Non-Parametric Estimation of Probability Density Functions via a Simple Interpolation Filter

P. Carbone* and D. Petri†

*University of Perugia - Department of Engineering, via G. Duranti, 93, 06125 Perugia – Italy

†University of Trento, Department of Industrial Engineering, Via Sommarive, 9 Povo, 38123 Trento – Italy

Abstract—In this paper we discuss non-parametric estimation of the probability density function (PDF) of a univariate random variable. This problem has been the subject of a vast amount of scientific literature in many domains: while statisticians are mainly interested in the analysis of the properties of proposed estimators, engineers treat the histogram as a ready-to-use tool for dataset analysis. By considering histogram data as a numerical sequence, a simple PDF estimator is presented in this paper. It is based on basic notions related to the reconstruction of a continuous-time signal from a sequence of samples and it is as accurate as kernel-based estimators, widely adopted in the statistical literature. The major properties of the proposed PDF estimator are discussed and then verified by simulations related to the common case of a normal density function.

I. INTRODUCTION

The estimation of the Probability Density Function (PDF) of a random variable is a well known problem analyzed both in the statistical and engineering literature with a great amount of technical details. The simple histogram-based estimator yields a discontinuous stepwise behavior and provides roughly quantized estimates of the original PDF [1]. Conversely, the less popular, at least in the engineering domain, kernel-based (KB) estimators may provide smoother estimates at the price of increased bias and additional complexity in data processing [2], [3]. Moreover they leave the user with the additional requirement of setting the kernel type [4], [5], [6]. While the expert user may easily interpret the estimator output so to select the best combination of all possible estimator parameters, the less advanced practitioner may not be aware of the risk in interpreting PDF estimation results uncritically. Thus, he/she might be interested in automating data processing so to mitigate the occurrence of this risk. The issue of automatically selecting the optimal estimator parameters was analyzed for instance in [7], [8] with respect to optimal bin width selection and in [9] for the optimization of kernel-based estimation. In [6] a different method was proposed that is blind and adaptive, so that a smooth estimate is achieved.

The purpose of this paper is to show how very simple digital signal processing techniques, such as numerical filtering and linear interpolation, may provide PDF estimates with improved statistical properties over histogram and close to what can be obtained using kernel-based estimators. The main idea behind the proposed approach is firstly introduced, then its properties are analyzed both through theoretical and simulation results. It is shown that the main advantage of the presented approach is

its simplicity both in terms of needed theoretical background and in computational complexity required to process the data.

II. NON-PARAMETRIC ESTIMATION OF A PDF

While several methods were proposed to provide PDF estimate, the histogram is certainly the simplest and also the most readily available in the majority of signal processing software applications. However, its application may result in poor estimates. As an example, in Fig. 1(a) it is shown the mean value of a set of histograms related to of a zero-mean normal PDF with standard deviation $\sigma = 0.8$. While the x -axis is continuous, the histogram discretizes it, so that the bias (that is the difference between the mean of the estimated PDF and the true PDF) exhibits the erratic behavior, typical of discretization, as shown in Fig. 1(b) [10], [11], [12], [13].

Fig. 1. Histogram of the PDF of a zero-mean Gaussian random variable with standard deviation equal to 0.8, (a) mean value of the estimator and true PDF with details in the inset; (b) estimator bias. Estimates are based on 1000 records of 500 independent realizations.

Clearly, a smaller value of the bin width would result in an overall smaller PDF estimator bias, but also in a larger variance. To remove this problem various smoothing techniques were proposed, the most widely used ones are probably those

based on kernel functions [2]. It is shown next how to get results very close to those obtained using kernels but applying a very simple numerical interpolation filter.

III. A SIMPLE PDF ESTIMATOR BASED ON AN INTERPOLATION FILTER

In this section we firstly analyze the main properties of the classical histogram and we then propose a simple filtering approach resulting in estimates that, in the average, are close to those coming from the kernel-based approach [2]–[6]. The classical histogram estimates the PDF of a random variable by first partitioning the expected range of values in non-overlapping subintervals of constant-size dimension Δ , representing the histogram bin width and defined by

$$I_n = \left(n\Delta - \frac{\Delta}{2}, n\Delta + \frac{\Delta}{2} \right], \quad (1)$$

where n is an integer value. Then, the N outcomes $x_0, \dots, x_{N-1}$ of the random variable are collected. By defining N_n as the count of such outcomes falling in the n -th subinterval I_n , an expression for the conventional histogram-based estimator of the PDF $f(x)$ is:

$$h(x) = \frac{1}{N\Delta} N_{[\frac{x}{\Delta}]} \quad (2)$$

where $[x]$ is obtained from a real value x by rounding it to the closest integer value. While (2) does not include possible offset values in the positioning of adjacent intervals, the analysis provided here can easily be extended to also include this case. Because of the discontinuous behavior of the rounding function, (2) is a piecewise constant function of x . While $f(x)$ might *truly* exhibit such behavior, the user is frequently required to estimate continuous PDFs. So, a much smoother estimate than the one provided by (2) is of interest. To this aim, various approaches were proposed, that make use of *kernels* [2].

Conversely, since $h(\cdot)$ in (2) is stepwise constant it can be represented without loss of information, by the sequence of its amplitudes $h(n\Delta)$, called in the following *discrete histogram*. A smoothed PDF estimate can then be obtained by applying a moving average filter to this sequence followed by linear interpolation provided by a first-order hold (FOH), as shown in Fig. 2. Clearly, not to *delay* the filter output with respect to its input, a zero-phase FOH filter must be used. Assuming

Fig. 2. Proposed signal processing chain for pdf estimation: the classical histogram is considered as a numerical sequence, filtered by a zero-phase moving average and linearly interpolated by a first-order hold.

that the number of filtered samples is $2M + 1$, the filter output sequence becomes

$$\hat{f}_M(x) = \frac{1}{2M+1} \sum_{n=-M}^M h\left(\left(\left[\frac{x}{\Delta}\right] + n\right)\Delta\right) \quad (3)$$

and the interpolation filter-based (IFB) PDF estimator is:

$$\hat{f}(x) = (\hat{f}_M * g)(x), \quad (4)$$

where

$$g(x) = \begin{cases} 1 - \frac{|x|}{\Delta} & |x| < \Delta \\ 0 & \text{otherwise} \end{cases}$$

As (3) shows, any filter output sample is the result of the mean value of $2M + 1$ histogram samples symmetrically chosen around the considered one. Observe also that, as it happens when filtering discrete time sequences, M is a *smoothing* parameter whose value is the result of a trade-off between smoothness and resolution requirements. Indeed when M is increased a beneficial estimator variance reduction is achieved but, at the same time, estimation bias may become excessive. Obviously, while (3) and (4) remain one of the possibly most simplest processing activities that can be performed, the proposed approach can be generalized to any other zero-phase interpolation filter, so that improved estimator performance can be achieved.

IV. PDF ESTIMATOR BIAS AND VARIANCE

The basic performance characterization of an estimator requires assessment of its bias and variance as well as computational complexity and robustness with respect to the assumed hypotheses. To gain insight in the behavior of the interpolation-filter-based estimator we firstly assess the bias and variance of the *filtered discrete* histogram $\hat{f}_M(x)$. It is expected that the interpolation-filter-based estimator will have improved properties, so that what is derived in subsection A and B will provide useful insight on the bias and variance of $\hat{f}(x)$, which will be explored in subsection C, by using Montecarlo simulations.

A. Bias of the filtered discrete histogram

The properties of $\hat{f}_M(x)$ depend on those of the sequence of random variables $N_{[\cdot]}$ in (2). By defining $\bar{N} = [\dots N_{-1} N_0 N_1 \dots]^T$ as the column vector containing all filter *input samples* it can be proven that $\bar{N}$ is a random vector having multinomial distribution, for which [14]:

$$E(\bar{N}) = [\dots Np_{-1}, Np_0, Np_1 \dots]^T$$

where $p_{[\frac{x}{\Delta}]} = \int_{\Delta[\frac{x}{\Delta}]-\frac{\Delta}{2}}^{\Delta[\frac{x}{\Delta}]+\frac{\Delta}{2}} f(x)dx$. It is also known that the variance of N_i is $Np_i(1-p_i)$ and that the covariance between N_i and N_j with $i \neq j$, is $-Np_i p_j$ [14].

Fig. 3. Verification of (7) and its approximated version. Estimates were obtained by using 15000 Montecarlo records, each one containing 500 realizations of a Gaussian random variable with zero-mean and standard deviation 0.8.

Thus, we obtain:

$$\begin{aligned}
 E(\hat{f}_M(x)) &= \frac{1}{(2M+1)\Delta} \sum_{n=-M}^M p\left(\left[\frac{x}{\Delta}\right] + n\right) \\
 &= \frac{1}{(2M+1)\Delta} \sum_{n=-M}^M \int_{\Delta\left[\frac{x}{\Delta}\right] + (n-\frac{1}{2})\Delta}^{\Delta\left[\frac{x}{\Delta}\right] + (n+\frac{1}{2})\Delta} f(x) dx \\
 &= \frac{1}{(2M+1)\Delta} \int_{\Delta\left[\frac{x}{\Delta}\right] - (M+\frac{1}{2})\Delta}^{\Delta\left[\frac{x}{\Delta}\right] + (M+\frac{1}{2})\Delta} f(x) dx \\
 &= \frac{1}{(2M+1)\Delta} \left\{ F\left(\Delta\left[\frac{x}{\Delta}\right] + \left(M + \frac{1}{2}\right)\Delta\right) - F\left(\Delta\left[\frac{x}{\Delta}\right] - \left(M + \frac{1}{2}\right)\Delta\right) \right\}, \quad (5)
 \end{aligned}$$

where $F(\cdot)$ is the cumulative distribution function of the considered random variable. Thus, the expected value of the estimator $\hat{f}_M(x)$ becomes equal to the discrete derivative of the cumulative distribution function. Observe also that when $M = 0$, the filter remains unapplied and (5) returns the expected value of the histogram-based estimator. While (5) provides exact values for the estimator bias once $f(x)$ is subtracted, the result in the Appendix shows that it can be approximated as:

$$E(\hat{f}_M(x)) - f(x) \simeq \frac{(2M+1)\Delta}{12} f''\left(\Delta\left[\frac{x}{\Delta}\right]\right). \quad (6)$$

B. Variance of the filtered discrete histogram

By using the well known property of the variance of a summation of random variables, the variance of $\hat{f}_M(x)$ can be calculated on the basis of the variances and the covariances of the entries of vector $\bar{N}$:

$$\begin{aligned}
 \text{Var}(\hat{f}_M(x)) &= \frac{1}{N(2M+1)^2\Delta^2} \left(\sum_{n=-M}^M p\left(\left[\frac{x}{\Delta}\right] + n\right) \times \right. \\
 &\quad \left. (1 - p\left(\left[\frac{x}{\Delta}\right] + n\right)) - 2 \sum_{i=-M+1}^M \sum_{j=-M}^{i-1} p\left(\left[\frac{x}{\Delta}\right] + j\right) p\left(\left[\frac{x}{\Delta}\right] + i\right) \right) \quad (7) \\
 &\simeq \frac{1}{N(2M+1)\Delta} f(x) [1 - f(x)(2M+1)\Delta],
 \end{aligned}$$

where the approximation follows by observing that, since small values of $2M\Delta$ are considered, the PDF to be estimated can be assumed constant over intervals of width $2M\Delta$. In this case, $p\left(\left[\frac{x}{\Delta}\right] + n\right) \simeq f(x)\Delta$ independently of n . Montecarlo runs were carried out to verify (5), (7) and the approximate variance given by the rightmost term in (7). Obtained results are plotted in Fig. 3 where both the exact and the approximated versions of the variance of the filtered histogram are also graphed. In the inset it is shown the detail around the PDF peak. It is clear that a good matching between all curves is always achieved.

C. Basic performance of the interpolation-filter PDF estimator

To show how the simple approach proposed in this paper outperforms the accuracy of the histogram-based estimator and becomes a competing technique with the kernel based estimator, Montecarlo runs based on 1000 records, each one containing 500 realizations of a Gaussian random variable with zero-mean and standard deviation $\sigma = 0.8$, were performed. Results obtained for $2M+1 = 3$ or $2M+1 = 5$ are shown in Fig. 4 and Fig. 5, respectively. In the same figures, also the PDF estimates obtained by applying the Gaussian kernel approach are reported for comparison. In subfigures (a) of each figure, the mean values of the histogram, interpolation-filter-based and Gaussian kernel-based estimators are shown along with the behavior of the true normal PDF.

In subfigures (b) the bias of each estimator is reported, while subfigures (c) show the corresponding Root-Mean-Square Errors (RMSE). The choice of the histogram resolution Δ was made according to the Silverman rule [7], [15]

$$\Delta = 1.06\sigma N^{-\frac{1}{5}},$$

which provides $\Delta = 0.2447$ in the considered case.

Subfigures (a) show that both the interpolation-filter-based and the kernel-based PDF estimators underestimate the PDF peak, as also shown in subfigures (b). This underestimation is more severe when $2M+1 = 5$, as the PDF peak is smoothed more sharply by the moving average filter. Subfigures (a) also show that both the interpolation-filter-based and the kernel-based estimators provide a smoother behavior than the stepwise behavior associated to the histogram. If the RMSE is considered as the optimality criterion, the comparison of subfigures (b) and (c) shows that tolerating some *bias* can results in the reduction of estimator variance.

Fig. 4(b) and 4(c) show that, by choosing properly the smoothing constant M , the simple interpolation-filter-based

Fig. 4. Estimation of the PDF of a zero-mean Gaussian random variable with standard deviation equal to 0.8: histogram-based (point-dash line), Gaussian kernel-based (dashed line) and interpolation-filter-based (solid line, $2M + 1 = 3$ -tap moving average): (a) mean value of the estimator and true PDF along with details in the inset; (b) estimator bias; (c) estimator RMSE. Estimates were obtained by using 1000 records, each containing 500 realizations of a Gaussian random variable with zero-mean and standard deviation 0.8.

estimator performs very similarly to the kernel-based estimator. Conversely, when M is made too large, as in Fig. 5, it smooths too much the histogram so introducing a larger bias than the kernel-based estimator, even if it remains overall superior in performance with respect to the simple histogram (see Fig.5(c)).

Anyway the choice of M is critical for ensuring accurate results. Given that M determines the bandwidth of the filter applied to the histogram (that is, to a quantized version of the original PDF) as a rule of thumb M should be chosen so that the filter bandwidth allows for the major portion of the characteristic function associated to the PDF falling within the filter pass-band. As a final remark consider that linearly interpolating the histogram provides RMSE values close to the lower envelope of the curves in subfigures (c), and so not

Fig. 5. Estimation of the PDF of a zero-mean Gaussian random variable with standard deviation equal to 0.8: histogram-based (point-dash line), Gaussian kernel-based (dashed line) and interpolation-filter-based (solid line, $2M + 1 = 5$ -tap moving average): (a) mean value of the estimator and true PDF along with details in the inset; (b) estimator bias; (c) estimator RMSE. Estimates were obtained by using 1000 each containing 500 realizations of a Gaussian random variable with zero-mean and standard deviation 0.8.

as good as those obtained by the interpolation-filter-based PDF estimator. Indeed the filter, while increasing the bias, removes some of the variance of the histogram and thus yields an overall better RMSE.

D. Main drawback of the proposed estimator

Simulations showed that the interpolation-filter-based and the kernel-based estimators exhibit very similar performance. Since both estimators smooth the data, they both tend to behave poorly when the PDF has sharp discontinuities as in the case of uniform or arcsine PDFs. To support this statement consider for instance the arcsine distribution, whose PDF has the following expression:

$$f(x) = \begin{cases} \frac{1}{\pi\sqrt{1-x^2}} & -1 < x < 1 \\ 0 & \text{otherwise.} \end{cases} \quad (8)$$

Fig. 6. Arcsine PDF given by (8): mean values of the three considered estimators (histogram, interpolation-filter-based and kernel-based estimators) obtained using Montecarlo simulations based on 1000 records of 500 samples each and $\Delta = 0.1529$. All estimators poorly follow the discontinuous behavior of the true PDF at $x = -1$, shown using a solid line.

which exhibits a sharp discontinuity when $x = \pm 1$. A comparison between the three estimators (histogram, interpolation-filter-based and kernel-based estimators) was carried out using Montecarlo simulations based on 1000 records of 500 samples each. The value of $\Delta = 0.1529$ obtained by using the Silverman's rule was chosen because it minimizes the histogram RMSE, as also verified through simulations based on the PDF (8). The mean values of the obtained estimators are plotted in Fig. 6. As anticipated, none of the estimators follow the bath-tube behavior of the true PDF. In this case other more involved techniques can be adopted as described in [6].

V. CONCLUSION

By treating histogram data as a numerical sequence a very simple and computationally light PDF estimator has been proposed in this paper that is based on an intuitive filtering and interpolating approach. The main advantage of the proposed estimator with respect to other state-of-the-art estimators is that it requires the use of only basic signal processing concepts that are expected to be available in the toolset of any engineer.

In particular, a zero-phase moving average filter has been applied to reduce estimator variance, followed by a first-order hold performing linear interpolation between filtered samples. It is shown that this simple estimator offers comparable accuracy to the well-known kernel-based estimator. While the main idea is presented in this paper, higher PDF estimator performances are expected when applying more sophisticated signal processing procedures proposed in the literature for the reconstruction of an analog signal from a sequence of samples [16].

APPENDIX

APPROXIMATION OF THE INTEGRAL OF A FUNCTION

Consider that by expanding $f(x)$ using a Taylor series about $x = x_0$ we have:

$$f(x) \simeq f(x_0) + f'(x_0)(x - x_0) + \frac{1}{2}f''(x_0)(x - x_0)^2,$$

so that when $\delta \simeq 0$,

$$\int_{x_0-\delta}^{x_0+\delta} f(x)dx \simeq 2\delta f(x_0) + \frac{1}{3}f''(x_0)\delta^2. \quad (\text{A.1})$$

Applying (A.1) to (5), (6) can easily be derived.

REFERENCES

- [1] B. Widrow, I. Kollár, Statistical Theory of Quantization, *IEEE Trans. Instrum. Meas.*, vol. 45, no. 2, pp. 353–361, April 1996.
- [2] E. Parzen, “On the estimation of a probability density function and the mode,” *Annals of Math. Stats.*, vol. 33, pp. 1065–1076, 1962.
- [3] A. Azzalini, “A note on the estimation of a distribution function and quantiles by a kernel method,” *Biometrika*, vol. 68, no. 1, pp. 326–328, 1981.
- [4] V. A. Epanechnikov, “Non-parametric estimation of a multivariate probability density,” *Theory Probab. Appl.*, vol. 14, no. 1, pp. 153–158, 1969.
- [5] K. Barbé, L. Gonzales Fuentes, L. Barford, W. Van Moer, “Eliminating User-Interaction in Probability Density Estimation,” *proc. Interna. Instrum. Meas. Conf.*, Minneapolis, pp. 1295–1300, 2013, 10.1109/I2MTC.2013.6555622.
- [6] K. Barbé, L. G. Fuentes, L. Barford, L. Lauwers, “A Guaranteed Blind and Automatic Probability Density Estimation of Raw Measurements,” *IEEE Trans. Instrum. Meas.*, vol. 63, no. 9, pp. 2120–2128, Sept. 2014.
- [7] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London, 1986.
- [8] I. Lira, “Assigning a probability density function for the value of a quantity based on discrete data: The resolution problem,” *Metrologia*, vol. 49, no. 6, pp. 765–771, 2012.
- [9] H. Shimazaki, S. Shinomoto, “Kernel bandwidth optimization in spike rate estimation,” *Journal of Comput Neuroscience*, vol. 29, pp. 171–182, 2010, DOI 10.1007/s10827-009-0180-4.
- [10] D. Petri, P. Carbone, “Effect of additive dither on the resolution of ideal quantizers,” *IEEE Trans. Instrum. Meas.*, vol. 43, no. 3, pp. 389–396, June 1994.
- [11] P. Carbone, C. Narduzzi, D. Petri, “Dither signal effects on the resolution of nonlinear quantizers,” *IEEE Trans. on Instr. Meas.*, vol. 43, n. 2, pp. 139–145, April 1994.
- [12] P. Carbone, D. Petri, “Performance of stochastic and deterministic dithered quantizers,” *IEEE Trans. on Instrumentation and Measurement*, Vol. 49, no. 2, pp. 337–340, Apr. 2000.
- [13] M. Bertocco, C. Narduzzi, C. Paglierani, D. Petri, “A noise model for digitized data,” *IEEE Trans. on Instrumentation and Measurement*, Vol. 49, no. 1, pp. 83–86, Feb. 2000.
- [14] M. Evans, N. Hastings, B. Peacock, *Statistical Distributions – 3rd ed.*, Wiley, New York, USA 2000.
- [15] S. J. Sheather, “Density Estimation,” *Statistical Science* vol. 19, No. 4, pp. 588–597, 2004, DOI 10.1214/088342304000000297.
- [16] A. Papoulis, S.U. Pillai *Probability, Random Variables and Stochastic Processes, 4th ed. Intern. ed.*, McGraw-Hill Publ.Comp, Jan. 2002.