ISSN 2464-4617(print)
ISSN 2464-4625(CD)
Computer Science Research Notes
CSRN 2803
Poster's Proceedings
http://www.WSCG.eu

The Fast Semi-bounded Kernel-Diffeomorphism Estimator

Ibtissem Ben Othman
GRIFT Research Group
CRISTAL Laboratory
School of Computer Sciences
2010, Manouba, Tunisia
ben.othman.ibtissem@gmail.com

Molka Troudi
ECSTRA Laboratory
Department of Quantitative Methods
IHEC Carthage
Carthage 2016, Tunisia
molkaghorbel@gmail.com

Faouzi Ghorbel
GRIFT Research Group
CRISTAL Laboratory
School of Computer Sciences
2010, Manouba, Tunisia
faouzi.ghorbel@ensi.rnu.tn

ABSTRACT

We introduce, by this work, a fast method to estimate probability density functions in the semi-bounded case. This new technique is a simplified version of the kernel-diffeomorphism estimator which requires complexity in the calculations. It is based on a logarithmic transformation of the data which will be estimated by the conventional kernel estimator. Thus, the algorithm complexity is reduced from $O(N^2)$ to $O(N)$.

Keywords

Kernel density estimator, Kernel-diffeomorphism Plug-in algorithm, Fast semi-bounded kernel-diffeomorphism Estimator.

1 INTRODUCTION

The estimation of probability density functions (pdf) is often required to study the complex technological systems and scientific phenomena. The various examples of statistics correspond to distributions with bounded or semi-bounded supports. To estimate the pdf, there are two classes of methods: parametric or non-parametric. In most situations, probability densities are unknown, such operation can be done by the non-parametric methods which are more precise. The histogram method [Silverman86], the orthogonal functions [Hall82] and the kernel method [Fukunaga13] are among the most frequently used non-parametric procedures. The histogram method has the disadvantage of discontinuity. Although the method of orthogonal functions is suitable for any type of support, however it produces the Gibbs effect.

In our research, we have opted for the non-parametric kernel method. In order to ensure a good quality of estimation, it is important to maximize the value of the smoothing parameter by minimizing the mean integrated squared error. The optimization of the bandwidth is performed by the diffeomorphism Plug-in algorithm. A direct resolution of the equation for calculating the optimal value of the smoothing parameter seems very difficult [Saoudi09]. The Plug-in method suggested in [Saoudi09] presents the iterative resolution of its equation. Moreover, a fast variant of this algorithm was developed in [Saoudi09].

In the case of densities with bounded or semi-bounded support, the conventional kernel method is no longer adequate and may present convergence problems at the edges: the Gibbs phenomenon. Several authors have attempted to solve this problem and have presented some methods for estimating distributions with bounded or semi-bounded supports. Among them, we can cite the orthogonal functions [Hall82] and the kernel-diffeomorphism estimator [Saoudi09]. This last method, which is derived from the conventional kernel estimator, is based on a suitable variable change by a C1-diffeomorphism. Inspired by the kernel method, it is important to maximize the value of the smoothing parameter in order to ensure good quality of the estimate. The optimization of the bandwidth is performed by the diffeomorphism Plug-in algorithm, which is a generalization of the conventional Plug-in algorithm [Jain00]. However, its implementation presents additional difficulties compared to the classical version.

In order to surpass the complexity reasons of implementation, we introduce in this work a new variant of the diffeomorphism Plug-in algorithm in the semi-bounded case. This version is based on the logarithmic variable change. The divergence to zero problem of the logarithmic function is solved by the addition of a small strictly positive rate.

The rest of this paper is organized as follows. We start by recalling the kernel method for density estimation in section 2 . The third section presents the contribution of

ISBN 978-80-86943-42-8
https://doi.org/10.24132/CSRN.2018.2803.8

the current research which constitutes the fast version of the kernel-diffeomorphism Plug-in algorithm in the semi-bounded case. And in section 4, we have opted for a comparative study between our new variant, the diffeomorphism and the conventional Plug-in algorithms in the semi-bounded case.

2 THE KERNEL ESTIMATE METHOD

The non-parametric Kernel Density Estimator (KDE) has been introduced by Rosenblatt in 1956 [Rosenblatt56] and developed by Parzen in 1962 [Parzen62]. In the case of data with bounded or semi-bounded support , a recent method based on a C1-diffeomorphism variable change has been developed by Saoudi and al. in [Saoudi94, Saoudi97] and called the Kernel-Diffeomorphism density Estimate (KDE). Similarly to the kernel method, the bandwidth optimization is necessary. The Plug-in adapted to this kernel variant is called the Kernel-Diffeomorphism Plug-in (KDP) algorithm.

2.0.1 Kernel-Diffeomorphism Estimator

In this paper, we deal with the semi-bounded densities case. Therefore, a more accurate estimate will be obtained using the KDE which reduces significantly the Gibbs effect [Saoudi94, Saoudi97]. This estimator is a generalization of the KDE. It is suitable for the functions defined on the interval $[a,b]$. The density function is expressed by:

$$\hat{f}_N(x) = \frac{|\Phi'(x)|}{Nh_N} \sum_{i=1}^{N} K(\frac{\Phi(x) - \Phi(X_i)}{h_N}) \qquad (1)$$

where Φ is a C1-diffeomorphism which has the infinity for limit as x approaches a or b. The problematic of optimizing the smoothing parameter can be resolved by using the same methods as those used for conventional kernel analysis. However, as shown in [**?**], an asymptotic study of the KDE, allows better approach to optimal smoothing parameter in the Mean Integrated Squared Error (MISE) sense. Then, its expression becomes the following:

$$h_N^* = [M_\Phi(K)]^{1/5}[J_\Phi(f)]^{-1/5}N^{-1/5} \qquad (2)$$

where $M_\Phi(K) = M(K)\int_R |\Phi'(x)|f(x)dx$ and $J_\Phi(f) = \int_R \frac{F^2(x)}{[\Phi'(x)]^8}dx$

$$\text{F(x)} = [\text{f(x)}[3\Phi''(x)^2 - \Phi'(x)\Phi'''(x)]] - 3f'(x)\Phi'(x)\Phi''(x) + f''(x)[\Phi'(x)]^2 \qquad (3)$$

2.0.2 Kernel-diffeomorphism Plug-in algorithm

The implementation of this extended version presents further difficulties compared to the classical Plug-in algorithm. Indeed, for the conventional Plug-in, $M(K)$ is a constant which can be determined analytically or numerically. As for the KDE adapted plug-in algorithm, $M(K)$ depends on unknown pdf. Similarly, $J(f)$ depends not only on f'', but also on f and f'. Therefore, the complexity of the KDP is increasing. We describe below the kernel-diffeomorphism Plug-in algorithm and its computing complexity:

Step 1: Initialize arbitrary $M_\phi(K)$. In practice $M_\phi^0(K)$ can be equal to $M(K)$.

Step 2: Fix arbitrary $J_\phi^0(f)$, then deduce the first value of the optimal bandwidth; h_N^0. Estimate $f^{(0)}$.

Step 3: Approximate the different quantities: $M_\phi^{(k)}(K)$, $f^{(k)'}$ and $f^{(k)''}$ for each iteration k.

Step 4: Estimate $J_\phi(f^{(k)})$. The value of $h_N^{(k)}$ is so deducted from the k^{th} iteration.

Step 5: Approximate $f^{(k)}$. Stop the algorithm when the difference between $h_N^{(k)}$ and $h_N^{(k-1)}$ is relatively low (below 1%).

Let N be the sample size and p the resolution defined as the point number for which f is estimated. The number of elementary operations is of the order of N^2p which involves a polynomial complexity of $O(N^2)$. Whereas, the conventional Plug-in complexity is linear in the order of $O(N)$.

3 CONTRIBUTION

For simplicity implementation, fast computation and convergence reasons, we introduce, in this section, the Fast Semi-bounded Kernel-Diffeomorphism Estimator (FSKDE). The optimization of the smoothing parameter is performed by the proposed Fast Semi-bounded Kernel-Diffeomorphism Plug-in algorithm (FSKDP). The FSKDP provides a significant improvement in the estimation of the semi-bounded densities. The idea consists on using the logarithmic change of the data qualified by their semi-infinity support: $\Gamma = Log(\zeta)$. Thus, the conventional Plug-in algorithm can now be applied to the transformed data.

3.1 The fast semi-bounded kernel-diffeomorphism estimator

Let's consider independent and identically distributed random variables: $\zeta : \Omega \to R$. Let f_ζ be their probability density function with semi-bounded support: $support(f_\zeta) \subset R$. We recall that the probability densities with semi-bounded support present estimation

 ISBN 978-80-86943-42-8

difficulties due mainly to the Gibbs effect. In order to bypass this divergence limitation at the edge of the semi-bounded interval, we consider a logarithmic variable change for the data: $\Gamma = Log(\zeta)$. The data is then transformed into the following new random variable: $\Gamma : \Omega \to R$. In this case, the distribution function of Γ can be expressed as follows:

$$F_\tau(\Gamma) = P[\Gamma < \tau] = P[Log(\zeta) < \tau] \quad (4)$$

$$F_\tau(\Gamma) = P[\zeta < exp(\Gamma)] = F_\zeta(exp(\Gamma)) \quad (5)$$

The probability density function of Γ can be written as follows:

$$f_\Gamma(\tau) = \frac{dF_\tau(\Gamma)}{d\tau} = f_\zeta(exp(\Gamma))exp(\Gamma) \quad (6)$$

We try to estimate f_ξ as a function of f_Γ. After the change of variable $t = \exp(\Gamma)$, we find f_ζ :

$$f_\zeta(t) = \frac{f_\Gamma(Log(t))}{t} \quad (7)$$

3.2 Convergence problem

The FSKDE presents a specific problem for zero. Indeed, the logarithmic function is defined only on $]0,+\infty[$. It is therefore imperative to submit the data onto a translation before carrying out the logarithmic transformation. This problem is illustrated in figure 1(a). Figure 1(a) represents the density of the exponential distribution $f(x) = \exp(x)$,

The estimation of the exponential density by the FSKDE illustrates the problem described above. Moreover, figure 1(a) shows the divergence at 0 of the estimated exponential distribution.

The idea consists in adding a small positive coefficient ε and thus using the logarithmic variable $Log(x+\varepsilon)$. We notice the resolution of the problem of divergence in 0 as shown in figure 1(b).

3.3 The fast semi-bounded kernel-diffeomorphism Plug-in algorithm

For simplicity implementation, fast computing and convergence reasons, we have introduced a new version of the kernel-diffeomorphism Plug-in algorithm in the case of semi-bounded support. The FSKDP provides a significant improvement in the estimation of the semi-bounded densities. The idea consists on using the logarithm change of the error rates values qualified by their semi-infinity support: $\Gamma = Log(\zeta + \varepsilon)$. Thus, the conventional Plug-in algorithm can now be applied to the transformed data. In order to specify a new classification quality measure, we perform a sequence of three steps:

- **Step 1:** calculate the kernel estimator of the changed variables in the logarithm space: $\Gamma = Log(\zeta + \varepsilon)$
- **Step 2:** iterate the conventional Plug-in algorithm for the transformed data.
- **Step 3:** return to the original space and compute the density kernel estimator: $\hat{f}_\zeta(t) = \frac{\hat{f}_\Gamma(Log(t))}{t}$

The fast diffeomorphism Plug-in algorithm has the same complexity as the conventional one. Thus, the complexity of the FSKDP is linear in the order of $O(N)$.

4 PERFORMANCE STUDY OF THE FAST SEMI-BOUNDED KERNEL-DIFFEOMORPHISM ESTIMATOR

In this part, a comparative study between the KDE and the FSDKE performance is presented in the semi-bounded case. Three semi-bounded distributions are simulated:

- an exponential law of mean 1,
- a first mixture of two laws:
  - a uniform law U(0,1) with a proportion $p_1 = 0.6$.
  - a Gaussian law N(0.8,0.2) with a ratio $p_2 = 0.4$.
- and a second mixture of three laws:
  - a uniform distribution of parameters $U(0,1,5)$ with a proportion $p_1 = 0.4$.
  - a Gaussian law $N(1.3,0.3)$ with a ratio $p_2 = 0.3$.
  - a Gaussian law $N(2.5,0.4)$ with a proportion $p_3 = 0.3$.

Figure 2 below illustrates the theoretical and estimated distributions of the three simulated laws already cited. The estimation is performed using the conventional kernel method. We notice that the estimate overflows from its natural support; the Gibbs phenomenon.

Figure 3 presents the estimation of these probability densities by the KDE. We note that the problem of the Gibbs effect is solved using this method.

Figure 4 presents the estimation of these probability densities by the FSKDP.

The divergence problem in 0 of the FSKDP based on the change of logarithmic variable $Log(x)$ is clearly observed in the first illustration (a) of figure 1. However, we can notice the resolution of this problem by using the logarithmic variable $Log(x+\varepsilon)$. Indeed, the estimation of the probability densities by the FSKDP is almost perfect. These observations are confirmed by the values of the MISE given in the following table. Moreover, the FSKDP has a speed of calculation (having the low execution times) and a better accuracy of estimation (with minimal MISE) with respect to the KDP.

 ISBN 978-80-86943-42-8

5 CONCLUSIONS

To conclude, we have suggested a fast method to estimate probability density functions in the semi-bounded case; the fast semi-bounded kernel-diffeomorphism estimator. This new technique is based on the fast semi-bounded kernel-diffeomorphism Plug-in with a complexity reduced to $O(N)$.

In our future works, we intend to study the case of bounded support distributions. We will also test the new estimators performance in real data.

6 REFERENCES

[Fukunaga13] Fukunaga, K. Introduction to statistical pattern recognition. Academic Press, 2013.

[Hall82] Hall, P. Comparison of two orthogonal series methods of estimating a density and its derivatives on an interval. Journal of multivariate analysis, 12, pp. 432-449, 1982.

[Silverman86] B. W. Silverman. Density estimation for statistics and data analysis. Chapman and Hall, London; New York, 1986.

[Jain00] Jain, A.K., Duin, R.P.W. and Mao, J. Statistical pattern recognition: A review. IEEE Transactions on pattern analysis and machine intelligence, 22, pp. 4-37, 2000.

[Rosenblatt56] Rosenblatt, M., et al. Remarks on some non-parametric estimates of a density function. The Annals of Mathematical Statistics, 27, pp. 832-837, 1956.

[Saoudi97] Saoudi, S., Ghorbel, F. and Hillion, A. Some statistical properties of the kernel-diffeomorphism estimator. Applied Stochastic Models in Busines and Industry 13, pp. 39-58, 1997.

[Saoudi94] Saoudi, S., Hillion, A. and Ghorbel, F. Non-parametric probability density function estimation on a bounded support: Applications to shape classification and speech coding. Applied Stochastic Models in Business and Industry, 10, pp. 215-231, 1994.

[Saoudi09] Saoudi, S. and Troudi, M., Ghorbel, F. An iterative soft bit error rate estimation of any digital communication systems using a non-parametric probability density function. EURASIP Journal on Wireless Communications and Networking, 4, 2009.

[Parzen62] Parzen, Emanuel. On estimation of a probability density function and mode. The annals of mathematical statistics, JSTOR, 33(3), pp. 1065-1076, 1962.

 ISBN 978-80-86943-42-8

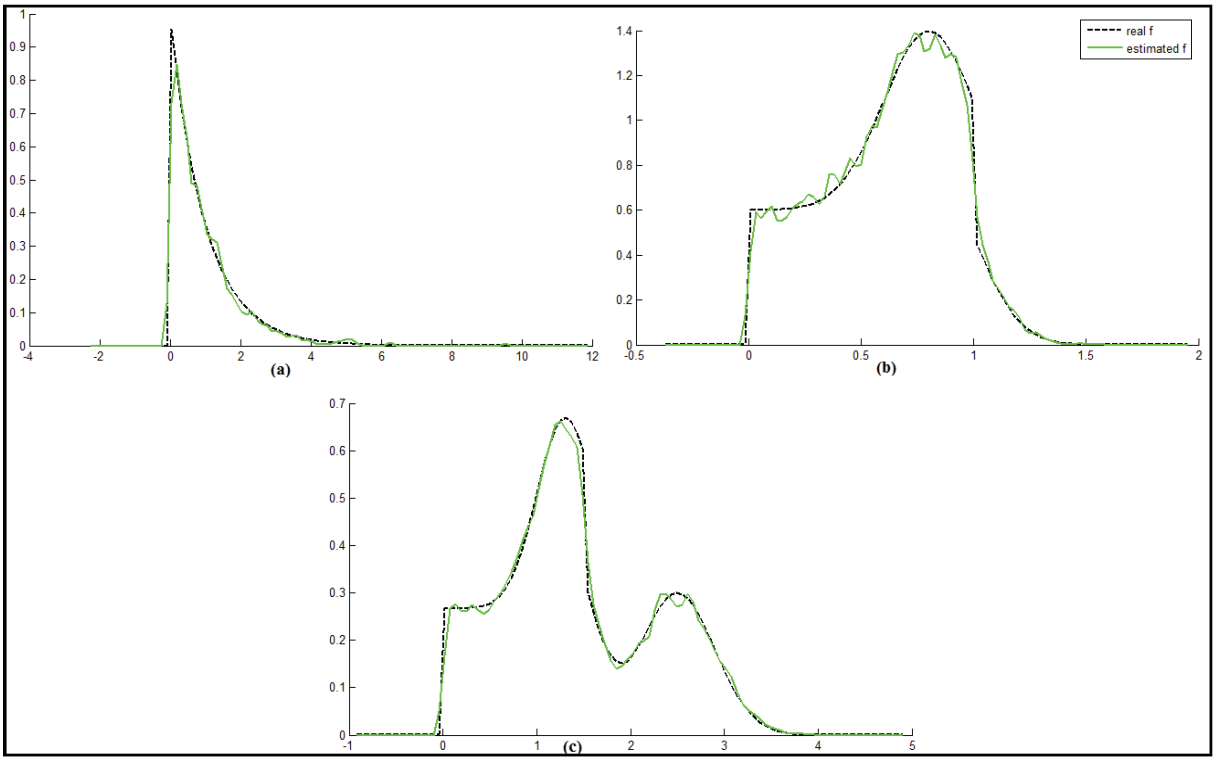

Figure 1: Estimation of the three distributions by the conventional kernel method: (a) the exponential distribution $\exp(x)$, (b) the first mixture and (c) the second mixture.

Method	MISEx10^{-5}	Execution time
KDP	0.1909	0.1248
FSKDP	0.1415	0.0624

Table 1: MISE and execution time of the conventional (KDP) and fast (FSKDP) kernel methods for the estimation of the exponential distribution $\exp(x)$.

Method	MISEx10^{-5}	Execution time
KDP	3.7218	1.0608
FSKDP	2.6252	0.7332

Table 2: MISE and execution time of the conventional (KDP) and fast (FSKDP) kernel methods for the estimation of the first mixture of the uniform law $U(0,1)$ and the Gaussian law $N(0.8,0.2)$.

Method	MISEx10^{-5}	Execution time
KDP	0.4739	1.0764
FSKDP	0.4027	0.7800

Table 3: MISE and execution time of the conventional (KDP) and fast (FSKDP) kernel methods for the estimation of the second mixture of the uniform law $U(0,1,5)$ and two Gaussian laws $N(1,3,0,3)$ and $N(2,5,0,4)$.

ISBN 978-80-86943-42-8

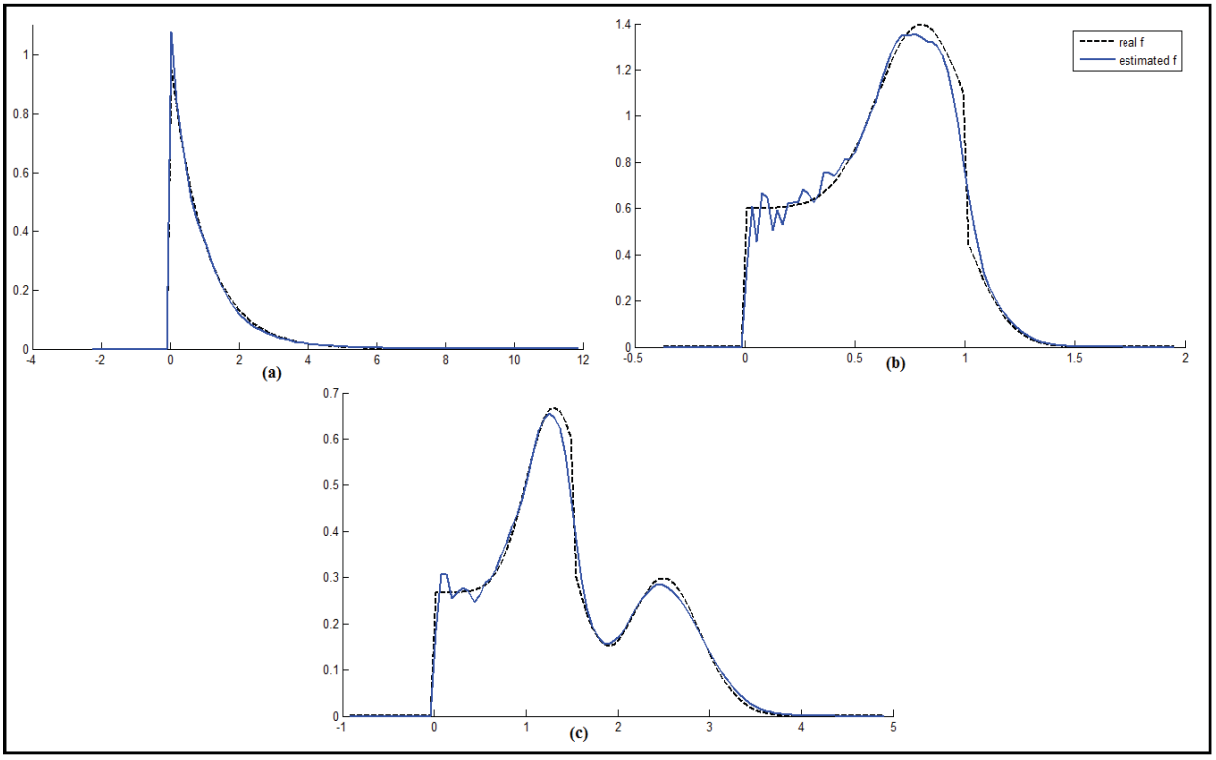

Figure 2: Estimation of the three distributions by the KDP: (a) the exponential distribution $\exp(x)$, (b) the first mixture and (c) the second mixture.

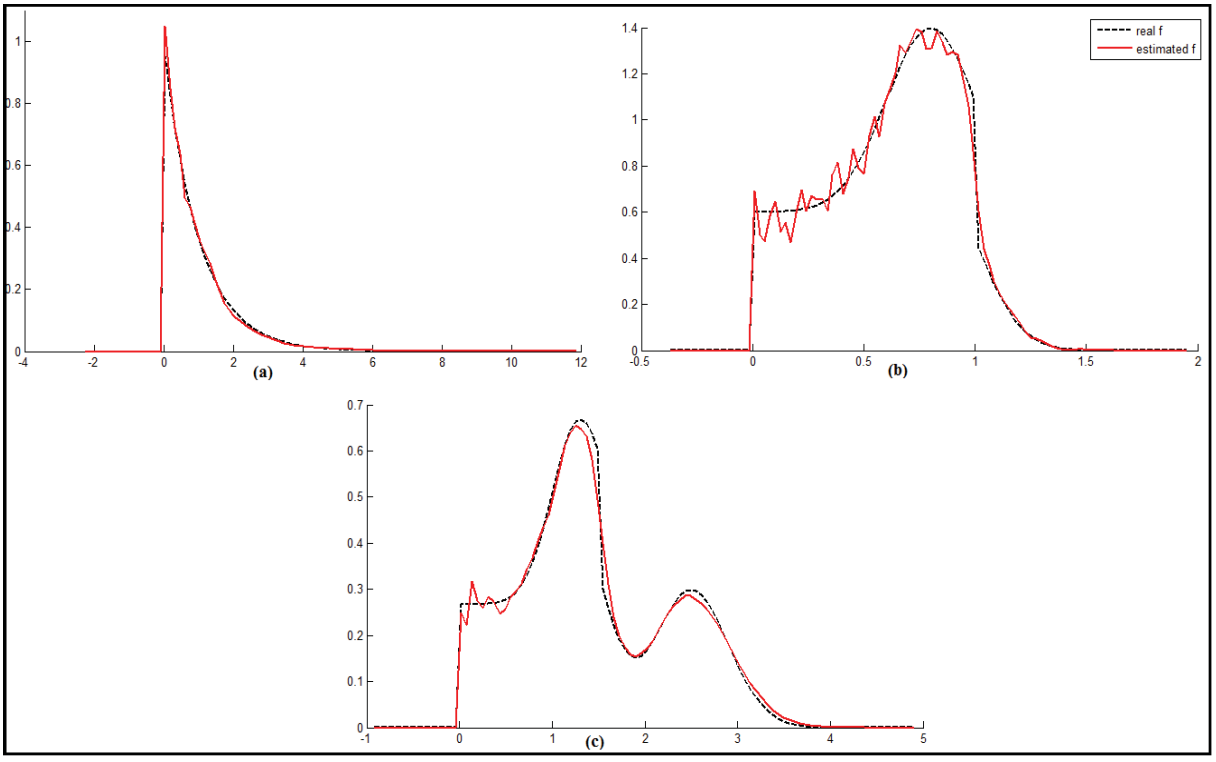

Figure 3: Estimation of the three distributions by the FSKDP: (a) the exponential distribution $\exp(x)$, (b) the first mixture and (c) the second mixture.

ISBN 978-80-86943-42-8