Comment

Frederic Udina[*]

September 4, 1997

1 Introduction

I would like first to thank the editors for setting up this forum around the very interesting paper written by Luc Devroye and giving myself the opportunity to participate in it.

As I knew about the new methodology for bandwidth selection developed by Devroye and Lugosi (DL), I was attracted by both the elegance and simplicity of the involved ideas and, of course, by the universality they achieve. My second thought was about the practical performance of these selectors. Luc Devroye has made a really great work by comparing the new methods with the classical ones and showing that they are very competitive. It would be very interesting to extend it to include sample sizes other than $N = 100$.

In the following lines I want to discuss how Devroye-Lugosi bandwidth selectors behave in practice when facing real data sets. This concern two main issues: how to adapt the method to work in a binned data computational setting and how to protect the method against typical data manipulation such as rounding or sorting.

2 Binned version of Devroye-Lugosi selector

Using kernel estimation techniques in practice means to use some discretization technique like binning as described, for example, in Fan and Marron (1993) or, under the name of *warping* in Härdle and Scott (1992). It

*Universitat Pompeu Fabra, Ramon Trias Fargas 25–27, 08005 Barcelona. e-mail: `udina@upf.es`

is the only practical way to deal with big or moderately big data sizes. For a given grid g_k, $\{j = 1 \dots G\}$, bin counts c_k are computed for the first $n - m$ data points using linear binning. Then approximated estimates for density values are computed for every g_k needing only a few kernel function values. The finite convolution needed can be computed even faster using fast Fourier transform techniques. We use in this quick study the global version of the "First bandwidth" as described in Devroye's paper. The slight changes we make to the algorithm will be commented. In this binned context, the approximated Yatracos set $\tilde{\mathcal{A}}_{ij}$ is simply a union of bins (the ones where $f_{h_i}(g_k) > f_{h_j}(g_k)$) that can be coded as a bit vector of length G. Yatracos distance computation is really fast and simple in this situation.

A potential problem that would need some study appear: although binning gives very good approximated density values we have no guarantee of having good approximation to the true Yatracos sets (whatever can this mean) and this can distort the distances to be minimized. Our experience show that the choice of G and the binning width w do not affect the selected bandwidth, provided that G is big enough and the width is small. We usually take G in the hundreds. Once w is fixed, we use it to decide the minimum bandwidth to be considered, because it has no sense to take bandwidth values that would include less than several grid values in the computation.

3 Dealing with *real* data sets

When we deal with real data sets we must be careful about how data has been transferred to the analyst. It's quite usual, for example, that data have been collected in rounded form or have been rounded a posteriori. Some bandwidth selectors are robust against any small change in data values, but the Devroye-Lugosi universal selector can have problems with repeated values. We will discuss some examples below.

Another source of problems can come from data being arranged in some way before arriving to the algorithm. Some of my time was spent investigating strange behavior in the algorithms proposed by the paper. The reason was that the data were sorted in increasing order in the file where it came to me. Obviously the method doesn't work at all with sorted data if split is done in the usual way, taking the first m values. Similar problems can appear if data is collected or transmitted in a non-random ordering, by different categories, for example. Even in some Monte-Carlo simulations I found that some of my algorithms were generating data from normal mixtures in a component-by-component way. This normally doesn't have any consequence but when splitting the data set to apply the Devroye-Lugosi procedures, problems arose.

The strategy to follow when data is suspect to have been sorted is not easy to devise. One possibility is to shuffle randomly the data set and then to take the first m data values to build the empirical measure. This has the problem of producing a different selected bandwidth every time the algorithm is applied. As we will see in the examples, changing the m length subsample introduces a lot of variability in the selected bandwidth. In the paper, Devroye introduces the rotation idea to compensate it. The obvious drawback is that computation time grow, as the method itself is quite time consuming. Another possibility could be extracting the m data points in some systematic way. For example, a simple scheme could be, to take m elements out of n, sort the data set and take those with ordered index $\lfloor ni/m \rfloor$ for $i = 1 \dots m$. But this need some theoretical justification because violates the independence assumption and in case of repeated values would favor even more coincidences between the two subsamples.

4 Some real data sets

I was trying first to work with the Miguel Hidalgo stamps data set brought to the density estimation area by Izenmann and Sommer (1988). It's a data set that present both a rich structure to study and a reliable sample size (n=485). Data comes from thickness measured from a collection of old stamps issued in 1872-74 in Mexico. But unfortunately, data is heavily rounded to thousandth of millimeters resulting in a high proportion of repeated data. This results in the impossibility to apply the described techniques to it. The Sheather-Jones plug-in bandwidth, for example, is 0.00272 for this data set, so it is in the same order of the resolution of the data. To select this bandwidth, DL would need to fix a minimum bandwidth so small that the estimates will result a series of single value peaks. When splitting the data, the m subsample results to be a subsample of the $n - m$ one. The result is that the minimum bandwidth is selected most of the times the algorithm is run. Probably Devroye-Lugosi theory might be modified to work with smoothed discrete distributions and then some work could be done with this data set.

Postman, Huchra and Geller (1986) studied the velocities of 82 galaxies moving away form our galaxy. The data set was deeply analyzed from the kernel smoothing point of view in Park and Turlach (1992). They gave bandwidth values selected by several automatic selectors. Translated to our kernel, they found from 1.37 for least squares cross-validation to 3.48 for biased cross-validation. For the Sheather-Jones plug-in they gave 1.44 and 2.72 for the normal-reference rule-of-thumb. The data set arrived to me sorted in increasing order, so the re-shuffling strategy was applied.

	# of h values			
m/n	10	20	30	40
0.1	1.339	1.686	2.543	2.626
0.2	1.928	1.418	1.811	1.724
0.3	2.777	3.365	3.189	3.107
0.4	2.777	2.382	2.271	2.414
0.5	1.928	1.686	1.811	1.724

Figure 1: Values selected by the DL algorithm as described in the text, for several values of m/n and of the number of estimators in contest.

Applying the version of DL selector described above we obtained values around 2.0, depending on parameter choice. To show how the value selected depends on these parameters, we shuffled the data set once (it arrived to me sorted in increasing order again) and we run the algorithm with m being 10%, 20%, 30%, 40%, and 50% of the data size $n = 82$ and taking the number of different values for h in the range $[0.15, 4.0]$ as 10, 20, 30, and 40. The minimum bandwidth is fixed as 1.5 times the binning width (so ensuring that at least three bin are involved in any computation). The maximum bandwidth is chosen to be the interquartilic range. The resulting values are shown in the table in figure 1. The variability of the selected parameter is noticeable. As expected, higher values of the ratio m/n give less variability. But even using the same pair of parameters, shuffling data can give very different values for the selected bandwidth. For example, with $m/n = .3$ and 20 bandwidth values, we obtained 0.423, 1.418, 1.418, 1.686, and 1.193 in five runs of the algorithm with re shuffling. To test the rotation idea suggested by Devroye, we took $m = n/2$ and computed the geometric mean of the two selected bandwitdhs obtained using both halves alternately as m subsample. Number of bandwidths in contest were 20 in the range [0.154.0]. The values obtained were 0.7751.0041.0941.5461.546. We see that variability has not improved so much.

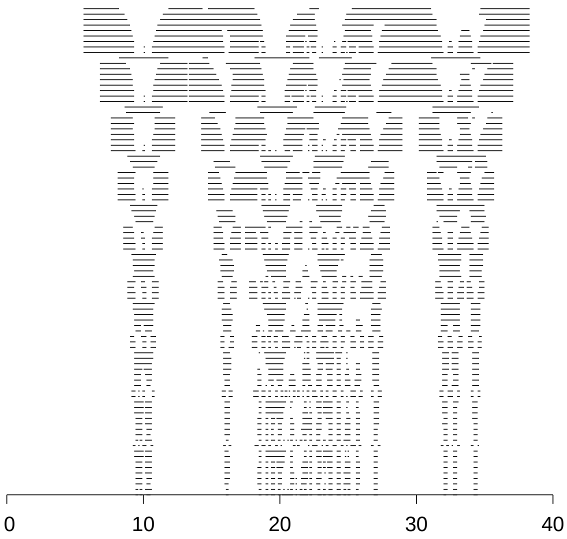

Figure 2: Yatracos sets for galaxies data

As a matter of curiosity, we also show in figure 2 the shape that Yatracos sets appear to have. For 10 bandwidth values (from .15 to 4.0 in geometric steps) we generated 10 density estimates $f_1, \ldots, f_{10}$ and computed the Yatracos sets $A_{ij} = \{x | f_i(x) > f_j(x)\}$ Each horizontal line in figure 2 represent a set. The lower band composed by nine lines corresponds, as going up, to $A_{1j}, j = 2..10$ and so the other 10 bands.

Most of the variability found in the selected bandwidths for the galaxies data can be explained by the sample size being too small. We decided to contrast it with the case of a very big data set, also well known in the smoothing literature. The net income of 7201 British households in the year 1975 was studied in Schmitz and Marron (1992) and also discussed in Wand, Marron and Ruppert (1991). Data range goes from 0.026 to 9.1225 the unit being the average income. Despite the data range and resolution, the set contains a quite high number of repeated values, namely 1357. We took a grid size of 800 in the range -1 to 10, so the minimum bandwidth to be considered is approx .022. The maximum was taken here as half the interquartilic range and 25 bandwidth values were taken in the resulting interval [.022.4]. m was chosen to be 25% of the sample size. In figure 3 we plot Yatracos distance over (logarithm of) bandwidth values. 8 different runs are plotted resulting from re shuffling before split. The selected bandwidths were 0.051, 0.058 twice, 0.065 three times, 0.094 twice, 0.106 and 0.152. We see that variability is somewhat reduced but it's still quite high. Applying the systematic split discussed in section 2 selected a bandwidth of 0.025, clearly too small. The reason can be the relatively high percentage of repeated values. Applying the rotation algorithm in this setting resulted to more reasonable and stable values, we got 0.10, 0.094, .081 in three different runs.

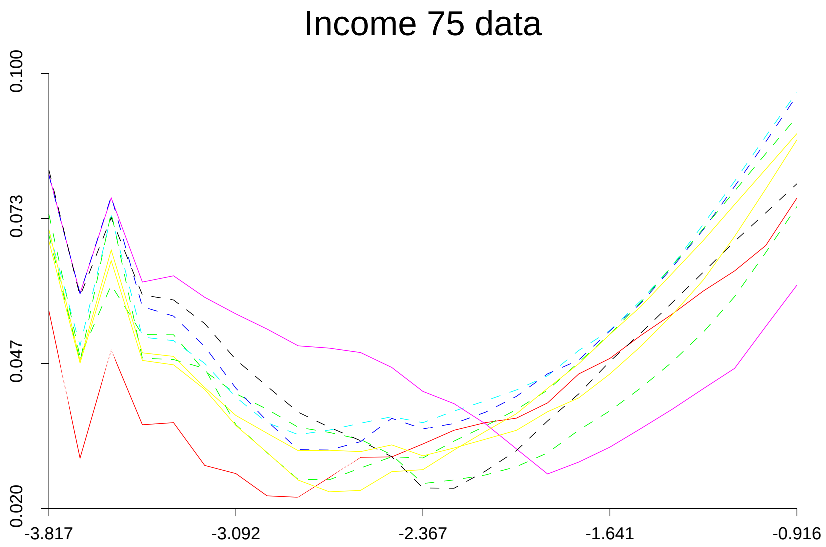

Figure 3: Values for the Yatracos distance in vertical axis. In horizontal axis, log of bandwidth. Each line is a diferent run of the algorithm with previous data shuffling.

5 Conclusion

We have seen that to use the DL selectors with real data, binning techniques can be used but some problems must be faced. The main one is how to deal with repeated values – this seems to need some theoretical adjustment. The other problem is to protect the algorithm against data that can be sorted o arranged in some non-random way. It looks like the shuffling-then-rotate approach is the safer one but further study is needed.

6 Additional references

FAN, J., MARRON, J. S. (1992) "Best possible constant for bandwidth selection" *Annals of Statistics,* **20**, 2057–2070

HÄRDLE, W. K., SCOTT, D. W. (1992) "Smoothing by Weighted Averaging of Rounded Points" *Computational Statistics,* **7**, 97–128

IZENMAN, A. J.; SOMMER, C. (1988) "Philatelic Mixtures and Multimodal densities" *JASA,* **83**, 941–953

POSTMAN, M., HUCHRA J.P. AND GELLER, M.J. (1986) "Probes of large scale structures in the Corona Borealis region" *The Astronomical Journal,* **92**, 1238–1247

PARK, B. U., TURLACH, B. A. (1992) "Practical Performance of Several Data driven bandwidth selectors" *Computational Statistics,* **7**, 251–270

SCHMITZ, H.P. AND MARRON, J.S. (1989) "Simultaneous estimation of several Size Distributions of Income" *Discussion paper A-186, SFB 303, University of Bonn, Dept. of Economics,*

WAND, M. P., MARRON, J. S., RUPPERT, D. (1991) "Transformations in Density Estimation" *JASA,* **86**, 414, 343–361