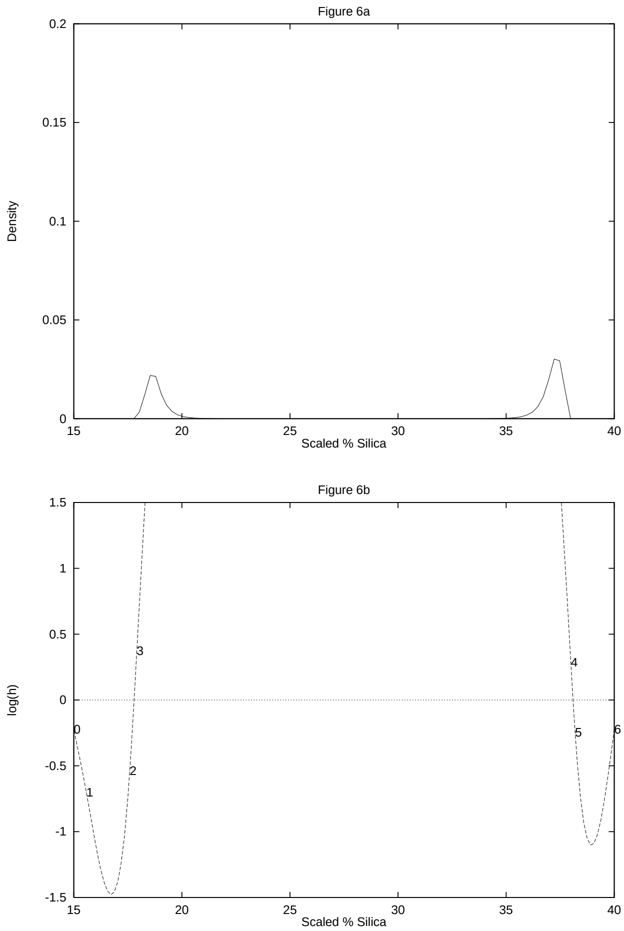

FIGURE 6: *Density estimate for the Chrondite data in (a), bandwidth function in (b). Estimate is 0 in the region containing the data, has two bumps well outside the range of the data.*

Figure 7 gives yet another misrepresentation of the data. Now the second mode of Figure 3 appears much bigger than the others. Figure 7 also shows that there is a difference between allowing the bandwidth to depend on location, $h(x)$, and on the data points $h(X_i)$. In general, we found that the $h(x)$ type of bandwidth variation was somewhat more flexible (not surprising in view of the fact that it's area is not constrained), but also more prone to spurious modes, as seen here. Especially important is that it does not seem possible to create modes that appear outside of the range of the data using $h(X_i)$. But the most important lesson is that both types of location varying bandwidth deserve to be treated with a good deal of healthy skepticism, because the impression one receives from the estimate is easily manipulated in any of many possible directions.

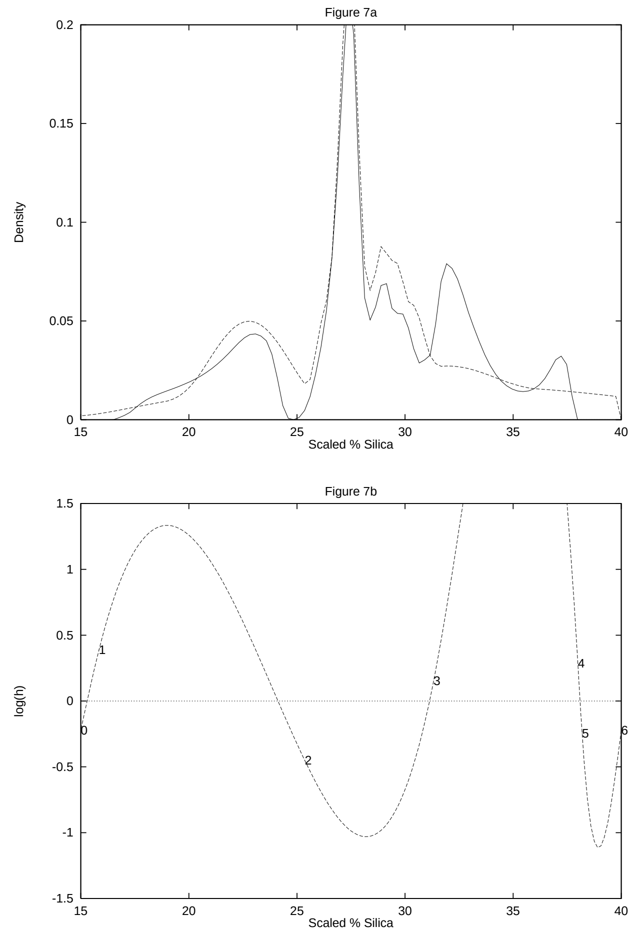

FIGURE 7: *Density estimate for the Chrondite data in (a), bandwidth function in (b). Much different shape from other examples, also shows varying bandwidths depending on location, $h(x)$ (solid curve), and depending on data points, $h(X_i)$ (dashed curve).*

4 Which curves can be estimates?

The example in Section 3 suggests that for a given set of data and a given kernel, the set of possible kernel density estimates is very large. Here this set is investigated, when the kernel function is a symmetric probability density.

For bandwidth variation depending on the data values, the set of possible kernel density estimators is a set of mixture densities, of the form:

$$\left\{ f(x) = \sum_{i=1}^{n} \frac{1}{n} K_{h_i}(x - X_i) : \mathbf{h} = (h_1, ..., h_n) \in \Re^n \right\}.$$

This allows a wide range of possible features, especially number and location of modes. For example, when the kernel is Gaussian, given any point x_0 in the range of the data, a member of this family has a unique mode at x_0 (choose very large bandwidths for all data points, except for one point on each side of x_0, adjust those bandwidths to put the mode at x_0).

For bandwidth variation depending on x, even more is possible. To understand this, for a given continuous kernel K and data set $\mathbf{X} = \{X_1, ..., X_n\}$, call the set of points in $\Re^2$, that some kernel estimate passes through, the "envelope" of the estimate:

$$Env(K, \mathbf{X}) = \left\{ \left(x, \hat{f}_h(x)\right) : x \in \Re, h > 0 \right\}.$$

Note that any function whose graph is contained in $Env(K, \mathbf{X})$ is a varying bandwidth kernel estimate for some bandwidth function $h(x)$.

To study the set $Env(K, \mathbf{X})$, let $\overline{b}_{K,\mathbf{X}}(x) = \sup_{h>0} \hat{f}_h(x)$ denote the upper boundary. It is straightforward to show that $Env(K, \mathbf{X})$ consists of all points between the x-axis and the curve $\overline{b}_{K,\mathbf{X}}$. This shows that the area under this type of bandwidth varying density estimate can be arbitrarily close to 0. The curve $\overline{b}_{K,\mathbf{X}}(x)$ is studied next.

To find an upper bound, note that from

$$\sup_h \left| \frac{x}{h} \right| K\left(\frac{x}{h}\right) = \sup_t tK(t)$$

it follows that

$$\sup_h K_h\left(x - X_i\right) = \frac{\overline{K}}{|x - X_i|},$$

where $\overline{K} = \sup_{t>0} tK(t)$. Hence,

$$\overline{b}_{K,\mathbf{X}}(x) = \sup_h n^{-1} \sum_{i=1}^{n} K_h\left(x - X_i\right) \leq n^{-1} \sum_{i=1}^{n} \frac{\overline{K}}{|x - X_i|}, \tag{2}$$

i.e. is bounded by a kernel estimate with a nonintegrable kernel proportional to $1/x$.

This same shape also appears in a lower bound. Assume that "K contains an ϵ rectangle" in the sense that $K(x) \geq \epsilon$ for all $x \in [-\epsilon, \epsilon]$. Then for each $x \in \Re$ and for $i = 1, ..., n$, the bandwidth choice $h = \left|\frac{x - X_i}{\epsilon}\right|$ gives

$$\overline{b}_{K,\mathbf{X}}(x) \geq \widehat{f}_{\left|\frac{x-X_i}{\epsilon}\right|}(x) = \frac{1}{n\left|\frac{x-X_i}{\epsilon}\right|}\sum_{i'=1}^{n} K\left(\frac{x - X_{i'}}{\left|\frac{x-X_i}{\epsilon}\right|}\right) \geq \frac{\epsilon^2}{n\left|x - X_i\right|}. \qquad (3)$$

This makes it clear that the area under this type of bandwidth varying kernel density estimate can be infinite.

The upper and lower bounds in (2) and (3) show that the boundary of the envelope is intimately connected with the function $1/x$. An approximation to $\overline{b}_{K,\mathbf{X}}(x)$ for the Chrondite data is shown in Figure 8. The envelope in Figure 8b was constructed by taking the maxima of a large family of Gaussian kernel density estimates, shown in Figure 8a. Figure 8b also contains a good constant bandwidth estimate, in particular the method of Sheather and Jones (1991) [18], which shows the area of the envelope is much bigger.

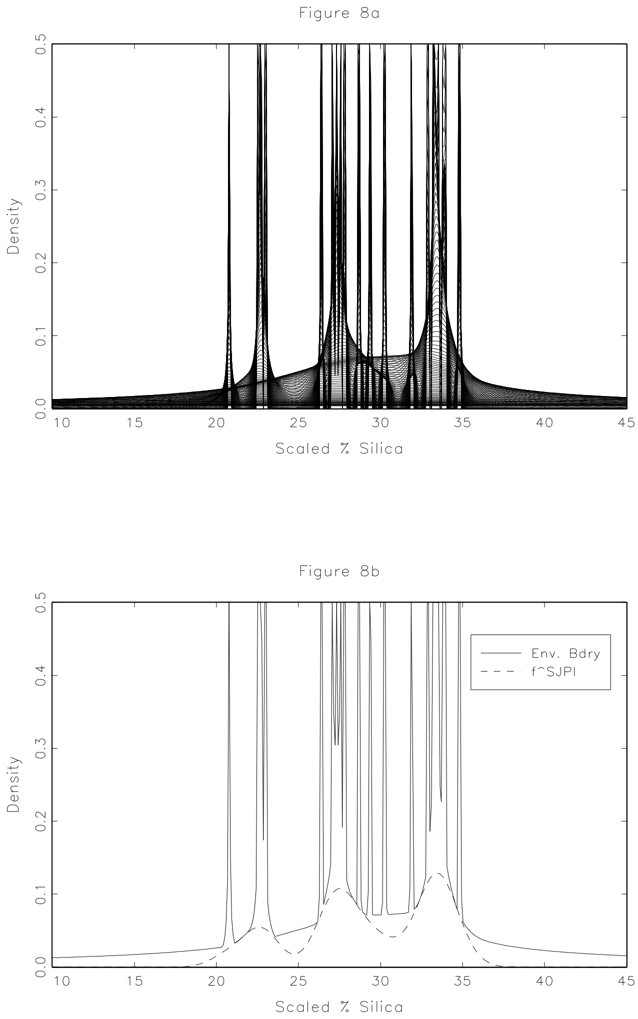

FIGURE 8: *Envelope for the Chrondite data in Figure 3. Large family of kernel density estimates are overlaid in (a). Maxima of these gives envelope boundary in (b), together with a single good global bandwidth estimate.*

Figure 8a shows where the various features in the envelope come from. The thin spikes correspond to single data points. The fatter spikes are several data points very close together. The flat spots come from relatively large bandwidths. The rounded corners near the fatter spikes come from a fairly wide range of different bandwidths. As expected from the above, the tails of the envelope go down very slowly ($\sim \frac{1}{|x|}$ for $|x| \to \infty$).

5 Details of the implementation

We have used LISP-STAT to implement the techniques of this paper. There are two kind of programs involved: interaction programs and computation programs. Interaction programs are for building the graphical objects that show the curves on the screen, read the mouse actions, and interpret what the user wants to do with these actions. Computation programs are used to compute density estimates given the bandwidth function (constant or variable, variable on data or on location).

LISP-STAT provides useful tools for the development of graphical interfaces, see Tierney (1990) [20]. We have used this object oriented environment to build `KDE objects`. This provides the user with easy adjustment of the parameters and options of kernel density estimation, see Udina (1995) [21]. More information about the software is available, using some WWW reader, in `http://libiya.upf.es/` or by `ftp` in `halley.upf.es:pub/stat/`. It is available also in *statlib*, $\star$ we will put the reference here when we have it $\star$. The variable bandwidth method shown in figures 2-7 has been implemented in KDE objects. By means of the mouse, the user can move the knots. From the knots, a cubic B-spline (see deBoor (1978)[3]) passing through the knots is computed and it is used to compute the bandwidth for each location. A possible extension of the methods given here is to replace the knot controlled cubic splines with Bézier-like splines (see Newman and Sproull (1979) [14] or Foley and van Damm (1982)[6]). This will allow more convenient user manipulation of the bandwidth function, via *handles* that are not points on the curve. In addition to better manipulation of the bandwidth function it would be more difficult to reach values outside a reasonable range as in figure 6.

We use binning methods to compute the density estimates. This means that data are discretized to a grid of values, usually a grid size of some hundreds is used. For a comprehensive discussion of these discretized methods, see Fan and Marron (1994)[5]. Here we use the linear binning method to discretize the data, though our software allows other methods. This binning process is of great relevance when dealing with big data sizes as in the income data example presented in section 1. With these preprocessed data, updating methods as described by Fan and Marron (1994)[5] are used to compute the density estimate. This method requires that the kernel function be a polynomial, so the beta family is quite useful. The updating method consists of rearranging the polynomial in such a way that the coefficients involved can be *updated* when passing from one grid point to the next one with few operations.

6 Acknowledgement

Research of the first author was partially supported by National Science Foundation Grant DMS-9203135. Research of the second author was done while visiting the University of North Carolina at Chapel Hill and was partially supported by spanish DGICYT grants 91.0814 and PB92.1037.

References

[1] Abramson, I. S. (1982) On bandwidth variation in kernel density estimation - a square root law, *Annals of Statistics*, 9, 168-176.

[2] Breiman, L., Meisel, W. and Purcell, E. (1977) Variable kernel estimates of probability densities, *Technometrics*, 19, 135-144.

[3] De Boor, C., *A Practical guide to splines*, Springer, New York.

[4] Eubank, R. L. (1988) *Spline smoothing and Nonparametric Regression*, Dekker, New York.

[5] Fan, J. F. and Marron, J. S., (1994) Fast implementation of nonparametric curve estimators, *Journal of Computational and Graphical Statistics*, 3:35-56.

[6] Foley, J. and van Dam, A., (1982) *Fundamentals of interactive computer graphics.* Addison-Wesley, REading, MA.

[7] Good, I. J. and Gaskins, R. A. (1980) Density estimation and bump hunting by the penalized likelihood method exemplified by scattering

and meteorite data, *Journal of the American Statistical Association*, 75, 42-73.

[8] Härdle, W. (1990) *Applied Nonparametric Regression*, Cambridge University Press.

[9] Jones, M. C. (1990) Variable kernel density estimates and variable kernel density estimates, *Australian Journal of Statistics*, 32, 361-371.

[10] Jones, M. C., Marron, J. S. and Sheather, S. J. (1992) Progress in data-based bandwidth selection for kernel density estimation, unpublished manuscript.

[11] Jones, M. C., McKay, I. J. and Hu, T. C. (1994) Variable location and scale density estimation, *Annals of the Institute of Statistical Mathematics*, 46, to appear.

[12] Loftsgaarden, D. O. and Quesenberry, C. P. (1965) A nonparametric estimate of a multivariate density function, *Annals of Mathematical Statistics*, 36, 1049-1051.

[13] Müller, H. G. (1988) *Nonparametric Analysis of Longitudinal Data*, Springer Lecture Notes in Statistics 46, Springer, New York.

[14] Newman, W. M. and Sproull R. F. (1979) *Principles of interactive computer graphics* McGraw-Hill, New York.

[15] Samiuddin, M. and El-Sayyad, G. M. (1990) On nonparametric kernel density estimates, *Biometrika*, **77**, 865-874.

[16] Schmitz, H. P. and Marron, J. S. (1992) Simultaneous estimation of several size distributions of income, *Econometric Theory*, **8**, 476-488.

[17] Scott, D. W. (1992) *Multivariate density estimation: theory, practice and visualization.* Wiley, New York.

[18] Sheather, S. J. and Jones, M. C. (1991) A reliable data-based bandwidth selection method for kernel density estimation, *Journal of the Royal Statistical Society, Series B*, **53**, 683-690.

[19] Silverman, B. W. (1986) *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London.

[20] Tierney, L. (1990) *LISP-STAT: An Object Oriented Environment for Statistical Computing and Dynamic Graphics*, John Wiley and Sons, New York.

[21] Udina, F. (1995) KDE Objects, *Economics Working Paper Series*. Universitat Pompeu Fabra, Barcelona.

[22] Wahba, G. (1990) *Spline Models for Observational Data*, SIAM, Philadelphia.

[23] Wand, M. P., Marron, J. S. and Ruppert, D. R. (1991) Transformations in density estimation, *Journal of the American Statistical Association*, **86**, 343-361.