A SHORT INTRODUCTION TO NONPARAMETRIC CURVE ESTIMATION

Ricardo Cao
Universidade da Coruña
Spain

Some ideas about how basic aspects of nonparametric curve estimation can be introduced to students at a post secondary level will be discussed here. The idea of estimating population curves, like the density or the regression function, is studied from a nonparametric viewpoint. Starting from well-known estimators as the histogram or the regressogram, the discussion will then go to some of the smoothing methods developed in the last four decades, mainly focusing on the kernel density and regression estimators. Some ideas about the important problem of smoothing parameter selection will also be presented.

PARAMETRIC VERSUS NONPARAMETRIC FITS IN CURVE ESTIMATION

One of the risks after a first course in a statistics at a post secondary level is that some of the students start believing that every continuous population has a normal distribution. Using very well-known real data sets, like the turtle data or the Old Faithful geyser data in Silverman (1986) it is quite easy to make a student understand that there are plenty of real examples in which the normal and also some other classical parametric families are not reasonable (see Figure 1).

As a possible consequence of this lack-of-fit "problem" of parametric models in real world examples a large number of nonparametric techniques have arose during the last forty years. These nonparametric procedures estimate population curves without assuming any particular parametric form. One may think of estimating the probability density function in the examples mentioned above by some nonparametric density estimator, like a simple histogram or the more elaborated kernel method, for instance.

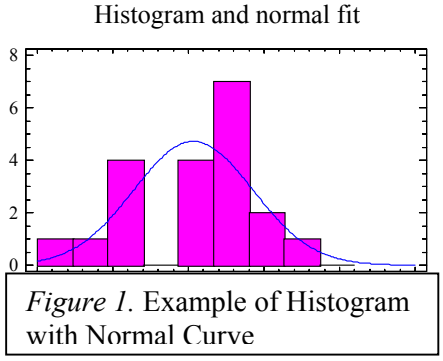

Figure 1. Example of Histogram with Normal Curve

Similar facts appear when dealing with different probability curves, as the regression function. After an intuitive introduction of the concept of regression via the conditional expectation, several examples can be introduced to see how unreasonable the classical linear assumption for the regression function may be. Some of these examples could be the motorcycle data and the coronary risk-factor study data (see, for instance, Fan & Gijbels, 1996).

Although the nonparametric fits exhibit several benefits when compared to parametric fits and have been widely used as exploratory techniques, it is also important to point out, from the beginning, two of the practical problems of the nonparametric techniques: the smoothing parameter selection and the curse of dimensionality.

FROM THE HISTOGRAM TO THE KERNEL DENSITY ESTIMATOR

The key point to explain how the nonparametric estimators work for the probability density function is to focus first on the definition of this curve. The limit concept of ratio between probability mass in a neighbourhood of a point and the "size" of that neighbourhood plays an important role when explaining what a simple histogram is doing.

This idea can be easily extended to obtain the moving histogram (or naive estimator) and, finally, the kernel density estimator by just giving different weights to the data according to their proximity to the interest point. A mathematical expression for the well-known Parzen-Rosenblatt kernel density estimator can be then introduced:

$$\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right).$$

Its practical performance can be explored by using it with real data coming from some of the examples mentioned above. In particular, the great influence of the bandwidth, or smoothing parameter, and the small influence of the kernel function are easily shown when viewing some plots of the estimator applied to the data (see Figure 2).

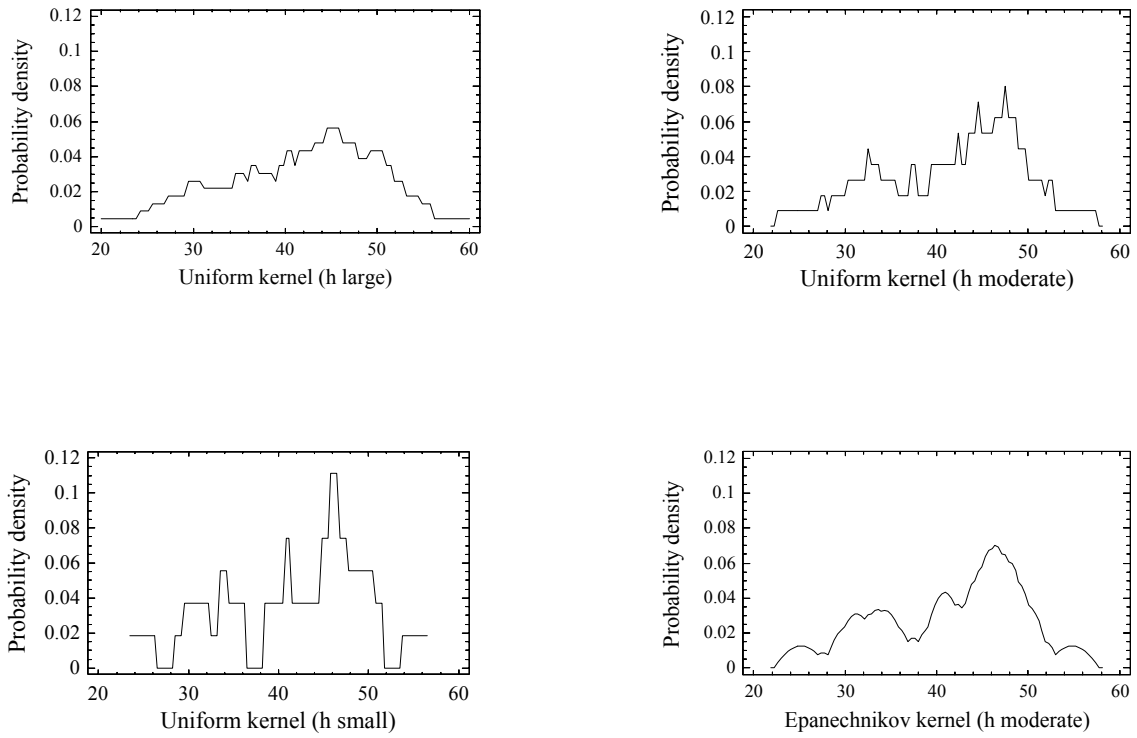

Figure 2. Estimator Plots.

Some standard problems, as the undersmoothing and oversmoothing effects caused by the increasing of variance and bias of the estimator, can also be presented from a practical viewpoint using some real example. The asymptotic expressions for the bias and the variance of the estimator give an easy explanation of the crucial role of the amount of smoothing used in the estimation process.

QUICK IDEAS ABOUT OTHER NONPARAMETRIC APPROACHES

Still in the context of density estimation, some other nonparametric techniques are available. We will just give some intuitive ideas about how several of these nonparametric density estimators deal with the problem. The nearest neighbour estimator will be introduced by recalling the relationship between the length of the interval in which the k nearest neighbour to a point lay and the value of the probability density at that point. When regarding the density in a functional vector space, the orthogonal series estimator appears as some finite linear combination approximation to the infinite linear combination representation of the density. The splines estimator will be introduced as an attempt to use piecewise smooth polynomial functions for estimating the density. Some practical application of these estimators could be useful to compare their performance with the histogram or the kernel density estimator.

KERNEL AND LOCAL POLYNOMIAL ESTIMATORS FOR REGRESSION

The kernel estimator can be easily introduced in the nonparametric regression context via the regressogram, see Figure 3. This is parallel to what happens when going from the histogram to the kernel estimator in the density case. The undersmoothing and oversmoothing problems, as well as some introductory bias and variance asymptotic expressions will be presented in a completely analogous way as done for the density set-up. It is also worth mentioning that the kernel density estimator is the least squares local linear constant estimator.

The previous property can be taken as a definition in order to extend local constant estimators to the so-called local polynomial regression estimators. The behaviour of these estimators will be illustrated by means of the motion picture analysis of smoothing (see Marron, Ruppert, Smith & Conley, 2000).

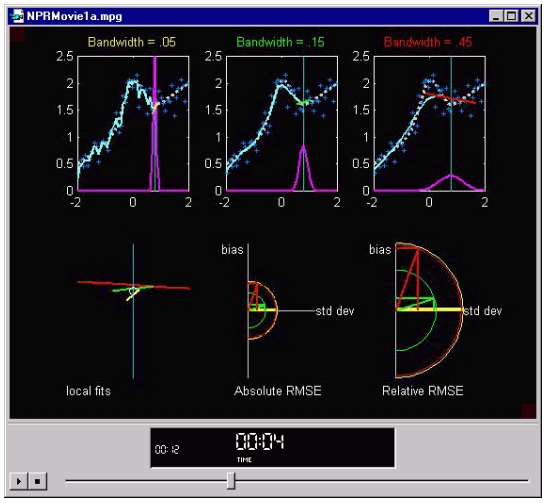

Figure 3. Regressogram.

HOW TO CHOOSE THE SMOOTHING PARAMETER?

After having pointed out how important is the bandwidth selection problem in curve estimation, some available automatic bandwidth selectors will be presented for the density case. To do this, we will sketch some of the methods in the comparative study by Cao, Cuevas and González-Manteiga (1994). By looking at the plot of the MISE (mean integrated squared error), in figure 4 in a simulated context, as a function of the smoothing parameter, it is clear that any reasonable bandwidth selector has to take into account the mentioned trade-off between variance (or stochastic error) and squared bias (or deterministic error).

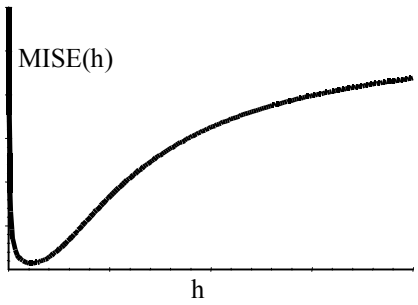

Figure 4. Plot of the MISE.

We first introduce, in a natural way, first-generation plug-in bandwidth selectors which essentially estimate the unknown (population) quantities appearing in the asymptotic formula for the mean integrated squared error (AMISE). Also much improved versions of these plug-in ideas (like the Sheather and Jones selector will be briefly introduced).

Starting from the expression of the integrated squared error (ISE) it is easy to derive the form of the least squares cross-validation function to be minimised to find the pertaining bandwidth selector. More refined and accurate methods based on the same idea, as the smoothed cross-validation, will be presented too.

Finally, some quick ideas about how the bootstrap can be used for the bandwidth selection problem will be given. These bootstrap bandwidth selectors are obtained by minimising some bootstrap version of MISE. One important feature that should be mentioned is the fact that closed expressions are available for these bootstrap versions, without any need of performing Monte Carlo simulations.

DISCUSSION

The lack-of-fit problems of many parametric models as well as the flexibility of application of nonparametric techniques as exploratory tools have made of nonparametric curve estimation procedures a very active research field in statistics. From the teaching statistics

viewpoint there exists the need of including more and more of these new techniques in the content of statistical courses at a post-secondary level. As an attempt to do it, a proposal is made in order to get the students on these concepts in an intuitive and not much formalistic way. Of course, some mathematical tools are needed, but our proposal mainly focuses on concepts, ideas and graphical visualisation.

REFERENCES

Cao, R., Cuevas, A., & González-Manteiga, W. (1994). A comparative study of several smoothing methods in density estimation. *Computational Statistics and Data Analysis, 17,* 153-176.

Fan, J., & Gijbels, I. (1996). *Local polynomial modelling and its applications*. London: Chapman & Hall.

Marron, J.S., Ruppert, D., Smith, E.K., & Conley, G. (2000). Motion picture analysis of smoothing. North Carolina Institute of Statistics, Mimeo Series #2367.

Silverman, B.W. (1986). *Density estimation for statistics and data analysis*. London: Chapman & Hall.