Hindawi Publishing Corporation
EURASIP Journal on Advances in Signal Processing
Volume 2008, Article ID 739082, 8 pages
doi:10.1155/2008/739082

Research Article

Analytical Plug-In Method for Kernel Density Estimator Applied to Genetic Neutrality Study

Molka Troudi,[1, 2] Adel M. Alimi,[1] and Samir Saoudi[2]

[1] *REGIM, Ecole Nationale des Ingénieurs de Sfax, Route de Soukra, Km 1.5, 3002, Sfax, Tunisia*
[2] *Département Signal et Communications, TELECOM Bretagne, Institut TELECOM, Technopôle Brest-Iroise, CS 83818, 29283 Brest Cedex 3, France*

Correspondence should be addressed to Samir Saoudi, samir.saoudi@enst-bretagne.fr

Received 21 November 2007; Accepted 1 May 2008

Recommended by Jar-Ferr Yang

The plug-in method enables optimization of the bandwidth of the kernel density estimator in order to estimate probability density functions (pdfs). Here, a faster procedure than that of the common plug-in method is proposed. The mean integrated square error (MISE) depends directly upon $J(f)$ which is linked to the second-order derivative of the pdf. As we intend to introduce an analytical approximation of $J(f)$, the pdf is estimated only once, at the end of iterations. These two kinds of algorithm are tested on different random variables having distributions known for their difficult estimation. Finally, they are applied to genetic data in order to provide a better characterisation in the mean of neutrality of Tunisian Berber populations.

1. INTRODUCTION

The problem of estimation of a probability density function $f(x)$ is interesting for many reasons, among which are the possible applications in the field of discriminant analysis or the estimation of functions of the density. The parametric approach to density estimation assumes a functional form for the density and then estimates the unknown parameters using techniques such as the maximum likelihood estimation or Pearson system based on the estimation of the skewness and the Kurtosis [1]. However, unless the form of density is known a priori, assuming a functional form for a density very often leads to erroneous inference. On the other hand, nonparametric methods do not make any assumptions as to the form of the underlying density. Today, a rich basket of nonparametric density estimators (Kernel, orthogonal series, histogram, etc.) exists [2–4].

This work focuses on kernel density estimators (KDE) as introduced by Rozenblatt [5] and Parzen [6]. These estimators are defined by

$$\widehat{f}_n(x) = \frac{1}{nh_n}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right), \tag{1}$$

where $(X_i)_{1\leq i\leq n}$ is the observed data with length equal to n; h_n is called the bandwidth; and K is a probability density function called the Kernel. K is assumed to be an even regular function with unit variance and zero mean. The Kernel K is called regular if it is a square integrated density.

For a practical implementation of KDE, the choice of the bandwidth h_n is very important. Small h_n leads to an estimator with a small bias and large variance, whereas large h_n leads to a small variance at the expense of increase: the bandwidth has to be optimally chosen.

Several techniques have been proposed for optimal bandwidth selection. The best known of these include rules of thumb, oversmoothing, least squares cross-validation, direct plug-in methods, solve-the-equation plug-in method, and the smoothed bootstrap [7].

Here, a fast version of the plug-in method which gives a good approximation of the optimal bandwidth in the mean integrated square error (MISE) sense is considered. The plug-in method achieves approximation of the bandwidth h_n by an iterative approximation of second derivative of the density f, noted by $J(f)$. Thus, a sequence of positive numbers $h_n^{(k)}$ is constructed through the iterations with n as the sample size, and k as the number of iterations.

The analytical approximation of $J(f)$ enables us to estimate the pdf only once, whereas the numerical approximation of $J(f)$ requires the estimation of the pdf for each iteration.

The present paper is organized as follows. We recall, in Section 1, the principle of the convergence theorem of such an estimator in the mean integrated square error (MISE) sense. A description of the plug-in algorithm is proposed in Section 2. In Section 3, the fast plug-in algorithm is introduced. Therefore, an experimental comparison between the numerical and analytical plug-in KDE is presented. The last section is devoted to the study of the distribution of the statistic D obtained from simulated neutral populations which is described in the same section.

2. THEORETICAL STUDIES

2.1. Notations and recalls

To evaluate the performance of the KDE, it is necessary to choose a measure of distance between the true density f and its estimate $\widehat{f}_n$. Especially, common choices are the integrated square error (ISE) and its expected value, the mean integrated square error (MISE):

$$\text{ISE}(f,\widehat{f}_n) = D_2(\widehat{f}_n, f) = \int_{-\infty}^{+\infty} |\widehat{f}_n(x) - f(x)|^2 dx,$$

$$\text{MISE} = E[\text{ISE}(f,\widehat{f}_n)] = E\left[\int_{-\infty}^{+\infty} |\widehat{f}_n(x) - f(x)|^2 dx\right]. \quad (2)$$

The convergence of $\widehat{f}_n$ depends on the choice of both the kernel function and the bandwidth h_n. However, the choice of h_n is much more important for the behavior of $\widehat{f}_n$ than the choice of K.

The optimal kernel K_o and the optimal bandwidth are those which minimize the mean integrated square error (MISE). However, the condition of convergence required by MISE is as follows: $n(h_n)^2$ has to tend toward 0 when n tends toward the infinity.

2.2. Convergence in the MISE sense

The minimization of MISE with respect to the bandwidth, for a fixed size n of the sample, implies the following asymptotic study.

Let us consider the expression of mean square error (MSE):

$$\text{MSE} = E\left[|\widehat{f}_n - f|^2\right] = \text{var}(\widehat{f}_n) + (f - E[\widehat{f}_n])^2. \quad (3)$$

The development of this expression gives the following formula (see the appendix):

$$E\left[|\widehat{f}_n - f|^2\right] = \frac{1}{nh_n}\int K^2(u) f(x-h_n u)\,du + \left[\int K(u)(f(x-uh_n) - f)\,du\right]^2 - \frac{1}{n}\left(\int K(u) f(x-h_n u)\,du\right)^2. \quad (4)$$

Firstly, let us consider the Taylor pdf expansion:

$$f(x-h_n u) = f(x) - h_n u f'(x) + \frac{u^2}{2}h_n^2 f''(x) - \frac{u^3 h_n^3}{6} f^{(3)}(x - \theta h_n u), \quad (5)$$

where $0 < \theta < 1$.

By using the following notations:

$$M(K) = \int_{-\infty}^{+\infty} K^2(u)\,du, \quad (6)$$

$$J(f) = \int_{-\infty}^{+\infty} (f''(x))^2 dx, \quad (7)$$

where f'' is the second derivative of f.

$\Delta(h_n)$, which is the Taylor expansion of the MISE (and consequently an approximation of MISE), is given by (more details are given in the appendix)

$$\text{MISE} \approx \Delta(h_n) = \frac{M(K)}{nh_n} + \frac{J(f)h_n^4}{4}. \quad (8)$$

The minimum value of the function $\Delta(h_n)$ is obtained by annulling its derivative $\Delta'(h_n) = 0$:

$$\Delta'(h_n) = -\frac{M(K)}{nh_n^2} + h_n^3 J(f) = 0. \quad (9)$$

Therefore, the optimal value of h_n noted by h_n^* becomes

$$h_n^* = n^{-1/5} \cdot (J(f))^{-1/5} \cdot (M(K))^{1/5}. \quad (10)$$

This gives the minimum for the MISE formulated in this expression:

$$\text{MISE} = \frac{5}{4} n^{-4/5} (M(K))^{4/5} (J(f))^{1/5}. \quad (11)$$

On the other hand, the optimal kernel K_o, in MISE sense, has the following expression:

$$K_o(x) = \begin{cases} 0 & \text{if } |x| > \sqrt{5}, \\ \dfrac{3}{4\sqrt{5}}\left(1 - \dfrac{x^2}{5}\right) & \text{if } |x| \le \sqrt{5}. \end{cases} \quad (12)$$

2.3. Plug-in algorithm

Even if the asymptotic study gives the expression of the optimal bandwidth, it seems to be difficult to use it in practice as it depends on the unknown density f. Thus, several methods have been developed to estimate the optimal bandwidth from a given data set $X_1, \dots, X_n$. In this paper, we are interested in the plug-in method. Such a method is an iterative algorithm which converges to the optimal bandwidth.

In this section, we will recall the different steps of the plug-in algorithm. In Section 3, the new idea which reduces it to another algorithm having a less complex computation is described.

Step 1. Determination of $M(K)$ which is computed using an analytical integration (6).

Step 2. Arbitrary initialization of $J^{(0)}(f)$ in order to determine $h_n^{(0)}$, the first value of h_n (10).

Step 3. Estimation of the pdf $f^{(0)}$ using $h_n^{(0)}$ and (1).

Step 4. At the Kth iteration, deduction of $J^{(k)}(f)$ value from the pdf $f^{(k-1)}$ by using (7). Therefore, $h_n^{(k)}$ is computed by using (10), and $f^{(k)}$ is re-estimated by using (1). The iterations will allow each time to re-estimate numerically $J(f)$, and therefore h_n and f.

Step 5. Stop criterion: $|h_n^{(k-1)} - h_n^{(k)}| < \varepsilon$.

3. FAST ITERATIVE PLUG-IN ALGORITHM

3.1. The analytical approximation of $J(f)$ in the case of optimal kernel

In this section, we intend to show how it is possible to compute analytically $J(f)$ in the case of the optimal kernel:

$$\hat{f}''(x) = \frac{1}{nh_n^3}\sum_{i=1}^{n} K''\left(\frac{x - X_i}{h_n}\right). \tag{13}$$

We have

$$K_o(x) = \begin{cases} 0 & \text{if } |x| > \sqrt{5}, \\ \dfrac{3}{4\sqrt{5}}\left(1 - \dfrac{x^2}{5}\right) & \text{if } |x| \leq \sqrt{5}. \end{cases} \tag{14}$$

Then,

$$K''(x) = \begin{cases} 0, & \text{if } |x| > \sqrt{5}, \\ \text{indefinite}, & \text{if } |x| = \sqrt{5}, \\ -\dfrac{3\sqrt{5}}{50}, & \text{if } |x| < \sqrt{5}, \end{cases} \tag{15}$$

Considering the following function which is constant in intervals and which forms a partition of the real line:

$$\beta(x) = \left[\sum_{i}^{n} K''\left(\frac{x - X_i}{h_n}\right)\right]^2 = \left[\sum_{i \in A_n(x)} K''\left(\frac{x - X_i}{h_n}\right)\right]^2, \tag{16}$$

where $A_n(x)$ is the following subset of natural integers:

$$A_n(x) = \left\{0 \leq i \leq n;\ \frac{|x - X_i|}{h_n} \leq \sqrt{5}\right\}. \tag{17}$$

$J(f)$ is composed of a finite sum of second derivatives of the optimal kernel. Therefore, the number of indefinite points for the function $\beta(x)$ is also finite.

The contribution of such a set of points to the integral value $(J(f))$ is, therefore, relatively marginal. This implies that

$$J(f) \approx \int_{-\infty}^{+\infty}\left[\frac{1}{nh_n^2}\sum_{i=1}^{n} K''\left(\frac{x - X_i}{h_n}\right)\right]^2 dx, \tag{18}$$

$$J(f) \approx \frac{1}{n^2 h_n^6}\int_{-\infty}^{+\infty}\beta(x)dx. \tag{19}$$

Nevertheless, as observed in the following simulations, the best power of the bandwidth h_n which optimizes the MISE belongs to the interval [4, 5] of the real line. This result has been deduced experimentally: several simulated distributions have been tested versus h_n powers.

Table 1 visualizes the evolution of MISE for some values of h_n power. The three distributions presented are selected from an important number of studied ones. Distribution 1 is a mixture from two uniform distributions; Distribution 2 is exponential, whereas Distribution 3 is a mixture between a uniform distribution and a Gaussian distribution. It is easy to observe that the power 4.5 seems to be that which allows the convergence.

From a theoretical point of view, this could be explained by the fact that the derivation proposed here could be seen as the well-known kernel approximation method. The variance parameter of the kernel approximation method corresponds to the bandwidth of the kernel estimator which needs to be adjusted. Consequently, the optimal expression of $J(f)$ becomes

$$J(f) \approx \frac{9}{500}\frac{1}{n^2 h_n^{4.5}}\int_{-\infty}^{+\infty}\beta(x)dx. \tag{20}$$

3.2. The fast plug-in optimal KDE

In this section, we intend to describe the different steps of the proposed fast iterative optimal kernel estimator algorithm.

Step 1. Determination of $M(K)$ which is computed using an analytical integration (6).

Step 2. Arbitrary initialization of $J^{(0)}(f)$ in order to determine $h_n^{(0)}$, the first value of h_n (10).

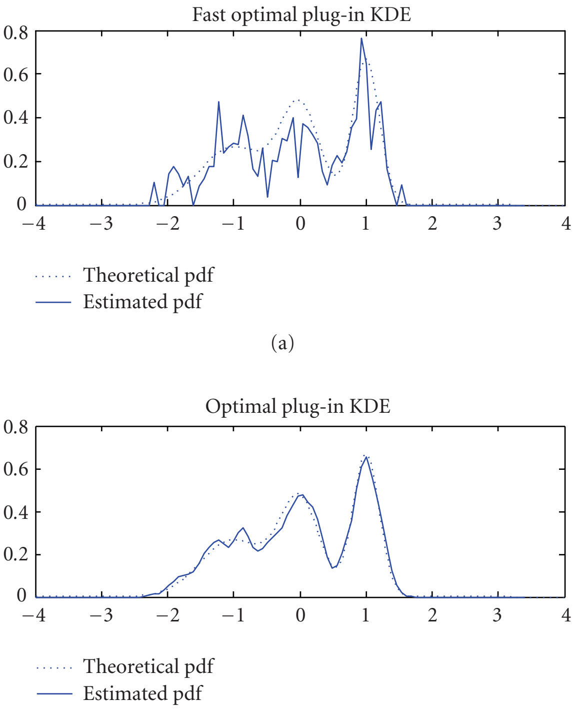
Step 3. At the kth iteration, computation of $J^{(k)}(f)$ value directly from the sample (X_i) using (20). Therefore, $h_n^{(k)}$ is computed from (10), and $J(f^{(k)})$ is recomputed from (20). The iterations will allow each time to re-estimate analytically $J(f)$.

Step 4. Stop criterion: $|h_n^{(k-1)} - h_n^{(k)}| < \varepsilon$.

The improvement in the complexity computation sense is obtained by the introduction of the analytical approximation of $J(f)$. For this reason, the proposed fast algorithm does not need the estimation of the pdf step.

Table 1: MISE versus power of h_n for three selected distributions.

Power of h_n	6.0	5.5	5.0	4.5	4.0	3.5	3.0
Distribution $1*10^{-4}$	58.0	11.0	5.5	4.3	4.4	5.9	5.9
Distribution $2*10^{-4}$	66.0	35.0	34.0	17.0	27.0	34.0	53.0
Distribution $3*10^{-4}$	64.0	5.8	3.1	1.1	4.8	6.0	9.8



(a)

(b)

Figure 1: Multimodal pdf estimated using optimal plug-in KDE and fast optimal plug-in KDE (power 6).

Table 2: MISE and variance of 100 simulations of the two studied KDE.

	MISE $*10^{-4}$	Variance $*10^{-8}$
Plug-in optimal KDE	2.8	9.9
Fast plug-in optimal KDE	2.7	1.1

3.3. Estimation of simulated Gaussian mixture pdf

In this section, the following Gaussian distribution mixtures are considered.

3.3.1. Multimodal distribution

$$f(x) = \pi_1 f_{\mu_1,\sigma_1}(x) + \pi_2 f_{\mu 2,\sigma_2}(x) + \pi_3 f_{\mu_3,\sigma_3}(x) \qquad (21)$$

with $\mu_1 = -1$, $\mu_2 = 0$, $\mu_3 = 1$, and $\sigma_1 = 0.5$, $\sigma_2 = 0.3$, $\sigma_3 = 0.2$, μ_i, and σ_i^2 are, respectively, the mean and the variance of each distribution.

The a priori probabilities π_1, π_2, and π_3 are, respectively, equal to $0.3, 0.35$, and 0.35. The data size is 1000. Then, plug-in optimal KDE and fast plug-in optimal KDE are applied. The two estimations are compared with the theoretical distribution in the mean of MISE criterion which is computed for each case.

Figure 1 represents each of the KDE pdf estimations using both the fast plug-in algorithm (power 6 and power 4.5) and the classical plug-in algorithm for estimating the optimal bandwidth. It is clear that the convergence is not reached with the bandwidth selected by using the fast plug-in algorithm.

However, the power of h_n has been adjusted experimentally. Figure 2 clearly shows that the choice of 4.5 as a power value of h_n in (20) instead of 6 enables the fast plug-in optimal KDE to give a pdf estimation as good as the classical plug-in optimal KDE. This result is corroborated by MISE values and their variance presented in Table 2.

Figure 3 plots the evolution of the MISE versus the sample size (100 iterations). The two curves are very close: the MISE, while having nearly the same values, are weaker, and tends toward 0 when the sample size is growing.

Table 3 summarizes the MISE values and their variances: the difference between the estimations issued from the two plug-in KDE tends toward 0.

3.3.2. Bimodal distribution

The study of another example which includes a uniform distribution known for its estimation difficulties is proposed below:

$$f(x) = \pi_1 f_{\mu,\sigma}(x) + \pi_2 f_{a,b}(x) \qquad (22)$$

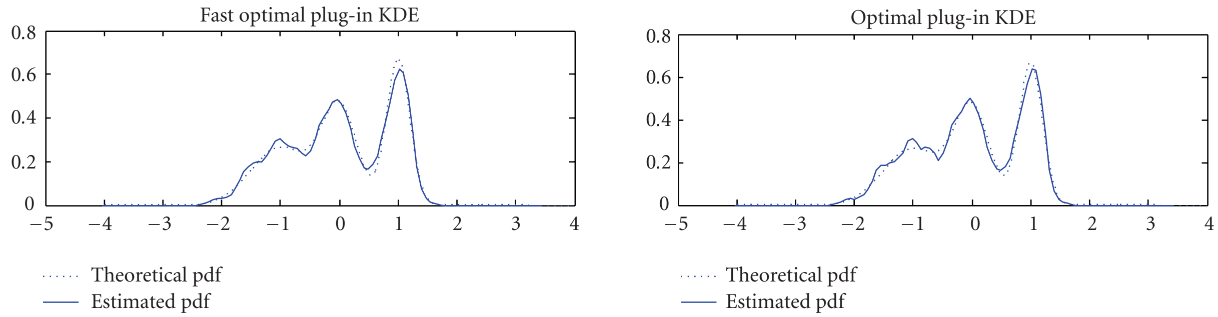
with $\mu = 0.3$ and $\sigma = 0.2$; μ and σ are, respectively, the mean and the variance of the Gaussian distribution. The parameters of the uniform distribution are $a = -0.3$ and $b = 0.2$.

The a priori probabilities π_1 and π_2 are, respectively, 0.75 for the Gaussian distribution and 0.25 for the uniform distribution. The sample size is 4000.

The results are the same as those obtained for the multimodal distribution. As observed in Figure 4, the convergence is obtained with (20), that is, a power value of 4.5 for h_n. The two KDE give close estimations of the pdf. The MISE and their variances confirm these observations (Table 4).

3.4. Complexity and convergence

For the plug-in optimal KDE algorithm, f and $J(f)$ are estimated k times, where k is the number of iterations. k has a small value due to the speed convergence of the KDE algorithm. We note by n the sample size and by p the



(a)

(b)

Figure 2: Estimation of simulated pdf using optimal plug-in KDE and fast optimal plug-in KDE (power 4.5).

Table 3: Convergence of MISE and variance values versus sample size computed from 1000 simulations of the theoretical distribution using the two studied KDE.

Sample size		500	1000	1500	2000	2500	3000	3500	4000
Fast plug-in KDE	$10^{-4}*$ MISE	8.3	6.3	5.5	5.3	4.7	4.4	4.4	4.3
	$10^{-8}*$ variance	27.5	13.6	9.6	8.3	6.2	5.2	3.4	3.0
Plug-in KDE	$10^{-4}*$ MISE	8.5	6.4	5.6	5.3	4.7	4.5	4.4	4.3
	$10^{-8}*$ variance	26.7	13.0	7.8	7.4	4.3	4.3	3.2	3.0

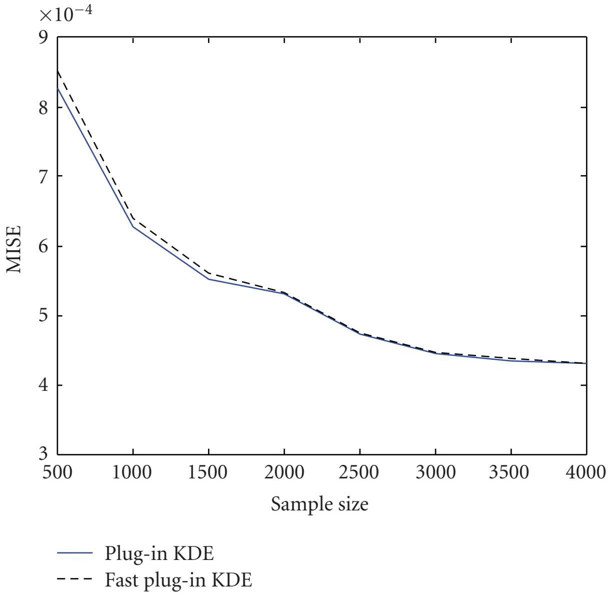
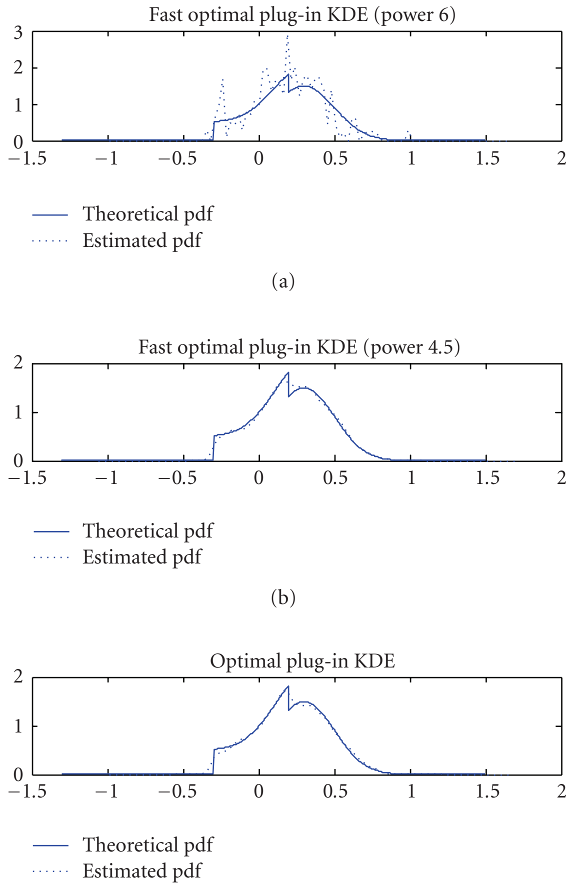


Figure 3: MISE estimated by plug-in optimal KDE and fast plug-in optimal KDE (power 4.5) versus the sample size.

Table 4: MISE and variance of 100 simulations of the theoretical distribution using the two studied KDE.

	MISE $*10^{-3}$	Variance $*10^{-6}$
Plug-in optimal KDE	21.8	26.1
Fast plug-in optimal KDE (power 4.5)	21.8	26.4



(a)

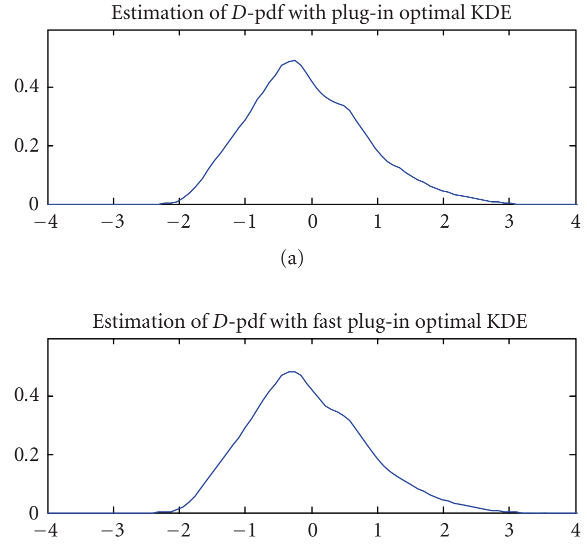
(b)

(c)

Figure 4: Bimodal pdf estimated by optimal plug-in KDE and fast optimal plug-in KDE (power 6 and 4.5).

Table 5: Characteristics of the studied populations.

Population	n	π	D_p
Sened	55	7.60471	−1.717
Testour	40	6.33846	−1.755



(a)

(b)

Figure 5: Estimation of the D-pdf of Sened by optimal plug-in KDE and fast plug-in KDE.

resolution. The estimation of f is $O(2np)$, and the estimation of $J(f)$ is $O(2p)$. The complexity of this iterative algorithm is consequently $O(2knp)$. In the fast plug-in optimal KDE algorithm, f is estimated only once. Thus, the computational cost of this algorithm is $O(2p(k+n))$. The value of k is small in comparison to the value of n : k is around 5 and n generally exceeds 500. It can, therefore, be neglected and the computational cost becomes $O(2np)$.

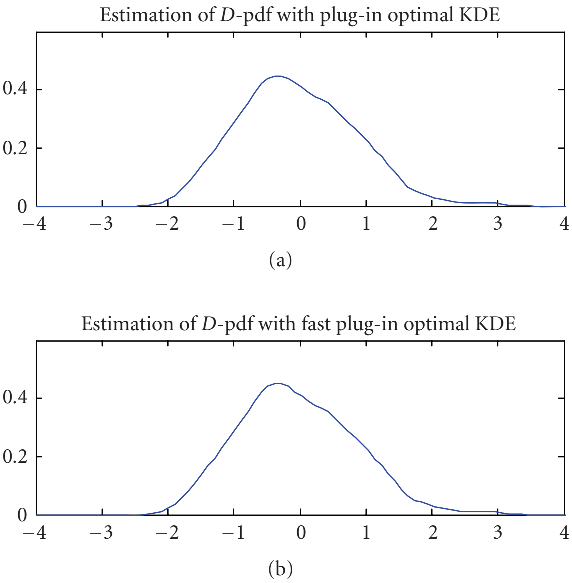
4. ESTIMATION OF PDF OF TUNISIAN GENETIC PARAMETER

In this section, we are interested in the genetics of populations and more specifically in Tajima's estimation of the pdf of the statistic D of Tajima. This is estimated in order to evaluate the neutrality of studied populations. The data are obtained by generating neutral populations, using parameters θ and n [8]. Such parameters are computed from the sample data. The parameter θ is defined as equal to $4N\mu$, N being the effective population size, μ the mutations number per generation, and n is the sample size.

Several methods have been proposed to estimate θ. We can cite here the number of segregation sites S [9] and the average number of (pairwise) nucleotides differences between the DNA sequences, called π [10].

π gives a direct estimation of θ, that is, $E(\widehat{\pi}) = \theta$.

On the other hand, we have $\widehat{\theta} = S/a_1$, where $a_1 = \sum_{i=1}^{n-1}(1/i)$.



(a)

(b)

Figure 6: Estimation of the D-pdf of Testour by optimal plug-in KDE and fast plug-in KDE.

The difference between S and π is the effect of selection. Deleterious mutants are maintained in a population with low frequency. If some of the mutants observed have a low frequency, $\widehat{\pi}$ may not be the same as S/a_1.

Tajima has proposed the following statistic:

$$D = \frac{d}{\sqrt{\widehat{V}(d)}} = \frac{\widehat{\pi} - S/a_1}{\sqrt{\widehat{V}(d)}}. \tag{23}$$

The mean and the variance of the statistic D are approximately 0 and 1, respectively. If the distribution of D is known, it is possible to use it in testing neutral mutation hypothesis. For this purpose, a computer simulation was conducted [8] using S and $\widehat{\pi}$ parameters computed from the actual data. The value of D obtained from the studied population sample using (23) is called D_p. $P[D < D_p]$ must be higher than 0.02 for declaring the population as neutral with a first-kind risk α equals to 5%.

The characteristics of the considered populations are introduced in Table 5.

Pdf of theoretical neutral distributions of D are estimated by using the two plug-in optimal KDE studied in this paper: the plug-in optimal KDE and the fast plug-in optimal KDE. It is noticed that there are no differences between the two estimations (Figures 5 and 6).

For a better estimation of the neutrality of the studied populations, we propose to compute a mean value of $P[D < D_p]$ deducted from 100 simulations of theoretical neutral populations. The obtained results are presented in Table 6.

The two KDE give the same conclusions about the neutrality of studied populations. Yet, it is possible to say without any ambiguity that Sened population is a neutral population, whereas Testour population is not neutral. The improvement given by the implementation of the

Table 6: Mean values of $P[D < D_p]$ obtained from the two optimal plug-in KDE studied.

	$P[D < D_p]$ (Mean value)	
Population	Plug-in optimal KDE	Fast plug-in optimal KDE
Sened	$21.3*10^{-3}$	$21.4*10^{-3}$
Testour	$18.3*10^{-3}$	$18.4*10^{-3}$

fast optimal KDE relatively to the optimal KDE concerns particularly the computational cost.

5. CONCLUSIONS

In this paper, we have presented a fast version of the plug-in algorithm which estimates the optimal KDE bandwidth as well as the classical plug-in algorithm. Such a method is based on the optimal kernel which is directly derived without taking the indefinite points into account. The convergence in the MISE sense is obtained with less complexity.

The entity noticed by $J(f)$ which represents the integral of the second-order derivative of the pdf is approximated analytically at each iteration in order to tend to the optimal bandwidth. However, the mathematical expression computed analytically (19) gives an incorrect estimation of the tested pdf. The value of the power of h_n was, therefore, experimentally adjusted to 4.5. The efficiency of this fast algorithm was tested on several simulations of multimodal densities which are known for being difficult cases in the mean of estimation problem, and the proposed algorithm allows estimation comparable to the plug-in optimal KDE in the MISE sense with an improvement of the computational cost.

In the field of genetic population, this fast algorithm allows a better estimation of the neutrality of studied populations in the computational cost sense, especially if this neutrality is not robust ($P[D < D_p] \approx 0.02$). The computation of a mean value of this probability needs a large computational time. The use of the fast plug-in optimal KDE provides an improvement of the computational cost without any loss of quality of results studied.

In fact, several applications could be concerned by the proposed fast algorithm. The estimation of the probability error for the CDMA communication systems should be mentioned as it will be considered in future work. The generalization to the multivariate case will also be dealt with, in addition to the consideration of densities confined to a bounded support [3, 11, 12] in order to study the same idea for the diffeomorphism KDE.

APPENDIX

$$\text{MSE} = E\left[|\hat{f}_n - f|^2\right] = E[\hat{f}_n^2] + f^2 - 2fE[\hat{f}_n],$$
$$E\left[|\hat{f}_n - f|^2\right] = E[\hat{f}_n^2] - (E[\hat{f}_n])^2 + (E[\hat{f}_n])^2 + f^2 - 2fE[\hat{f}_n],$$
$$E\left[|\hat{f}_n - f|^2\right] = \text{var}(\hat{f}_n) + (f - E[\hat{f}_n])^2. \tag{A.1}$$

Let us compute the variance of $\hat{f}_n$ for a kernel density estimator assuming that the kernel K is an even regular function with unit variance and zero mean:

$$\text{var}(\hat{f}_n) = \text{var}\left(\frac{1}{nh_n}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right)\right),$$
$$\text{var}(\hat{f}_n) = \frac{1}{n^2 h_n^2}\,\text{var}\left(\sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right)\right),$$
$$\text{var}(\hat{f}_n) = \frac{n}{n^2 h_n^2}\,\text{var}\left(K\left(\frac{x - X_1}{h_n}\right)\right) = \frac{1}{nh_n^2}\,\text{var}\left(K\left(\frac{x - X_1}{h_n}\right)\right),$$
$$\text{var}(\hat{f}_n) = \frac{1}{nh_n^2}\left(E\left[K^2\left(\frac{x - X_1}{h_n}\right)\right] - \left(E\left[K\left(\frac{x - X_1}{h_n}\right)\right]\right)^2\right),$$
$$\text{var}(\hat{f}_n) = \frac{1}{nh_n^2}\left[\int K^2\left(\frac{x - y}{h_n}\right) f(y)dy - \left(\int K\left(\frac{x - y}{h_n}\right) f(y)dy\right)^2\right]. \tag{A.2}$$

Let us consider the following change of variable: $u = ((x - y)/h_n)$,

$$\text{var}(\hat{f}_n) = \frac{1}{nh_n^2}\left[\int K^2(u) f(x - h_n u) h_n du - \left(\int K(u) f(x - h_n u) h_n du\right)^2\right], \tag{A.3}$$
$$\text{var}(\hat{f}_n) = \frac{1}{nh_n}\int K^2(u) f(x - h_n u) du - \frac{1}{n}\left(\int K(u) f(x - h_n u) du\right)^2.$$

Secondly, we are interested in the development of $(f - E[\hat{f}_n])^2$,

$$E[\hat{f}_n] = E\left[\frac{1}{nh_n}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right)\right],$$
$$E[\hat{f}_n] = \frac{1}{nh_n} E\left[\sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right)\right], \tag{A.4}$$
$$E[\hat{f}_n] = \frac{n}{nh_n} E\left[K\left(\frac{x - X_1}{h_n}\right)\right],$$
$$E[\hat{f}_n] = \frac{1}{h_n}\int K\left(\frac{x - y}{h_n}\right) f(y)dy.$$

Let us consider the following change of variable $u = ((x-y)/h_n)$,

$$E[\hat{f}_n] = \frac{1}{h_n}\int K(u) f(x-uh_n) h_n du = \int K(u) f(x-uh_n) du,$$

$$(f - E[\hat{f}_n])^2 = (E[\hat{f}_n] - f)^2 = \left[\int K(u) f(x-uh_n) du - \int K(u) f(x) du\right]^2,$$

$$(E[\hat{f}_n] - f)^2 = \left[\int K(u)(f(x-uh_n) - f(x)) du\right]^2. \tag{A.5}$$

Using (A.3) and (A.5), we have

$$\begin{aligned} E\left[\,|\hat{f}_n - f|^2\right] &= \frac{1}{nh_n}\int K^2(u) f(x-h_n u) du \\ &\quad + \left[\int K(u)(f(x-uh_n) - f(x)) du\right]^2 \\ &\quad - \frac{1}{n}\left(\int K(u) f(x-h_n u) du\right)^2 \\ &= A_n(x) + B_n(x) - C_n(x) \end{aligned} \tag{A.6}$$

with

$$\begin{aligned} A_n(x) &= \frac{1}{nh_n}\int K^2(u) f(x-h_n u) du, \\ B_n(x) &= \left[\int K(u)(f(x-uh_n) - f(x)) du\right]^2, \\ C_n(x) &= \frac{1}{n}\left(\int K(u) f(x-h_n u) du\right)^2. \end{aligned} \tag{A.7}$$

By introducing the following Taylor expansion,

$$\begin{aligned} &f(x-h_n u) \\ &\quad = f(x) - h_n u f'(x) + \frac{u^2}{2} h_n^2 f''(x) - \frac{u^3 h_n^3}{6} f^{(3)}(x - \theta h_n u), \end{aligned} \tag{A.8}$$

where $0 < \theta < 1$, $A_n(x)$ and $B_n(x)$ can be approximated by

$$\begin{aligned} A_n(x) &\approx \frac{f(x)}{nh_n}\int_{-\infty}^{+\infty} K^2(u) du, \\ B_n(x) &\approx \left[-f'(x) h_n^4 \int_{-\infty}^{+\infty} uK(u) du \right. \\ &\quad \left. + \frac{f''(x)}{2} h_n^2 \int_{-\infty}^{+\infty} u^2 K(u) du\right]^2, \\ C_n(x) &\approx \frac{f^2(x)}{n}. \end{aligned} \tag{A.9}$$

By considering only even kernels with unit variance and zero mean,

$$\begin{aligned} B_n(x) &\approx \frac{f''^2}{4} h_n^4, \\ \text{MISE} &= \int_{-\infty}^{+\infty} (A_n(x) + B_n(x) - C_n(x)) dx. \end{aligned} \tag{A.10}$$

By using the following notations:

$$M(K) = \int_{-\infty}^{+\infty} K^2(u) du, \qquad J(f) = \int_{-\infty}^{+\infty} (f''(x))^2 dx, \tag{A.11}$$

where f'' is the second derivative of f; MISE can be approximated by $\Delta(h_n)$; the Taylor expansion. Then,

$$\text{MISE} \approx \Delta(h_n) = \frac{M(K)}{nh_n} + \frac{J(f) h_n^4}{4}. \tag{A.12}$$

REFERENCES

[1] H. Oja, "On location, scale, skewness and kurtosis of univariate distribution," *Scandinavian Journal of Statistics*, vol. 8, no. 2, pp. 164–168, 1981.

[2] A. W. Bowman and A. Azzalini, *Applied Smoothing Techniques for Data Analysis*, Oxford University Press, Oxford, UK, 1997.

[3] P. Hall, "Comparison of two orthogonal series methods of estimating a density and its derivatives on interval," *Journal of Multivariate Analysis*, vol. 12, no. 3, pp. 432–449, 1982.

[4] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*, Chapman & Hall, London, UK, 1986.

[5] M. Rozenblatt, "Remarks on some non-parametric estimates of a density function," *Annals of Mathematical Statistics*, vol. 27, no. 3, pp. 832–837, 1956.

[6] E. Parzen, "On estimation of a probability density function and mode," *Annals of mathematical statistics*, vol. 33, no. 3, pp. 1065–1076, 1962.

[7] M. C. Jones, J. S. Marron, and S. J. Sheather, "A brief survey of bandwidth selection for density estimation," *Journal of the American Statistical Association*, vol. 91, no. 433, pp. 401–407, 1996.

[8] F. Tajima, "Statistical method for testing the neutral mutation hypothesis by DNA polymorphism," *Genetics*, vol. 123, no. 3, pp. 585–595, 1989.

[9] G. A. Watterson, "On the number of segregating sites in genetical models without recombination," *Theoretical Population Biology*, vol. 7, no. 2, pp. 256–276, 1975.

[10] F. Tajima, "Evolutionary relationship of DNA sequences in finite populations," *Genetics*, vol. 105, no. 2, pp. 437–460, 1983.

[11] S. Saoudi, A. Hillion, and F. Ghorbel, "Non-parametric probability density function estimation on a bounded support: applications to shape classification and speech coding," *Applied Stochastic Models and Data Analysis*, vol. 10, no. 3, pp. 215–231, 1994.

[12] S. Saoudi, F. Ghorbel, and A. Hillion, "Some statistical properties of the kernel-diffeomorphism estimator," *Applied Stochastic Models and Data Analysis*, vol. 13, no. 1, pp. 39–58, 1997.