Drawbacks in classic methods for multivariate mode detection problems

Susana Vegas-Azcárate
Statistics and Decision Sciences Group
Rey Juan Carlos University
28933 Móstoles, Spain
susana.vegas@urjc.es

Jorge Muruzábal
Statistics and Decision Sciences Group
Rey Juan Carlos University
28933 Móstoles, Spain
jorge.muruzabal@urjc.es

Abstract - *This paper shows how the SOM structure derived from a topographic map formation algorithm provides a powerful approach to multivariate mode detection. Three classic approaches to mode detection are studied in detail first. The bkde method for univariate data and the bkde2D and kde2d methods for bivariate data.*

Key words - self-organizing maps, mode detection

1 Introduction

The data sets considered in the univariate case are some of the Gaussian mixture densities developed by Marron and Wand (1992). The strongly skewed density (Figure 1b top) departs in the direction of skewness and was chosen to resemble to log-normal. The bimodal (Figure 1c top), skewed bimodal (Figure 2a top) and trimodal (Figure 2b top) densities are mildly multimodal and one expect to estimate them fairly well with a data set of moderate size. The claw density (Figure 2c top) is an interesting strongly multimodal density, and will be very hard to recover in full with classic methods. The smooth and discrete comb densities (Figures 3a,b top), are enhancements of the basic idea of the bimodal density. The data sets used in the bivariate case are our generalizations of the claw, discrete comb, smooth comb and strongly skewed densities.

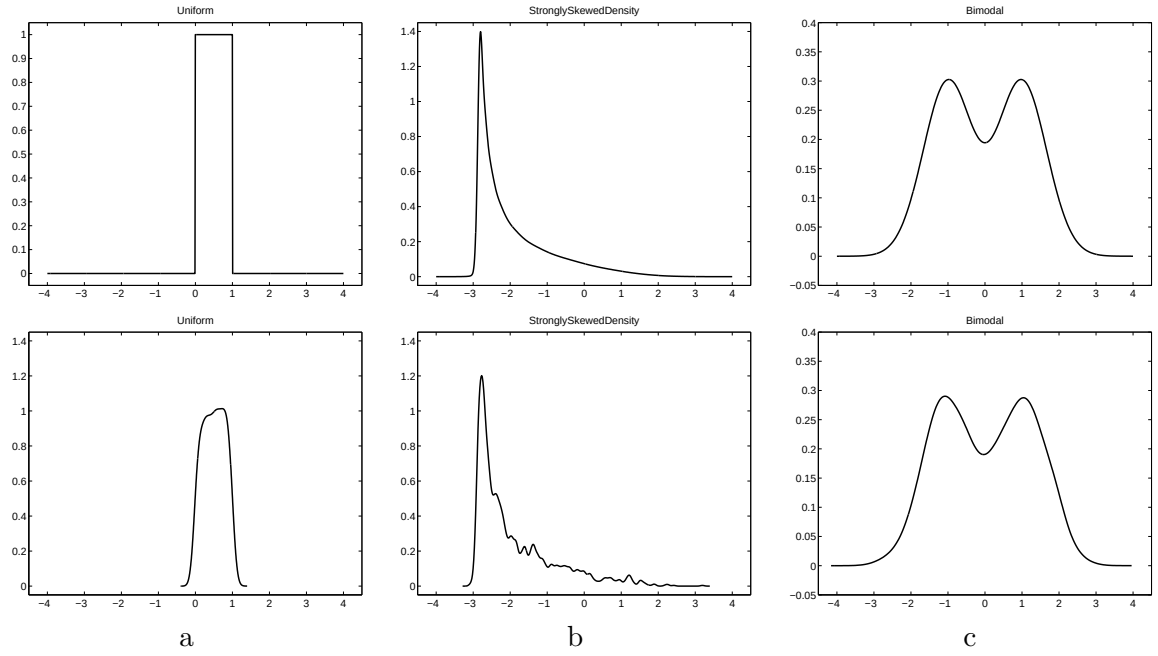

FIG. 1 – Marron and Wand data sets. Original (top) bkde (bottom). (a) Uniform ; (b) StronglySkewedDensity ; (c) Bimodal.

1.1 The *bkde* method

Wand and Jones (1995) have developed a binned approximation to the ordinary kernel density estimate, the *bkde* method for univariate data. Figures 1, 2 and 3 show how this method copes with Marron and Wand (1992) data sets. Good results are obtained using the uniform density. On the contrary, the unimodal strongly skewed density appears to be mildly multimodal, since *bkde* method is not able to reflect appropriately the smoothness of this density.

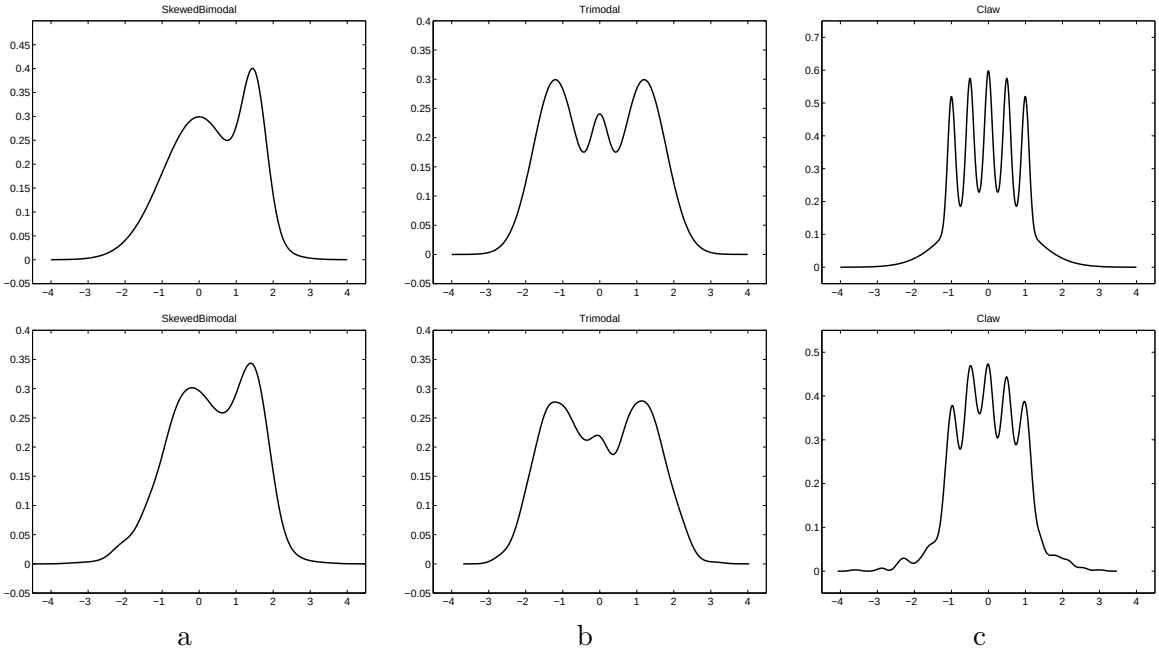

FIG. 2 – Marron and Wand data sets. Original (top) bkde (bottom). (a) SkewedBimodal ; (b) Trimodal ; (c) Claw.

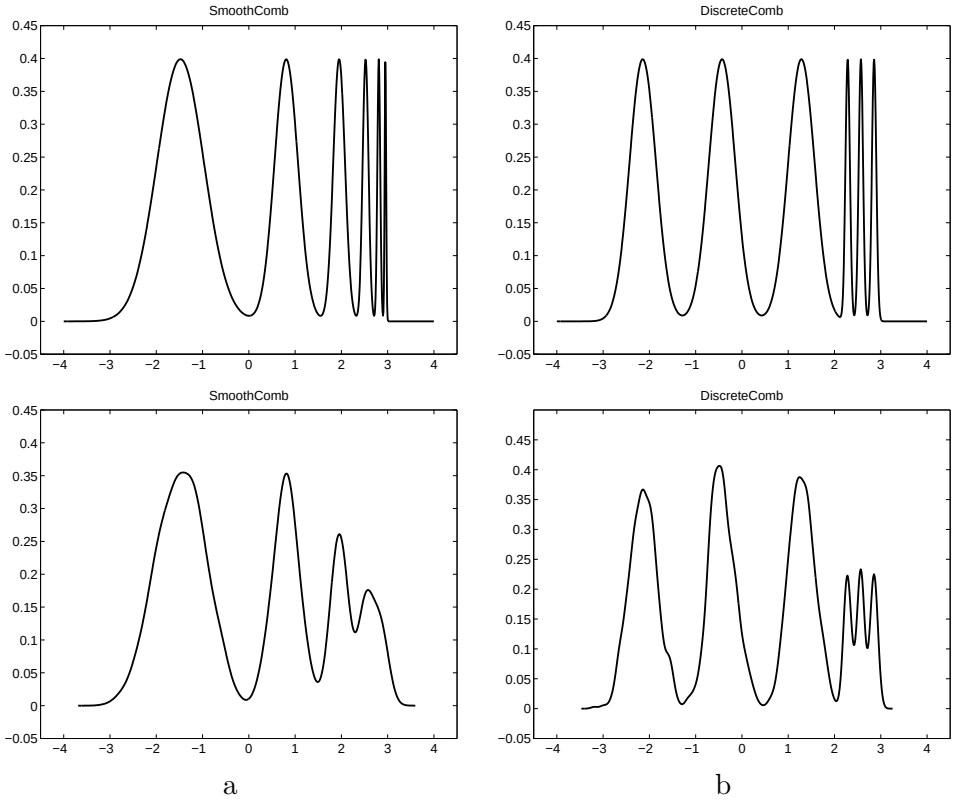

FIG. 3 – Marron and Wand data sets. Original (top) bkde (bottom). (a) SmoothComb ; (b) DiscreteComb.

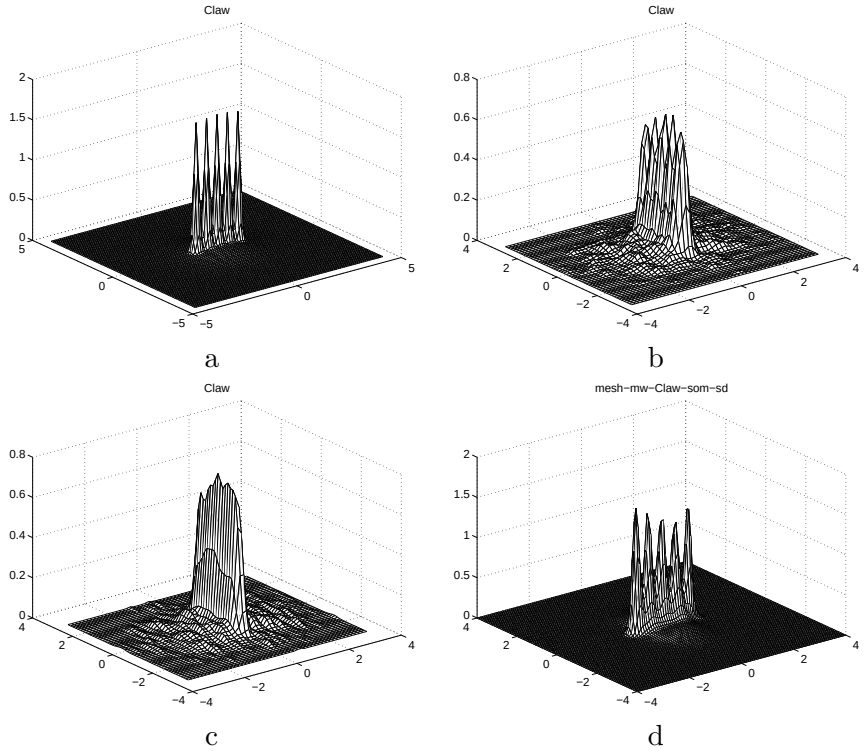

FIG. 4 – Bivariate claw density. (a) Original ; (b) bkde2D ; (c) kde2d ; (d) Kohonen's map.

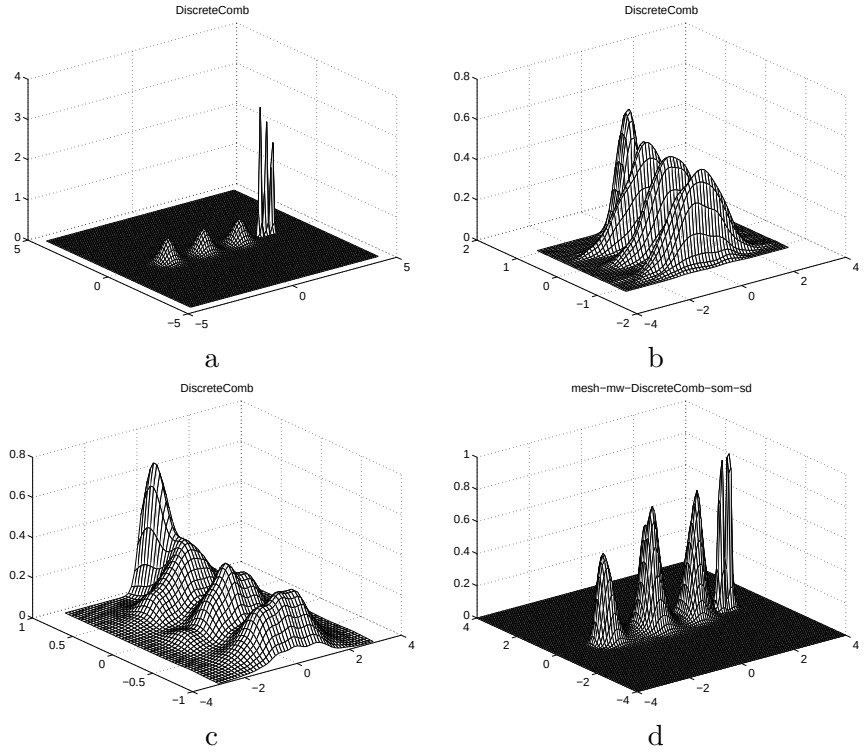

FIG. 5 – Bivariate discrete comb density. (a) Original ; (b) bkde2D ; (c) kde2d ; (d) Kohonen's map.

The two modes of the bimodal density are well placed, and also in the skewed bimodal one. The middle mode of the trimodal density in not well reflected, since small changes in the bandwidths will make it disappear. The same happens in the claw density, were the choice of the bandwidths values is crucial. The *bkde* method is able to catch 4 out of 5 modes in the smooth comb density. Although it is able to place the 6 modes of the discrete comb density, the last three components of this mixture are not properly weighted.

1.2 The *bkde*2*D* and *kde*2*d* methods

A bivariate extension of the *bkde* method was developed by Wand (1994); Wand and Jones (1995). The *bkde*2*D* method is the binned approximation to the 2D kernel density estimate. The kernel is the standard bivariate Gaussian density. Recently, Venables and Ripley (2002) has developed another bivariate method, the *kde*2*d* method. This is a two-dimensional kernel density estimation with an axis-aligned bivariate Gaussian kernel evaluated on a square grid. Both methods are studied here. Their behaviour on some bivariate data sets is analyzed and compared to the results obtained with the Kohonen's learning rule.

Figure 4 shows the bivariate claw density. The five modes of this density are well placed by the *bkde*2*D* method, but the lack of smooth of the resulting estimated density provides many local maxima that do not correspond to real modes. The *kde*2*d* method, besides the many false positives in the low density regions of the estimated density, it is not able to catch a single real mode. On the contrary, the SOM's estimated density resembles the real density. The five mode are accurately placed and no false local maxima appear in the low density regions.

The bivariate discrete comb density is illustrated in Figure 5. *bkde*2*D* method highlights the first three components of the mixture, but not properly weighted. Moreover, the last three components are mixed into one in the estimated density. In *kde*2*d* estimated density, the two first components are badly weighted, showing also local maxima that do not correspond to the real modes. Further, the last four components are mixed into one, showing also false positives. On the other hand, the SOM's estimated density is able to show the first three components with its real weight (note the difference in z-scale between Figures 5a and 5d) and with the appropriate smoothness. Moreover, each of the first three modes are well placed. Furthermore, the SOM algorithm is able to distinguish between two of the three last components, placing properly their modes and catching their smoothness. However, the SOM algorithm misses one component.

The bivariate smooth comb density is shown in Figure 6. In *bkde*2*D* estimated density, the two first components are shown, but wrong weighted. The third and fourth components are mixed into one, which shows two local maxima. The fifth component is missed with *bkde*2*D* method. In *kde*2*d* estimated density appears a variety of local maxima that do not corresponds to any of the real modes. Even the first and easier component is not properly showed with *kde*2*d* method. On the contrary, the SOM structure is able to imitate the real density. All components but the last one are properly placed and weighted in the SOM's estimated density.

Figure 7 presents the bivariate strongly skewed density. *bkde*2*D* estimated density is well placed but badly weighted. *kde*2*d* estimated density, besides that it is wrongly weighted, shows many local maxima in the low density regions. On the other hand, the SOM algorithm is able to reproduce accurately the real density.

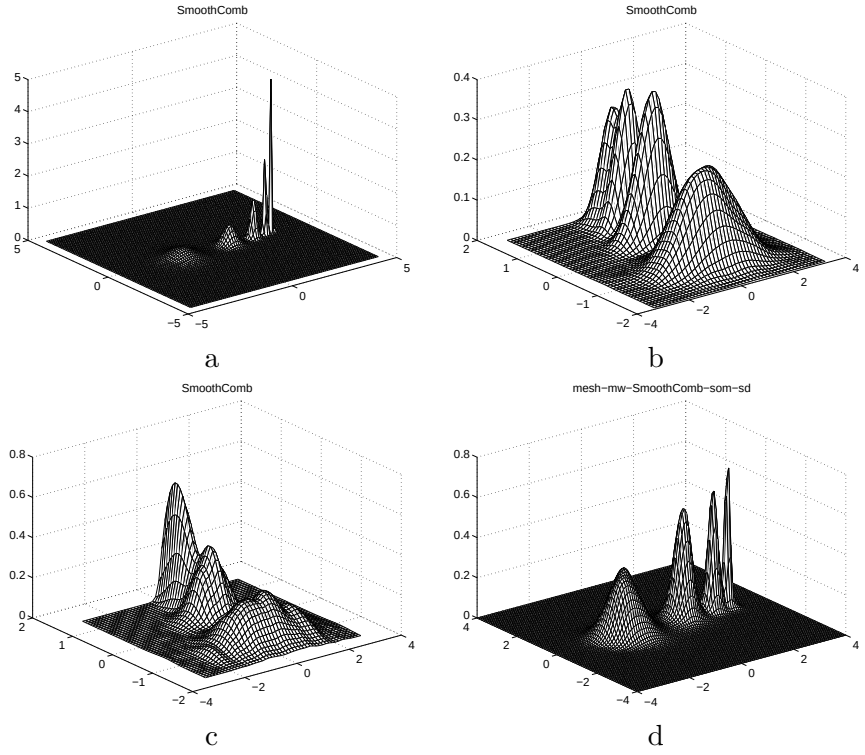

FIG. 6 – Bivariate smooth comb density. (a) Original ; (b) bkde2D ; (c) kde2d ; (d) Kohonen's map.

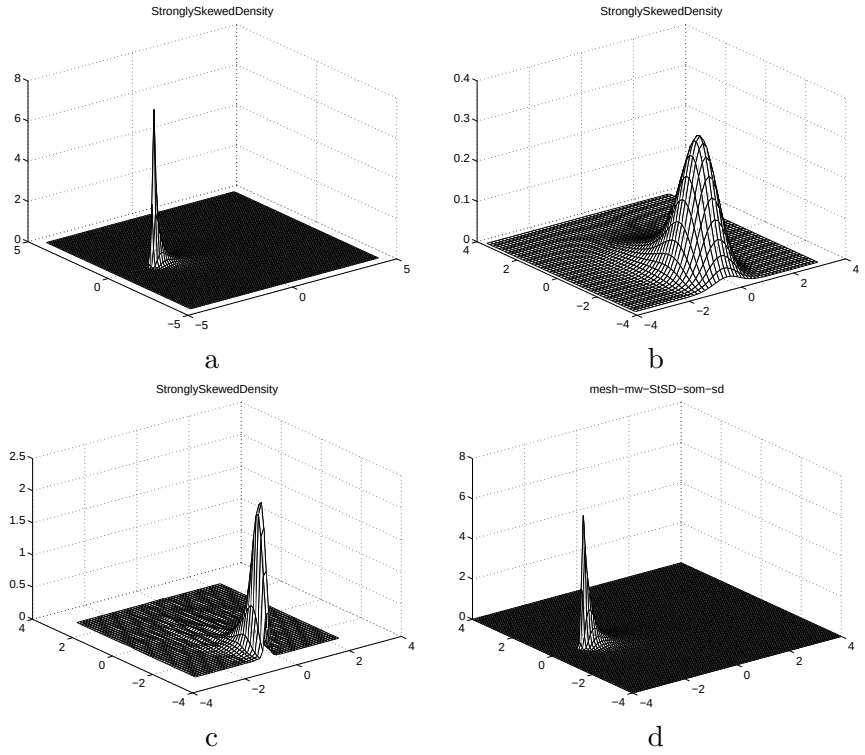

FIG. 7 – Bivariate strongly skewed density. (a) Original ; (b) bkde2D ; (c) kde2d ; (d) Kohonen's map.

2 Conclusions

The main drawbacks in the previously studied methods are the following. First, the choice of the bandwidths. These methods develop density estimation as a smoothing operation ; hence, there is a trade-off between bias in the estimate and the estimate's variability. In other words, large bandwidths will produce smooth estimates that may hide local features of the density, while small bandwidths may introduce spurious bumps into the estimate. These bandwidths are usually chosen ad hoc by the user. Moreover, when dealing with bivariate data the difficulties in choosing the appropriate bandwidth values increase.

Second, the difficulties in the post-processing of the outputs. *bkde* method returns the coordinates of the binned kernel density estimate of the probability density of the data, while *bkde*2*D* method returns the set of grid points in each coordinate direction, and the matrix of density estimates over the mesh induced by the grid points. This way, modes' location in the *bkde* and *bkde*2*D* estimated densities can only be assessed graphically, since the coordinates of the binned kernel and the set of grid points do not involve data directly.

And third, how to jump to higher dimensions ? A complete methodology on density estimation and mode detection in higher dimension is not established yet

The SOM structure is an excellent alternative to classic methods since, it provides better results even with bivariate data, it has no bandwidths to select before hand (SOM's crucial parameter is the neighbourhood range and the way it is decreased during training, and it can be automatically monitorized [Vegas-Azcárate and Muruzábal (2005); Muruzábal and Vegas-Azcárate (2005)]), the output post-processing possibilities are enormous (due to the topological relationship between neurons and data inputs) and, finally, the jump to higher dimensions is straightforward (the SOM structure has long been used for multivariate data visualization, clustering and classification).

Références

Marron, J. and Wand, M. (1992). Exact mean integrated squared error, *The Annals of Statistics*, **20** : 712–736.

Muruzábal, J. and Vegas-Azcárate, S. (2005). On equiprobabilistic maps and plausible density estimation, in *5th Workshop On Self-Organizing Maps, Paris*.

Vegas-Azcárate, S. and Muruzábal, J. (2005). On cluster analysis via neuron proximity in monitored self-organizing maps, in *Workshop on Biosignal Processing and Classification, Barcelona, Spain*.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S.*, Springer, Fourth edition.

Wand, M. and Jones, M. (1995). *Kernel smoothing*, London : Chapman and Hall.

Wand, M. P. (1994). Fast computation of multivariate kernel estimators, *Journal of Computational and Graphical Statistics*, **3** : 433–445.