A Solve-the-Equation Approach for Unidimensional Data Kernel Bandwidth Selection

Luís A. Alexandre
Dep. Informatics and Instituto de Telecomunicações,
University of Beira Interior, Portugal
Technical Report, ISBN 978-989-654-005-0

December 29, 2008

Abstract

In this paper we present a new approach to the automatic determination of the bandwidth for kernel based density estimation in the 1-D case. The approach is based on the used of the plug-in estimator in the optimal AMISE estimator and an iterative algorithm. We present comparisons of our proposal on both artificial and real data sets against some of the more common bandwidth estimators. The values obtained are usually smaller than the ones obtained with other methods, yielding higher peaks and deeper troughs in the estimated densities.

1 Introduction

The problem of density estimation is fundamental for many applications [8]. In the class of non-parametric methods, kernel based ones are probably the most used. A central issue when using kernel based approaches to density estimation is the bandwidth value to use. Several methods have been proposed for the optimal kernel bandwidth, and following [8], we can divide the approaches into: rules of thumb, cross-validation and plug-in based methods. Our proposal fits in the plug-in methods.

In our case, the need for density estimation came when searching for the good estimates of the entropy of an error variable while using neural networks trained with the Error Entropy Minimization principle [1, 5, 10].

The optimal h from the point of view of minimizing the asymptotic mean integrated square error (AMISE) is [11]

$$h_{opt} = k_2^{-2/5} \left[\int K(t)^2 dt \right]^{1/5} \left[\int (f''(x))^2 dx \right]^{-1/5} n^{-1/5} \tag{1}$$

where $f(x)$ is the density, $K(t)$ is the kernel used and

$$k_2 = \int t^2 K(t) dt. \tag{2}$$

The problem with expression (1) is that it depends on the unknown density.

The idea in this work is to first replace the unknown density with its kernel based plug-in estimator and then use an iterative algorithm to provide automatic estimates of h_{opt}.

The difference between this approach and, e.g., the one by Sheather and Jones [9], which is one of the most used methods, lies in the fact that we do not assume that f is sufficiently smooth so that

$$R(f'') = \int (f''(x))^2 dx$$

is approximated by

$$-\int f^{(4)}(x) f(x) dx.$$

Instead, we use the plug-in estimator for the f'' directly and solve the $R(f'')$ for the case of a Gaussian kernel.

The rest of the paper is organized as follows: the next section contains our proposal for the kernel bandwidth estimator; section 3 presents the iterative algorithm used; the experiments are in section 4 and the final section contains the conclusions.

2 The proposed estimator

Let $X_1, X_2, \ldots, X_n$ be a sample of size n from a random variable with density f. Our density estimator is

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h} \right) \tag{3}$$

where $K(x)$ is the kernel used. We use the Gaussian kernel

$$K(t) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2} \right). \tag{4}$$

The estimator of the i-th derivative of the density is

$$\hat{f}^{(i)}(x) = \frac{1}{nh^{i+1}} \sum_{j=1}^{n} K^{(i)} \left(\frac{x - X_j}{h} \right) \tag{5}$$

where $K^{(i)}$ is the i-th derivative of the kernel.

Since the kernel used is the Gaussian (4), k_2 is given by

$$k_2 = \int_{-\infty}^{\infty} \frac{t^2}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2} \right) dt = 1. \tag{6}$$

The second term on the rhs of expression (1) is

$$k_3 = \left[\int_{-\infty}^{\infty} K(t)^2 dt \right]^{1/5} = \left(\frac{1}{2\sqrt{\pi}} \right)^{1/5}. \tag{7}$$

The integral of the squared second derivative of f ($R(f'')$) will be estimated by

$$\begin{aligned} R(f'') \approx & \int (\hat{f}''(x))^2 dx \\ = & \frac{1}{(nh^3)^2} \int \left(\sum_{i=1}^{n} K'' \left(\frac{x - X_i}{h} \right) \right)^2 dx \\ = & \frac{1}{2\pi(nh^3)^2} \int \left(\sum_{i=1}^{n} (Y_i^2 - 1) \exp\left(-\frac{Y_i^2}{2} \right) \right)^2 dx \end{aligned} \tag{8}$$

where

$$Y_i = \frac{x - X_i}{h}. \tag{9}$$

Since

$$\left(\sum_{i=1}^{n} a_i \right)^2 = 2 \sum_{i=1}^{n} \sum_{j<i} a_i a_j + \sum_{i=1}^{n} a_i^2 \tag{10}$$

we have

$$\begin{aligned} \left(\sum_{i=1}^{n} (Y_i^2 - 1) \exp\left(-\frac{Y_i^2}{2} \right) \right)^2 &= \left(\sum_{i=1}^{n} Z_i \right)^2 \\ &= 2 \sum_{i=1}^{n} \sum_{j<i} Z_i Z_j + \sum_{i=1}^{n} Z_i^2 \end{aligned} \tag{11}$$

where

$$Z_i = (Y_i^2 - 1) \exp\left(-\frac{Y_i^2}{2}\right). \tag{12}$$

The integral of expression (11) can be written as

$$2\sum_{i=1}^{n}\sum_{j<i}\int Z_i Z_j dx + \sum_{i=1}^{n}\int Z_i^2 dx. \tag{13}$$

The first integral in this expression is

$$\int Z_i Z_j dx = \frac{\sqrt{\pi}}{16h^3}\left[\left((X_i - X_j)^2 - 6h^2\right)^2 - 24h^4\right] \exp\left(-\left(\frac{X_i - X_j}{2h}\right)^2\right) \tag{14}$$

and the second is

$$\int Z_i^2 dx = \frac{3h}{4}\sqrt{\pi}. \tag{15}$$

So, expression (13) becomes $\frac{\sqrt{\pi}}{4}k_4$ where

$$k_4 = 3nh + \frac{1}{2h^3}\sum_{i=1}^{n}\sum_{j<i}\left[\left((X_i - X_j)^2 - 6h^2\right)^2 - 24h^4\right] \exp\left(-\left(\frac{X_i - X_j}{2h}\right)^2\right). \tag{16}$$

Finally, our estimate of the best h is given by

$$\hat{h}_{opt} = \left(\frac{4nh^6}{k_4}\right)^{1/5} \tag{17}$$

Note that it still depends on the value of h, as the original expression. What we did was basically adapt it to the discrete case so that we can estimate its value from a sample. To solve this dependency we will use an iterative algorithm: see section 3 bellow.

3 An iterative algorithm for $\hat{h}_{opt}$

We use the following algorithm to determine $\hat{h}_{opt}$.

1. Initialize ϵ and set h_0=S1, $h_1 = h_0 + \epsilon$
2. While $(|h_1 - h_0| > \epsilon)$ do

(a) Set h_0=h_1

(b) Find h_1 using expression (17) with $h = h_0$

(c) Set $h_1 = (h_0 + h_1)/2$

3. Return $\hat{h}_{opt} = h_1$

The difference between consecutive estimates is compared against ϵ and is used to stop the iteration once a given precision is achieved. The $\hat{h}_{opt}$ estimate is initialized to the value of one of the estimators proposed by Silverman in [11], S1 (see below), plus ϵ.

We implemented another version based on the Newton-Raphson method, but the results were similar and this is a simpler algorithm.

4 Experiments

4.1 Experiments on artificial data sets

We built two types of data sets to test the above approximations to the best theoretical h given in expression (1), h_{opt}. The data sets are created with the same distribution but vary in the number of points from 10 to 280 in steps of 30. Also, for a given size, 30 repetitions are made, where naturally the data is different for each repetition. We compare the obtained $\hat{h}_{opt}$ against several common bandwidth estimators. Silverman's empirical expression for the case of a Gaussian kernel and assuming a Gaussian density is [11]

$$S0 = 1.06\sigma n^{-1/5} \tag{18}$$

where n is the number of data points and σ the data standard deviation. A more general proposal from Silverman is [11]

$$S1 = 0.9An^{-1/5} \tag{19}$$

where

$$A = \min(\sigma, \text{interquartile range}/1.34)$$

We will also compare the obtained values against the ones produced by two cross-validation approaches: biased cross-validation (BCV) [7] and unbiased cross-validation (UCV) [3]. Finally we will compare against the Sheather and Jones estimator (SJ) [9] as a member of the plug-in estimators. The label 'Our' represents the estimates produced with our proposal.

The data are generated from normal densities so it is possible to evaluate h_{opt} exactly.

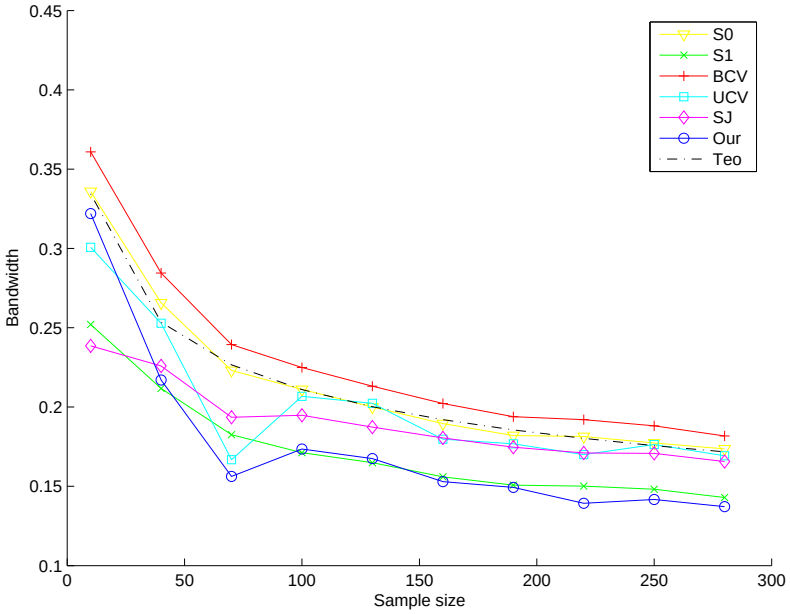

Figure 1: Average bandwidths for 30 repetitions and several sample sizes produced with 6 estimators and the optimal value for the data set 1.

4.1.1 Data set 1

In this case we use data from $f \sim N(0, \sigma)$. Expression (1) becomes

$$h_{opt} = k_3 \left(\frac{3n}{8\sigma^5 \sqrt{\pi}} \right)^{-1/5} = 1.06 \sigma n^{-1/5} \tag{20}$$

In this case, the optimal value is identical to expression (18), but since the latter is an empirical expression the values will be different from the theoretical ones due to the estimation of the standard deviation versus its true value. We used $\sigma = 0.5$ in the experiments.

Table 1 contains the mean squared error of the average estimates regarding the optimal value.

As would be expected, the best values are obtained with the $S0$ estimator. In these experiments our proposal presents the second worst performance. So if the densities to estimate are approximately normal, any of the other estimators (with the exception of S1) will perform better.

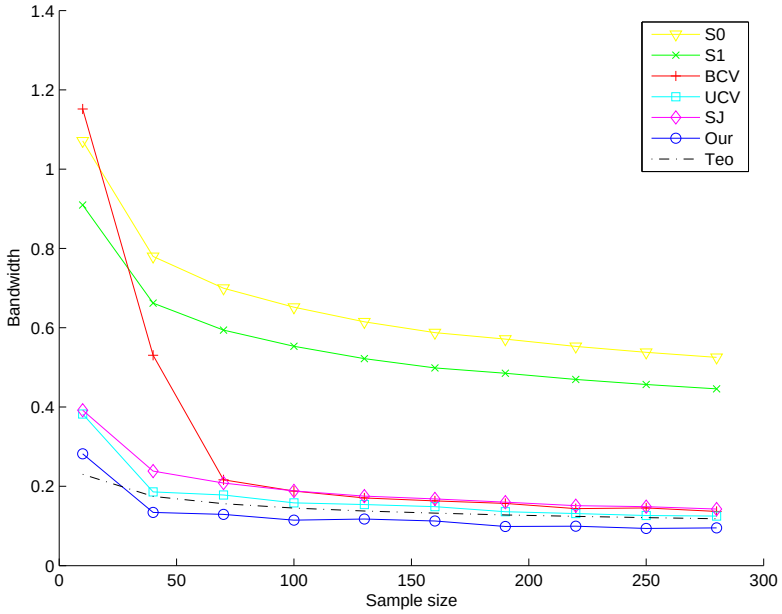

Figure 2: Average bandwidths for 30 repetitions and several sample sizes produced with 6 estimators and the optimal value for the data set 2.

4.1.2 Data set 2

For this data set we use a mixture of two normal densities with equal variances:

$$f(x) = 0.5f_1(x) + 0.5f_2(x),\ \ f_1 \sim N(-2, 0.3),\ \ f_2 \sim N(1, 0.3)$$

Expression (1) becomes in this case

$$h_{opt} = k_3 \left(43.5331n\right)^{-1/5} \tag{21}$$

Figure 2 presents the average for 30 repetitions of the estimated bandwidth obtained with the six estimators referred previously.

Analysing the MSE results in table 1, we observe that our proposal gives the smallest MSE followed closely by the UCV.

4.2 Experiments on real data sets

In this section we show the results of the proposed estimator and the respective densities on two real data sets. The first data set contains the data

Table 1: This table contains the mean squared error of the average estimates to the optimal value, for data sets 1 and 2.

	S0	S1	BCV	UCV	SJ	Our
Data set 1	0.0019	0.1829	0.0276	0.0508	0.1187	0.1577
Data set 2	0.0278	0.0181	0.0099	0.0002	0.0004	0.0001

Table 2: Bandwidth values estimated with different approaches for real data sets.

Data set	S0	S1	BCV	UCV	SJ	Our
Swiss bank notes	0.478	0.401	0.514	0.336	0.311	0.253
Buffalo snowfall	10.979	9.321	11.801	9.371	9.017	6.751

relative to the bottom margin, in millimeters, of 100 forged Swiss bank notes [2]. The second data set is the Buffalo snowfall data. This data set contains 63 observations representing the annual snowfall, in inches, in Buffalo (New York) from winter 1910/11 to 1972/73 [6].

Table 2 contains the kernel bandwidth estimates of the proposed estimator on these data sets, along with estimates from different methods.

Figures 3 and 4 present the estimated densities using our approach and the one by Sheather and Jones (SJ) on these two data sets.

We can see that since the values of the bandwidth obtained with our method are smaller than the ones obtained with any of the other methods tested, and in the case of the figures, smaller than the ones obtained with the SJ method. The peaks of the estimated densities are higher than the ones obtained by the SJ method and the troughs are deeper. This is an improvement since it is considered that the kernel approaches to density estimation usually produce flat peaks and insufficiently deep troughs [4].

5 Conclusions

In this paper we presented an approach to the determination of the optimal bandwidth for one dimensional kernel density estimation. This approach is based on the use of a plug-in estimator together with an iterative algorithm. We presented simulation results on both artificial and real data and compared these results against other common bandwidth estimators.

We concluded that our approach is appealing since it provides sharper peaks and deeper troughs in the estimated densities due to the fact that it yields smaller bandwidth estimates then the methods to which it was compared.

The asymptotic properties of the estimator are currently being studied and

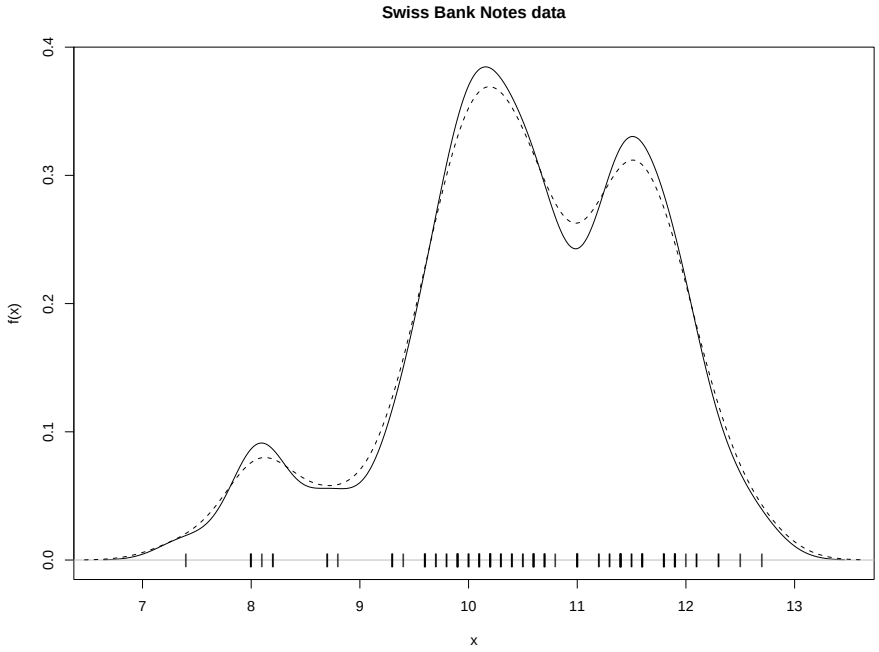

Figure 3: Swiss fraudulent bank notes (the bottom data only): the solid line presents the density estimate with the bandwidth obtained with our method and the dotted line using Sheather and Jones method.

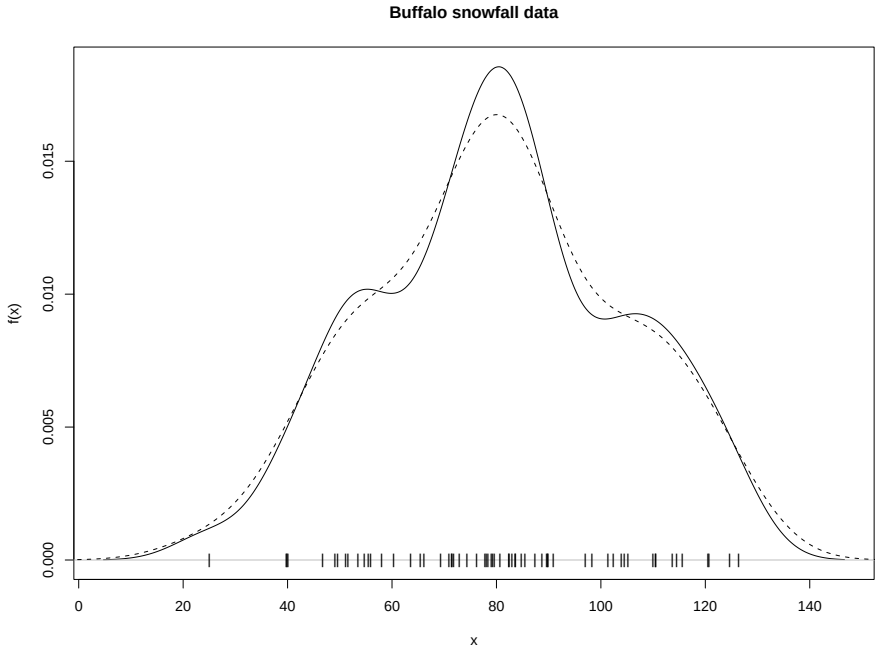

Figure 4: Buffalo snowfall data: the solid line presents the density estimate with the bandwidth obtained with our method and the dotted line using Sheather and Jones method.

will be the subject of a future publication.

Acknowledgments

This work was supported by the Portuguese FCT-Fundação para a Ciência e Tecnologia, POS_Conhecimento and FEDER (project POSC/EIA/56918/2004).

References

[1] L.A. Alexandre and J. Marques de Sá. Error entropy minimization for LSTM training. In *16th International Conference on Artificial Neural Networks - ICANN 2006*, volume LNCS 4131 - part I, pages 244–253, Athens, Greece, September 2006. Springer.

[2] B. Flury and H. Riedwyl. *Multivariate statistics: A practical approach.* Chapman & Hall, London, 1988.

[3] P. Hall. Large sample optimality of least-squares cross-validation in density estimation. *Ann. Statist.*, 11:1156–114, 1983.

[4] M.L. Hazelton and B.A. Turlach. Reweighted kernel density estimation. *Computational Statistics & Data Analysis*, 51:3057–3069, 2007.

[5] J. Santos, L.A. Alexandre, and J. Marques de Sá. The error entropy minimization algorithm for neural network classification. In Ahmad Lofti, editor, *Proceedings of the 5th International Conference on Recent Advances in Soft Computing*, pages 92–97, Nottingham, United Kingdom, December 2004.

[6] D.W. Scott. *Multivariate Density Estimation.* John Wiley & Sons, 1992.

[7] D.W. Scott and G.R. Terrell. Biased and unbiased cross-validation in density estimation. *J. Amer. Statist. Assoc.*, 82:1131–1146, 1987.

[8] S.J. Sheather. Density estimation. *Statistical Science*, 19(4):588–597, 2004.

[9] S.J. Sheather and M.C. Jones. A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society. Series B*, 53(3):683–690, 1991.

[10] L.M. Silva, C.A. Felgueiras, L.A. Alexandre, and J. Marques de Sá. Error entropy in classification problems: A univariate data analysis. *Neural Computation*, 18(9):2036–2061, September 2006.

[11] B.W. Silverman. *Density estimation for statistics and data analysis*. Chapman & Hall, 1986.