The Stata Journal (2003)
3, Number 2, pp. 133–147

Exploring the use of variable bandwidth kernel density estimators

Isaías H. Salgado-Ugarte
F.E.S. Zaragoza U.N.A.M. Biología &
Depto. de Biología, U.A.M. Iztapalapa
isalgado@servidor.unam.mx
Marco A. Pérez-Hernández
Depto. de Biología, U.A.M. Iztapalapa; México
maph@xanum.uam.mx

Abstract. Variable bandwidth kernel density estimators increase the window width at low densities and decrease it where data concentrate. This represents an improvement over the fixed bandwidth kernel density estimators. In this article, we explore the use of one implementation of a variable kernel estimator in conjunction with several rules and procedures for bandwidth selection applied to several real datasets. The considered examples permit us to state that when working with tens or a few hundreds of data observations, least-squares cross-validation bandwidth rarely produces useful estimates; with thousands of observations, this problem can be surpassed. Optimal bandwidth and biased cross-validation (BCV), in general, oversmooth multimodal densities. The Sheather–Jones plug-in rule produced bandwidths that behave slightly better in this respect. The Silverman test is considered as a very sophisticated and safe procedure to estimate the number of modes in univariate distributions; however, similar results could be obtained with the Sheather–Jones rule, but at a much lower computational cost. As expected, the variable bandwidth kernel density estimates showed fewer modes than those chosen by the Silverman test, especially those distributions in which multimodality was caused by several noisy minor modes. More research on the subject is needed.

Keywords: st0036, kernel density estimation, bandwidth, cross validation, multimodality test

1 Introduction

The variable bandwidth kernel density estimators (VBKDE) were briefly introduced in a previous article (Salgado-Ugarte et al. 1993). In this paper, we present some VBKDEs updated programs, and based on Sheather (1992) ideas, we explore their performance in conjunction with several rules and procedures for bandwidth choice on four real datasets previously analyzed.

The kernel density estimators (KDEs) are superior at recovering interesting structure and should be used instead of the traditional histograms or frequency polygons to analyze in detail data distributions. As stated in Salgado-Ugarte et al. (1993), the fixed bandwidth KDEs are vulnerable to noise in any low count interval of the distribution and miss distribution details in regions where data concentrate. In more technical terms,

the bandwidth (h) should increase with $f(x)$ to reduce variance and should decrease with $|f''(x)|$ to reduce bias. The ordinary kernel estimator lacks adaptivity and thus tends to oversmooth regions with high structure and to undersmooth the distribution tails or any data range with low structure (Simonoff 1996). To address this problem, one idea is to increase the window width in areas of low data densities and to decrease it at intervals with high counts. In this way, it is possible to recover detail where data concentrate and to eliminate noise where observations are sparse (Fox 1990).

There are several kernel density estimators, each using different algorithms and functions. Which KDE is the best? Currently, no method can claim to be "best", but one candidate would be the variable-bandwidth kernels (besides local kernels and transformation-based kernels), which achieve desirable local adaptivity and retain the intuitive appeal of the kernel. Nevertheless, they may share problems such as boundary bias (Simonoff 1996). A different approach has been suggested by Udina (1995), and more precisely by Marron and Udina (1998), who propose a subjective interactive procedure to vary the bandwidth. In what follows, we present some implementations of a variable bandwidth KDE in combination with several recommended procedures for data-based bandwidth choice.

2 New programs

The following programs are updated versions of the previously introduced (Salgado-Ugarte et al. 1993) programs `adgakern` and `adgaker2`, which use the algorithm adapted from Silverman (1986) by Fox (1990):

2.1 Syntax

`varwiker` *varname* [`if` *exp*] [`in` *range*], `bwidth(`#`)` [`gen(`*denvar*`)` `nograph` *graph_options*]

`varwike2` *varname* [`if` *exp*] [`in` *range*], `bwidth(`#`)` [`npoint(`#`)` `gen(`*denvar gridvar*`)` `numodes` `modes` `nograph` *graph_options*]

2.2 Description

`varwiker` estimates the density of *varname* using the variable bandwidth Gaussian kernel described in Fox (1990) modified from Silverman (1986) and draws the result.

`varwike2` estimates the density of *varname* using the variable bandwidth Gaussian kernel described in Fox (1990) modified from Silverman (1986), but at the second calculation stage only uses a uniformly spaced number of points (50 by default) to finish drawing the graph of the estimation.

2.3 Options

`bwidth(#)` specifies (as a geometric mean) the width of the window around each data point. `bwidth` is not optional. If not specified, the program halts and displays an error message on screen.

`npoint(#)` specifies the number of equally spaced points (grid) in the range of *varname* used for the density estimation. The default is 50 gridpoints.

`gen(`*denvar gridvar*`)` generates the variable *denvar* with the density values (`varwiker`) or generates the variable *denvar* with the density values estimated at the points given by *gridvar* (`varwike2`).

`numodes` displays the number of modes in the density estimation.

`modes` lists the estimated values for each mode. The `numodes` option must be included first.

`nograph` suppresses drawing of the graph.

graph_options are any of the options allowed with `graph, twoway` (Stata 7).

3 Methods

3.1 Variable bandwidth kernel density estimation

As a first step, `varwiker` estimates densities using a Gaussian kernel with fixed window, then uses these estimates to determine local weights inversely proportional to the preliminary density estimate. These local weights are used to adjust the window width so that it is narrower at high densities (retaining detail) and wider where density is low (eliminating noise). The algorithm to calculate the variable bandwidth KDE is (Silverman 1986; Fox 1990; Salgado-Ugarte et al. 1993) as follows:

1. Calculate a preliminary density estimate by using a fixed-bandwidth kernel function, $\widehat{f}_K(x)$.
2. At each observation X_i, calculate a local window factor, w_i, that is inversely related to the density estimate

$$w = \left\{ \frac{\widetilde{f}_\epsilon}{\widehat{f}_K(X_i)} \right\}^{1/2}$$

where

$$\tilde{f}_\epsilon = \left\{ \prod_{i=1}^{n} \widehat{f}_K(X_i) \right\}^{1/n}$$

is the geometric mean of the $\widehat{f}(X_i)$, and thus the w_i weights have a product and geometric mean of one.

3. Use the weights to calculate the adaptive-kernel estimator

$$\widehat{f}_A(x) = \frac{1}{nh}\sum_{i=1}^{n}\frac{1}{w_i}K\left(\frac{x - X_i}{w_i h}\right)$$

4. Iterate steps 2 and 3, using $\widehat{f}_A$ in place of $\widehat{f}_K$. In practice, iteration produces little change in the estimated densities (Fox 1990).

Because this implementation requires the calculation of local weights for each individual observation based on a preliminary density estimation, the time required is proportional to `_N`. With a lot of observations, you need to be patient.

3.2 Bandwidth selection

In comparison with histograms, the use of kernel density estimation solves the problems of origin and discontinuity; however, that of interval width choice remains. There are several ways to select an appropriate bandwidth. In general, these methods approximate the optimal bandwidth by minimizing an error measurement under specified conditions. We have included a collection of these procedures in `bandw.ado` (Salgado-Ugarte et al. 1995b).

Sheather and Jones (1991), extending the work of Park and Marron (1990), described a bandwidth selection procedure with excellent properties. This plug-in procedure (as it is named) uses an iterative approach and, by a numerical algorithm, finds the h value, which solves the formula for the optimal smoother parameter. In particular, it is more stable than the cross-validation approach. Cross-validation estimates the integrated squared error function and locates the minimum. The plug-in approach minimizes the function theoretically and then estimates this minimizing value directly. Due to its good performance, this method is subject to less variability (Bowman and Azzalini 1997).

In what follows, we present the application of VBKDEs employing the following six bandwidth selection procedures:

1. Oversmoothed (OS)
2. Optimal of Silverman (O)
3. Least-squares cross validation (LSCV)
4. Biased cross-validation (BCV)
5. Sheather–Jones plug-in (SJPI)
6. Silverman test (ST)

These methods have been discussed elsewhere in the related literature: Silverman (1986), Härdle (1991), Scott (1992), Wand and Jones (1995), Simonoff (1996), and Bowman and Azzalini (1997); STB inserts on the subject are Salgado-Ugarte et al. (1993, 1995a, 1995b, 1997).

Procedures 1–4 and 6 were calculated using the programs presented in Salgado-Ugarte et al. (1995b; 1997); procedure 5 was estimated by an XploRe program routine (Härdle et al. 1995).

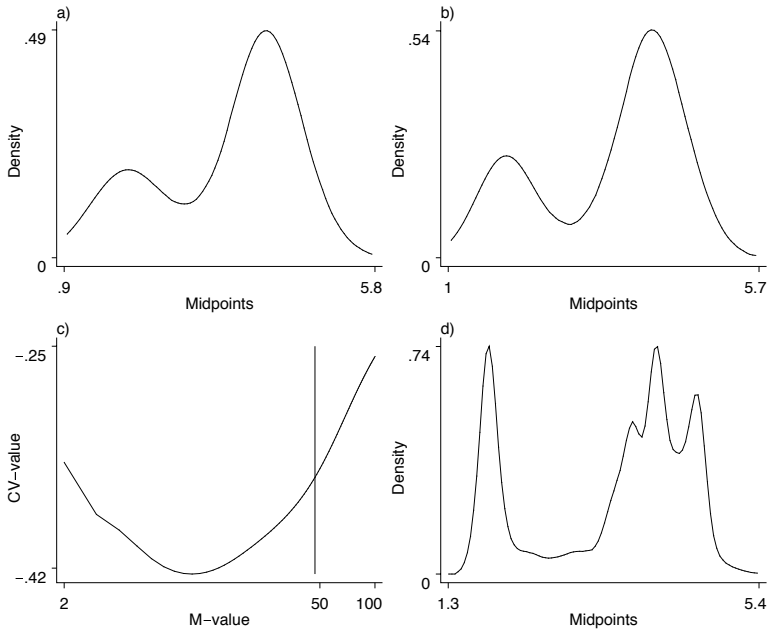

Figure 1: Geyser data: a) Oversmoothed VBKDE (G-mean of h = 0.47); b) Optimal VBKDE (G-mean of h = 0.37); c) LSCV score function. Minimum is located at M = 10, indicating a bandwidth h = 0.1; the line to the right represents the OS bandwidth (M = 47; h = 0.47); d) VBKDE with the LSCV bandwidth (G-mean of h = 0.1)

(*Continued on next page*)

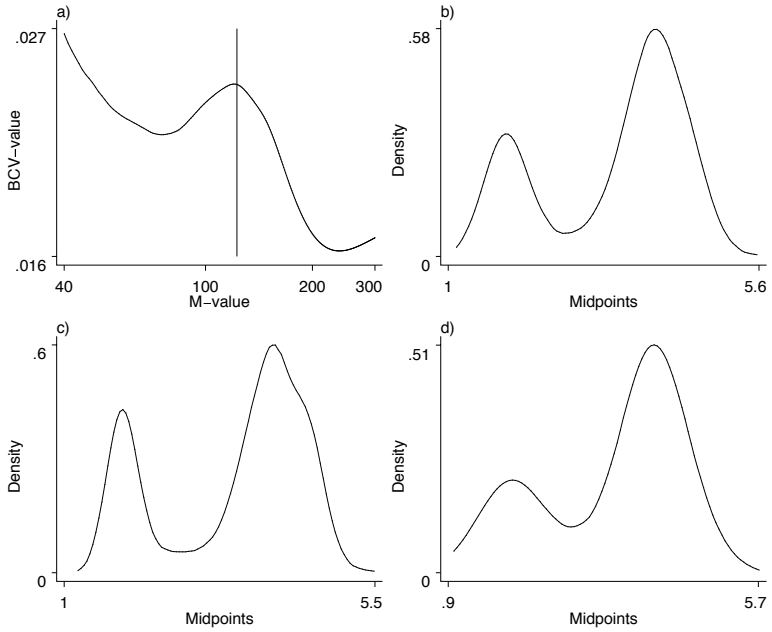

Figure 2: Geyser data: a) BCV score function. Vertical line on the right is the OS bandwidth (M = 122.6, h = 1.226). The local minimum to the left of the OS bandwidth is located at M = 75, h = 0.75; b) VBKDE with the BCV bandwidth, G-mean of h = 0.29 (rescaled to Gaussian by 0.75*0.381); c) VBKDE with the SJPI bandwidth (G-mean of h = 0.21); d) VBKDE with the ST bandwidth, G-mean of h = 0.43, resulting from (0.699 + 0.166)/2 (Table 1)

4 Datasets

To observe the performance of the variable bandwidth kernel, we use the following published datasets:

1. Old Faithful geyser data; in his monography on density estimation, Silverman (1986) considers the length of 107 eruptions of Old Faithful geyser in Yellowstone National Park, USA, taken from Weisberg (1980).

2. Old Mexican stamps data; the value of antique stamps in part is determined by the thickness of the paper they are printed on. In a classical report, Izenman and Sommer (1988) carried out a detailed analysis of the thickness of 485 specimens from the 1872 Hidalgo stamp issue of Mexico.

3. Galaxies data; in order to present a semi-parametric density estimation procedure, Röeder (1990) used the velocities of 82 galaxies in the Corona Borealis region, measured by Postman et al. (1986) to recognize clusters leading to a multimodal distribution.

4. Catfish data; in fisheries biology, the occurrence of multimodal size distributions can be used to estimate the growth of species. Salgado-Ugarte et al. (1997) employed data of standard body length measures of 1,116 catfish to present a bootstrap procedure to test multimodality.

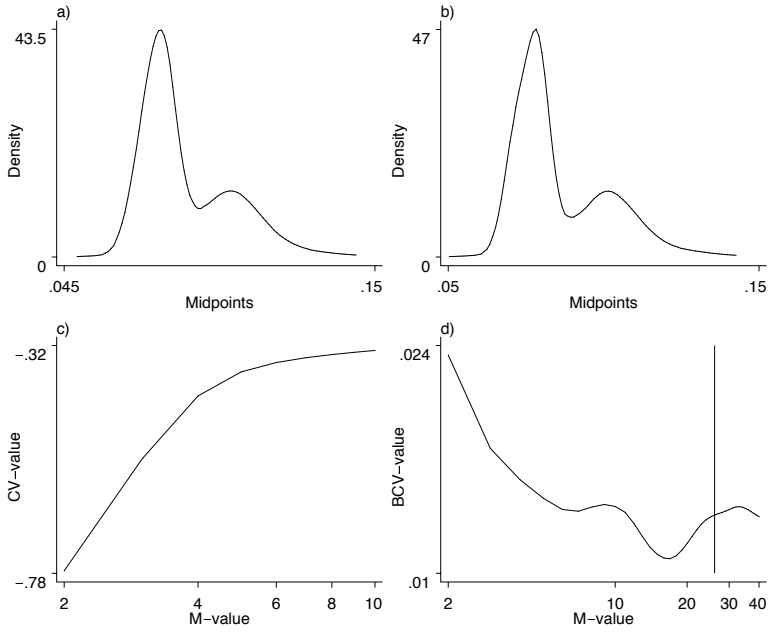

Figure 3: Mexican stamps data: a) Oversmoothed VBKDE (G-mean of h = 0.005); b) Optimal VBKDE (G-mean of h = 0.004); c) LSCV score function. There is no indication of a local minimum (bandwidth values are multiplied by 100) d) BCV score function. Vertical line on the right is the OS bandwidth (M = 26.072, h = 1.3036). There are two local minima, one at M = 7, h = 0.35 and other at M = 17, h = 0.85 (bandwidth values are multiplied by 100)

(*Continued on next page*)

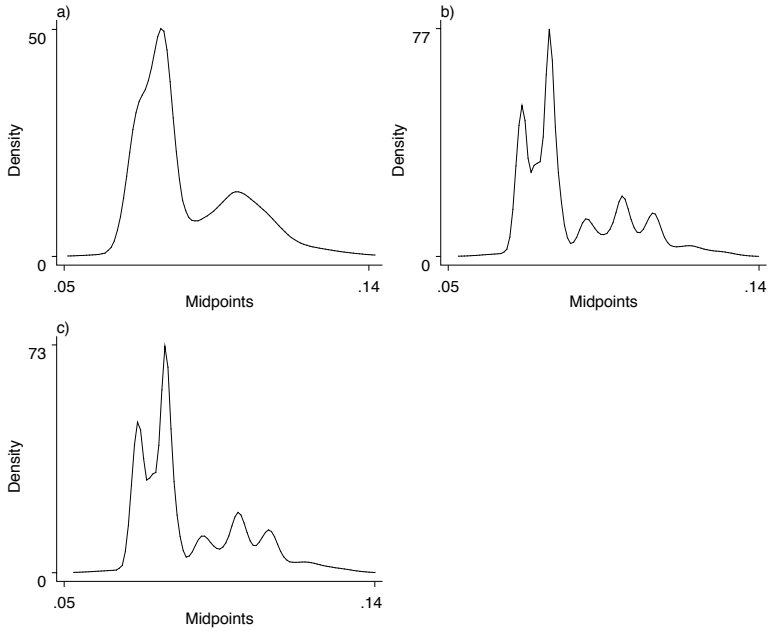

Figure 4: Mexican stamps data: a) VBKDE with the BCV bandwidth (rescaled to a Gaussian kernel) G-mean of h = 0.00324; b) VBKDE with the lowest value indicated by BCV = SJPI bandwidth (G-mean of h = 0.0013); c) VBKDE with the ST bandwidth, G-mean of h = 0.0015

5 Results

For all the cases, we present the VBKDEs employing the respective bandwidths to obviate a table with the numeric results. Included is one table with the Silverman's test results not previously reported, as the others are available in the papers on the subject.

5.1 Example 1: Geyser data

The bandwidths obtained were from 0.1 (LSCV) to 0.47 (OS) and the resulting estimates, excepting that from LSCV, are bimodal. The bimodal distribution obtained with the OS bandwidth makes evident the existence of (at least) two modes. Silverman (1986) smoothed these data using a trial-and-error scheme choosing a band of 0.25. This bandwidth value coincides with those produced by SJPI and BCV. The Silverman's test resulted in a bandwidth value located between the oversmoothed and the optimal. The LSCV bandwidth is very narrow, producing a very noisy density estimate. The BCV score shows two local minima, but only one of them is located to the left of the oversmoothed bandwidth.

Table 1: Critical bandwidths and estimated significance levels for geyser data Silverman (1986), n = 107

Number of modes	Critical bandwidth	*p*-value
1	0.700	0.00
2	0.166	0.53
3	0.133	0.50
4	0.116	0.47

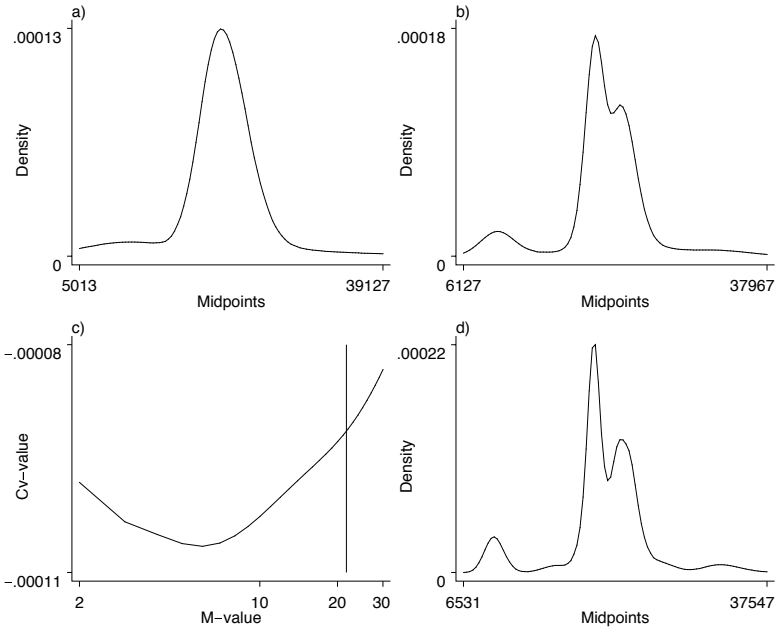

Figure 5: Galaxies data: a) Oversmoothed VBKDE (G-mean of h = 2164.7); b) Optimal VBKDE (G-mean of h = 1016); c) LSCV score function. Minimum is located at M = 6, indicating a bandwidth h = 600; the line to the right represents the OS bandwidth (M = 21.6 5; h = 2165); d) VBKDE with the LSCV bandwidth (G-mean of h = 600)

(*Continued on next page*)

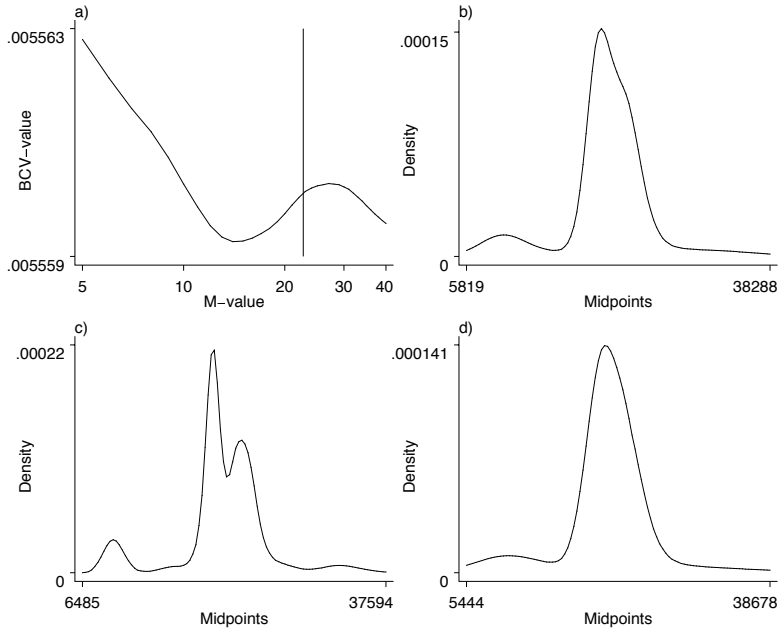

Figure 6: Galaxies data: a) BCV score function. Vertical line on the right is the OS bandwidth (M = 22.7, h = 5678). The local minimum to the left of the OS bandwidth is located at M = 14, h = 3500 (Quartic kernel); b) VBKDE with the BCV bandwidth (rescaled to a Gaussian kernel) G-mean of h = 1333.5; c) VBKDE with the SJPI bandwidth (G-mean of h = 647); d) VBKDE with the ST bandwidth, (G-mean of h = 1720)

5.2 Example 2. Old Mexican stamps

As reported by Sheather (1992), it was not possible to calculate the LSCV bandwidth due to the rounding of the original data. This can be observed in the LSCV function graph (Figure 3c), which does not show a local minimum. The other calculated bandwidths range from 0.005 (OS) to 0.0015 (PISJ and PS). The BCV score presented two local minima (Figure 3d); the lower bandwidth was equivalent to that of PISJ (0.0013) and the second had an intermediate value of 0.00324. The thick values were multiplied by 100 to run the cross-validation programs. The density estimates show from two to six modes. Izenman and Sommer (1988), based on exhaustive historical data, found explanation for the first five of their estimated seven modes. These authors also used the likelihood-ratio test (Wolfe 1970), which suggested a mixture of five Gaussian distributions. The larger modes could be artifacts derived from the fixed bandwidth density estimate used. This statement suggests that a variable KDE would be particularly useful to analyze these data. As expected, our results with a variable bandwidth estimator minimize the effect of the largest few values, emphasizing five main modes and only a

small mode towards larger thickness values. Note that the PISJ bandwidth produced a similar value to those from the Silverman's test or the lowest bandwidth from BCV, but at a much lower computational cost. These estimations suggest one small potential mode between the two largest modes found by Minnotte and Scott (1993) and clearly indicated by the Marron and Udina (1998) procedure. The OS bandwidth points out the multimodal nature of the data. As has been noted by other authors (Scott 1992; Salgado-Ugarte et al. 1997), the optimal bandwidth cannot show adequately the several modes in the distribution.

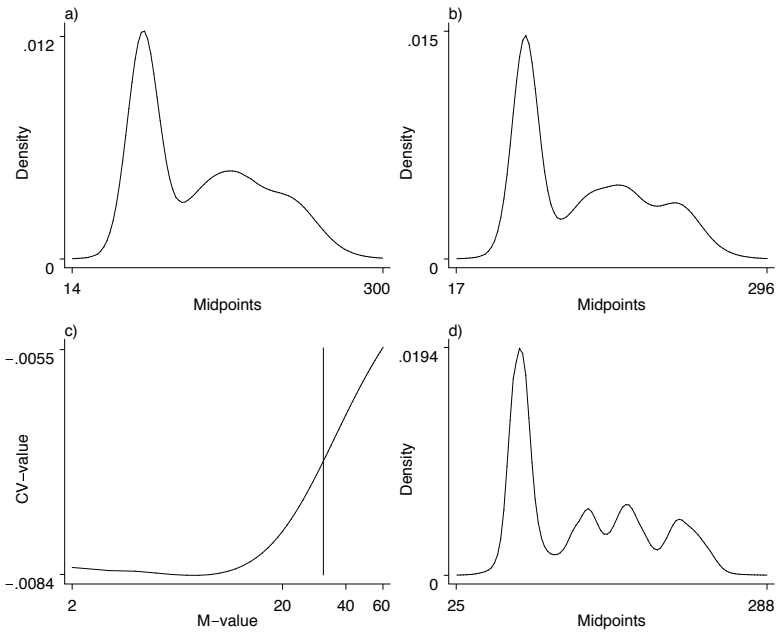

Figure 7: Catfish data: a) Oversmoothed VBKDE (G-mean of h = 15.6); b) Optimal VBKDE (G-mean of h = 12.2); c) LSCV score function. Minimum is located at M = 8, indicating a bandwidth h = 4; the line to the right represents the OS bandwidth (M = 31.26; h = 15.63); d) VBKDE with the LSCV bandwidth (G-mean of h = 4)

(*Continued on next page*)

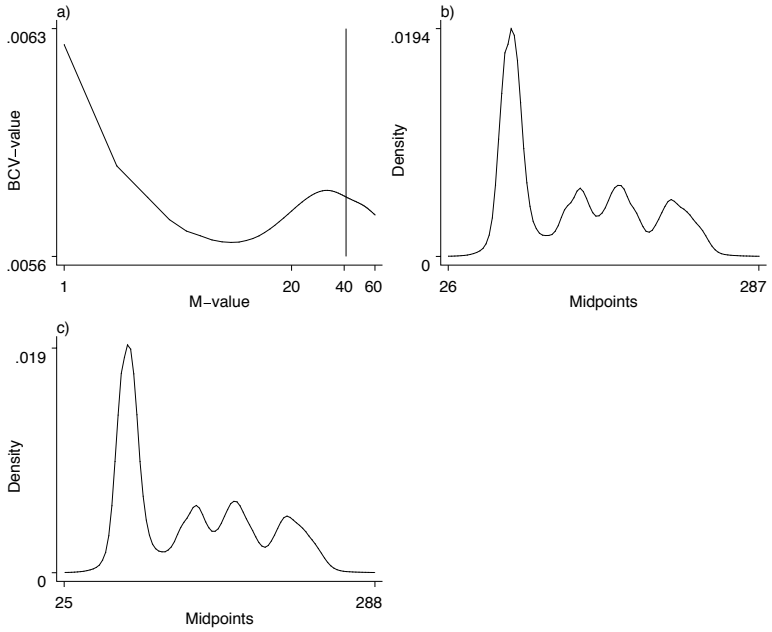

Figure 8: Catfish data: a) BCV score function. Vertical line on the right is the OS bandwidth (M = h = 41). The minimum is located at M = h = 9 (Quartic kernel); b) VBKDE with the BCV bandwidth rescaled to a Gaussian kernel (G-mean of h = 3.43); c) VBKDE with the SJPI bandwidth (G-mean of h = 3.9)

5.3 Example 3. Galaxy velocities data

The obtained bands range from 600 (LSCV) to 4762.5 (BCV), with five and one mode, respectively. For these data, Röeder (1990) developed a semiparametric estimate for the density based on the election of the Gaussian mixture that maximizes a nonparametric goodness-of-fit measure that is a function of the sampling spaces. This method produces a punctual estimate and a confidence set of plausible densities. This bandwidth applied to the galaxies velocity data produced a punctual estimate with five modes, but with no less than three or more than seven.

In this example, the OS bandwidth results in a main central mode and indications of one mode towards the left and another towards the right; the BCV and the Silverman test oversmoothed the data to the point in which any detail of the mode around 22,500 (as detected by the other methods) was lost.

The LSCV and SJPI gave very similar estimations, each one with four modes. Therefore, it seems that the OS, LSCV, and SJPI properly represent this data distribution. Silverman's bandwidth (as reported by Röeder (1990) produced a conservative result ((940 + 2500)/2 = 1720) that with the variable bandwidth KDE produced only two modes.

5.4 Example 4: Catfish data

The bandwidths ranged from 3.43 (BCV) to 15.6 (OS). This dataset has been used as an example for fixed width KDE application to detect cohorts and estimate fish growth. In this respect, it is worthwhile to note that the sample size is the largest from the reviewed datasets and that this large n resulted in stable bandwidth selection calculations. The OS bandwidth strongly suggests the existence of multimodality, a feature more clearly defined by the optimal bandwidth, indicating that this value is too wide to really represent the distribution of the data. The other bandwidths attained very similar magnitudes and lead to the presence of four modes. The coincidence between LSCV and BCV is a fact noted by Scott (1992), and in this case, the SJPI bandwidth coincides too. The Silverman test suggests a bandwidth compatible with the hypothesis of four modes in the distribution. This number of density maximum is obtained with bandwidths from 3.93 to 12.34, therefore, and in accordance with the above quoted results, we could choose 4 as the bandwidth representing data distribution.

6 Conclusions

As indicated in other similar analysis (for example, Sheather 1992), not knowing the true underlying distribution form does not permit one to make definitive statements on the distributions. Nevertheless, with the presented examples, it is possible to state several generalities.

Working with several tens or a few hundreds of observations, the variability of LSCV bandwidth seriously limits its practical use. However, it seems that as the number of observations is increased (thousands as in the catfish length dataset), this value becomes more stable. The optimal Silverman bandwidth, in general, oversmooths the densities with multimodality indication. The SJPI bandwidth was slightly better in this respect. The BCV oversmoothed multimodal densities, too.

As expected, in several cases, the variable bandwidth KDE estimates showed less modes than the number sustained by the Silverman's test. This is particularly true for the datasets having few dominant modes and some groups with few observations.

It is expected that a concordance among bandwidth selection methods working with high n values (in the order of thousands) will occur, even with multimodal distributions, although with modes represented with more or less equal number of observations.

Working with simulated distributions, Jones et al. (1996) found that the SJPI considered here had the best performance. Therefore, more work is needed to propose criteria for comparison of different bandwidth selectors by means of Monte Carlo simulations. In either case, the performance of the SJPI in all the examples is impressive, as sometimes it achieves the same result as the Silverman test, considered a very sophisticated and safe procedure to estimate the number of modes in univariate distributions.

It would be interesting to implement a variation of the Silverman test that uses the variable bandwidth KDE. More research on the subject is warranted.

7 Acknowledgments

The first author was supported by DGAPA-PAPIIT IN217596, and DGAPA-FES Zaragoza PAPIME 192031 at the Universidad Nacional Autónoma de México. Additional support was provided by the Department of Biology at Universidad Autónoma Metropolitana Iztapalapa during stays as visiting and titular professor (Dr. Ramón Riba y Nava Esparza professorship).

8 References

Bowman, A. W. and A. Azzalini. 1997. *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations*, vol. 18. Oxford, UK: Oxford University Press.

Fox, J. 1990. Describing univariate distributions. In *Modern Methods of Data Analysis*, eds. J. Fox and J. S. Long, 58–125. Newbury Park, CA: Sage Publications.

Härdle, W. 1991. *Smoothing Techniques with Implementation in S*. New York: Springer.

Härdle, W., S. Klinke, and B. A. Turlach. 1995. *Xplore: An Interactive Statistical Computing Environment*. New York: Springer.

Izenman, A. J. and C. Sommer. 1988. Philatelic mixtures and multimodal densities. *Journal of the American Statistical Association* 83(404): 941–953.

Jones, M. C., J. S. Marron, and S. J. Sheather. 1996. Progress in data-based bandwidth selection for kernel density estimation. *Computational Statistics* 11: 337–381.

Marron, J. S. and F. Udina. 1998. *Interactive local bandwidth choice*. Unpublished manuscript. Available at *http://libiya.upf.es/*.

Minnotte, M. C. and D. W. Scott. 1993. The mode tree: a tool for visualization of nonparametric density features. *Journal of Computational and Graphical Statistics* 2: 51–68.

Park, B. U. and J. S. Marron. 1990. Comparison of data-driven bandwidth selectors. *Journal of the American Statistical Association* 85: 66–72.

Postman, M., J. P. Huchra, and M. J. Geller. 1986. Probes of large-scale structures in the Corona Borealis region. *The Astronomical Journal* 92: 1238–1247.

Röeder, K. 1990. Density estimation with confidence sets exemplified by superclusters and voids in the galaxies. *Journal of the American Statistical Association* 85(411): 617–624.

Salgado-Ugarte, I. H., M. Shimizu, and T. Taniuchi. 1993. snp6: Exploring the shape of univariate data using kernel density estimators. *Stata Technical Bulletin* 16: 8–19. In *Stata Technical Bulletin Reprints*, vol. 3, 155–173. College Station, TX: Stata Press.

—. 1995a. snp6.1: ASH, WARPing, and kernel density estimation for univariate data. *Stata Technical Bulletin* 26: 23–31. In *Stata Technical Bulletin Reprints*, vol. 5, 161–172. College Station, TX: Stata Press.

—. 1995b. snp6.2: Practical rules for bandwidth selection in univariate density estimation. *Stata Technical Bulletin* 27: 5–19. In *Stata Technical Bulletin Reprints*, vol. 5, 172–190. College Station, TX: Stata Press.

—. 1997. snp13: Nonparametric assessment of multimodality for univariate data. *Stata Technical Bulletin* 38: 27–35. In *Stata Technical Bulletin Reprints*, vol. 7, 232–243. College Station, TX: Stata Press.

Scott, D. W. 1992. *Multivariate Density Estimation: Theory, Practice, and Visualization*. New York: John Wiley & Sons.

Sheather, S. J. 1992. The performance of six popular bandwidth selection methods on some real data sets. *Computational Statistics* 7: 225–250.

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

Silverman, B. W. 1986. *Density Estimation for Statistics and Data Analysis*. Monographs on Statistics and Applied Probability, London: Chapman & Hall.

Simonoff, J. S. 1996. *Smoothing Methods in Statistics*. New York: Springer.

Udina, F. 1995. *Interactive graphics for kernel density estimation*. Unpublished manuscript. Available at *http://libiya.upf.es/*.

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

Weisberg, S. 1980. *Applied Linear Regression*. New York: John Wiley & Sons.

Wolfe, J. H. 1970. Pattern clustering by multivariate mixture analysis. *Multivariate Behavioral Research* 5: 329–350.

About the Authors

Isaías H. Salgado-Ugarte is titular professor at the Biology Departments of Universidad Nacional Autónoma de México (FES Zaragoza) and Universidad Autónoma Metropolitana (Iztapalapa). He has been working on the application of statistical methods, especially nonparametric smoothing (kernel density and regression), and multivariate methods to the analysis of biological data.

Marco A. Pérez-Hernández is titular professor at the Biology Department of the Universidad Autónoma Metropolitana Iztapalapa Campus. His interests include ecosystem structure, fish communities diversity, distribution and abundance, and coastal lagoon habitat characterization.