

Variable Kernel Density Estimate: a new proposal

Stima di densità con il metodo del nucleo a parametro di liscio variabile: una nuova proposta

Giovanna Menardi

Dipartimento di Scienze Statistiche, Università degli Studi di Padova

e-mail: Giovanna.Menardi@econ.units.it

Riassunto: Nel lavoro si considera il problema della stima di densità con il metodo del nucleo a parametro di liscio variabile. Si propone una tecnica per determinare la costante di liscio basata sulla minimizzazione del MSE. Inoltre, poiché, in generale, i metodi che si basano sulla minimizzazione del MSE non permettono di definire la costante di liscio nei punti di flesso della densità da stimare, si propone una soluzione a tale problema.

Keywords: balloon estimator, kernel estimator, variable bandwidth.

1. Introduction

The univariate fixed bandwidth kernel estimator can be defined as follows:

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^n \frac{1}{h} K\left(\frac{x - x_i}{h}\right) \quad (1)$$

where $x_i, i = 1, \dots, n$ is a sample of i.i.d. observations from some unknown density f and the kernel K is taken to be a symmetric, nonnegative function centered at zero and integrating to one. Techniques for choosing the smoothing parameter have achieved a certain level of maturity (see e.g. Silverman, 1986) and fixed bandwidth kernel estimators work well for most regular densities. Though, when data exhibit multimodality or show strong skewness, they fail to respond appropriately, because they cannot cope with large variations in the magnitude of $f(x)$. A large bandwidth tends to oversmooth and so hide features by de-emphasizing or wiping out significant modes. On the other hand, a small bandwidth may introduce spurious peaks by undersmoothing the real density. For this reason there are proposals to vary the amount of smoothing. Intuitively, more smoothing is necessary where data are scarce and less smoothing is necessary around the modes.

The majority of the works concerning adaptive kernel density estimation can be grouped into two categories: *balloon* estimators and *sample point* estimators. The distinction between the two lies in how the bandwidth is varied. For balloon estimators, a different smoothing parameter is selected for each estimation point x . The estimate of f at x is an average of identically scaled kernels centered at each data point. A new value of the smoothing parameter is chosen to scale the kernels at an other estimation point. The form of the estimator is given by

$$\hat{f}_B(x) = \frac{1}{n} \sum_{i=1}^n \frac{1}{h(x)} K\left(\frac{x - x_i}{h(x)}\right) \quad (2)$$

The sample point estimator uses a different bandwidth for each data point and is given by

$$\hat{f}_{SP}(x) = \frac{1}{n} \sum_{i=1}^n \frac{1}{h(x_i)} K\left(\frac{x - x_i}{h(x_i)}\right). \quad (3)$$

The estimation of f at x is then an average of differently scaled kernels centered at each observation. We shall note that balloon estimators are pointwise and let a straightforward analysis because at each x they behave as fixed kernel estimators, but when considered as global estimators, they fail to integrate to one. Sample point estimators, on the other hand, are themselves densities but they suffer from a phenomenon referred as *nonlocality*, *i.e.* the estimate is strongly affected by data very far away from the estimation point.

In this paper we propose a technique to choose automatically the smoothing parameters of a balloon estimator. Since every continuous probability distribution can be well approximated by a mixture of Gaussian densities, the selection of the bandwidth at a point x is based on the minimization of the mean square error (MSE) after that a preliminary estimate of the components of the mixture is done. This technique is described in Section 2 and some asymptotic arguments are given to prove its validity. In Section 3 the results of a simulation study are reported. Some suggestions about how to proceed with the work are proposed in the final section.

2. Bandwidths selection

According to the MSE criterion, the form of the optimal $h(x)$ in (2) is, asymptotically,

$$h(x) = \left[\frac{R(K)f(x)}{n\mu_2(K)^2 f''(x)^2} \right]^{1/5} \quad (4)$$

for a second order kernel K . Here $R(K) = \int K^2(u)du$, $\mu_2(K) = \int u^2 K(u)du$, and $f''(x)$ is the second derivative of f at x . Two problems arise from the direct use of (4) to choose a good bandwidth: the expression contains the terms $f(x)$ and $f''(x)$ that are unknown and it is not defined at the points where $f''(x) = 0$. In this paper, solutions are proposed for both of these problems.

Let $\Phi_\sigma(x - \mu)$ be the normal density at x centered at μ and scaled by σ . Since every continuous probability distribution can be well approximated by a mixture of Gaussian densities, we easily get the expression of the optimal bandwidth at a point x , if we assume that $f(x) \approx \sum_{j=1}^m p_j \Phi_{\sigma_j}(x - \mu_j)$ for some m, p_j, σ_j, μ_j .

To estimate the components of the mixture, a pilot fixed kernel estimate of the density is performed. The signs of the differences in the estimates at adjacent points give an indication about the location of the modes and antimodes of the underlying density. The modes are taken to estimate the μ_j while the antimodes are used to get the p_j 's and the variances. Let's call k_j the number of data points between the $j - 1^{th}$ and the j^{th} antimode. The p_j 's are estimated by k_j/n , while the σ_j 's are the standard deviations of the k_j nearest neighbours of $\hat{\mu}_j, j = 1, \dots, m$.

The $h(x)$'s are not defined at points where the second derivative of the estimated mixture is zero. Schucany (1989) has proposed to take into account further terms in the asymptotic expansion of MSE but, in our findings, even with this correction the optimal bandwidth is not defined in some points. Since these points are supposed to be different

Table 1: *MSE values ($\times 10^5$) estimated at some critical points of the density.*

	$n = 1000$			$n = 250$		
	fixed h	sample point	balloon	fixed h	sample point	balloon
$x = 0$	0.563	0.393	0.463	1.378	0.552	0.295
$x = 8$	6.658	24.153	1.474	14.719	24.304	2.179
$x = -3$	0.324	0.237	0.127	0.810	0.169	0.076
$x = 3$	0.346	0.274	0.120	0.780	0.613	0.638
$x = 7.5$	0.497	3.729	0.485	1.884	3.806	0.367
$x = 8.5$	0.545	4.105	0.480	1.988	4.176	0.339
$x = -9$	0.008	0.017	0.007	0.024	0.028	0.009
$x = 10$	0.028	1.190	0.006	0.310	1.159	0.015

from the inflection ones, our suggestion is to evaluate the bandwidth at every estimation point according to both methods and to use the smaller one.

3. Simulation study

A simulation study was conducted to evaluate the performance of the balloon estimator described in Section 2. We were mainly interested in understanding if the estimate could reproduce the feature of multimodality of the data and the height of the peaks.

Samples of different size were extracted from a set of multimodal distributions. For brevity, we present here only the results of a single simulation study, based on 500 samples of size 1000 and 250 from the mixture of normal densities, $0.9N(0, 9) + 0.1N(8, 0.25)$. An average of MSE was computed over the samples to get a measure of global error. Moreover the MSE was computed at some critical points, namely the modes, the inflection points and two points on the tails of the distribution. The performance of the balloon estimator were compared with the fixed kernel and the sample point estimator. The fixed bandwidth was selected according to the plug-in rule suggested by Scheater and Jones (1991), while the square root law proposed by Silverman (1986) was used as bandwidth function in the sample point estimator ($h(x_i) = h_0 d_i / d$ with $d_i = 1/\sqrt{\hat{f}(x_i)}$, $\hat{f}$ pilot

Figure 1: *From left: fixed kernel estimate (plug in bandwidth), sample point estimate (Silverman bandwidth function), balloon estimate of $0.9N(0, 9) + 0.1N(8, 0.25)$ for $n=1000$ (dashed lines). The solid line represents the true density.*

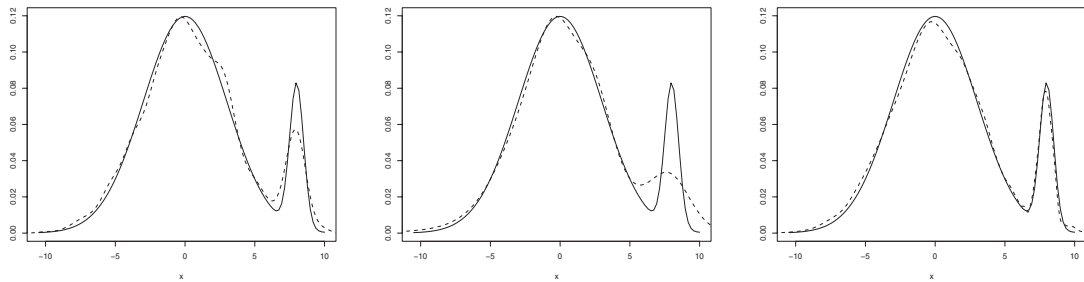
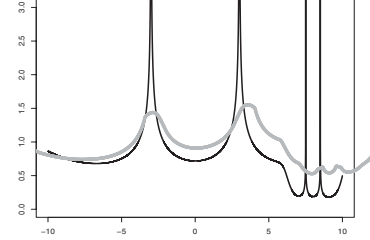


Table 2: Mean of the MSE ($\times 10^5$) over the samples of the fixed kernel density estimator, the Silverman sample point estimator and the balloon estimator described in Section 2

	fixed kernel	sample point	balloon
$n = 1000$	1.536	1.460	0.254
$n = 250$	1.183	1.487	0.334

Figure 2: Optimal (black line) and estimated (grey line) bandwidth function for density $0.9N(0, 9) + 0.1N(8, 0.25)$



estimate of f with bandwidth h_0 and d geometric mean of $d_i, i = 1, \dots, n$). The components of the mixture necessary to compute the $h(x)$ in (4) were estimated by using a pilot fixed kernel estimate of f with a smoothing parameter optimal for a Gaussian density. Anyway, results from a large simulation study have shown that the method is relatively insensitive to the choice of the pilot estimate. In all estimators a Gaussian kernel was used.

Some results are shown in Table 1 and 2. Both globally and locally the balloon estimator performs better than the fixed bandwidth and the sample point estimators. The solution proposed to get a local bandwidth in the inflection points is efficient and, differently from the other estimators, the balloon estimates do not oversmooth the density around the modes (see Figure 1). Moreover, the MSE is relatively low in the tails of the distribution.

4. Concluding remarks

The balloon estimator is a strong competitor against the standard kernel smoothing techniques when the data exhibit multimodality. The method could be successfully applied to problem of clustering or nonparametric discriminant analysis where it is not strictly necessary that the density estimates integrate to one. Further analysis is needed to investigate the validity of the method in estimating densities with features of irregularity different from multimodality. One point to consider is, however, whether the eventual advantage in using a variable window width is worth the computational effort necessary to calculate the bandwidth function.

References

- Schucany, W. R. (1989) Locally Optimal Window Widths for Kernel Density Estimation with Large Samples, *Statist. Prob. Letters*, 7, 401–405.
- Sheather, S. J. and Jones, M. C. (1991) A Reliable Data-based Bandwidth Selection Method for Kernel Density Estimation, *J. Roy. Statist. Soc. ser. B*, 53, 683–690.
- Silverman, B. W. (1986) *Density Estimation for Statistics and Data Analysis*, Chapman & Hall, London.