«دانش مدیریت»
سال سیزدهم - شماره ۴۹ - تابستان ۱۳۷۹
ص ص ۱۱۶-۱۳۹

Statistical Errors in Digital Estimation of Probability Density Functions

Nezameddin Faghih Ph.D.[1]

Abstract

An assessment of the statistical errors occurring in the digital estimation of probability density function maybe required in a variety of applied disciplines; eg. engineering, management and maintenance planning. In special, for such purposes, this paper studies the statistical errors in the computer estimation of probability density functions, by simulation of correlated and uncorrelated data. The estimate error, being composed of a random portion and a bias term is analyzed and applied to the cases of uniform and standard Gaussian density functions. Then further investigation is undertaken by employing simulated data with various conditions.

Introduction

The probability density function is one of the main types of statistical criteria used to describe the basic properties of random data. Hance, in the study of random data, an estimation of the

---

1 - Professor of Shiraz University

probability density function may be required. In practice, especially with the aid of digital computers, this is accomplished by dividing an appropriate range into a number of intervals and observing a finite sample size. The accuracy of the estimation then depends on the sample size and the window size.

The digital procedure for probability density estimation is given in various sources, e. g. [1], [2], [3] and a brief account of it may be found in Appendix A.

Consider an estimate p(x) obtained for a true probability density function p(x). For continuous signals of bandwidth B, the normalized mean square error of the estimate, ε_m , is given by the following equation [2]:

$$\varepsilon^2 \cong_m \frac{c^2}{2BTW\ p(x)} + \frac{w^4}{576}\left[\frac{\rho''(x)}{\rho(x)}\right]^2 \qquad (1)$$

Where W is the window size and p"(x) denotes the second derivative of p(x) with respect to x. In the above equation. Which is given for a continuous record of length T, c is a constant depending on the autocorrelation function of the signal. For discrete uncorrelated data. 2BT may be replaced by the number of samples and a value 1.0 has been given for c as would be expected theoretically [2].

The above formula, however, seems to require further study with regards to the effects of the second term and variations of the error with sample and window sizes, for probability density

estimates of discrete time series.

In this paper, the case of discrete data is considered. The estimate error is analysed and its expression is approached. This is, basically, in accordance with the analysis concerning the continuous case. It is then applied to the uniform and standard Gaussian probability densities and shown that the second term (which includes second order differential) is zero for the former case and negligible for the latter. Further, the probability density estimate error is studied by simulations on a digital computer and comparison of theoretical and empirical results.

Analysis of the Estimation Errors

The digital estimator for a probability density function is given by equation (A.2.). It is known that this estimator is biased and the bias of the estimate is given [2] as:

$$b\,[\hat{p}(x)] \cong \frac{W^2}{24}\,p''(x) \qquad (2)$$

It is also shown [2] that the variance of the estimate, based upon observing N independent sample values, is given by:

$$\mathrm{Var}\,[p(x)] = \frac{p(x)}{NW} \qquad (3)$$

The mean square error can be written as:

$$E\,[(\hat{p}(x) - p(x)^2] = Var\,[\hat{p}(x)] + b^2\,[\hat{p}(x)] \qquad (4)$$

Where E denotes the expectiation operator. This, after substitution from equations (2) and (3), yields:

$$E\,[(\hat{p}(x) - p(x)^2] = \frac{p(x)}{NW} + \frac{W^4}{576}\,[p''(x)]^2 \qquad (5)$$

Hance, the normalized mean square error of the estimate is:

$$\varepsilon^2_m = \frac{E\,[(\hat{p}(x) - p(x))^2]}{p^2(x)} \cong \frac{1}{NWp(x)} + \frac{W^4}{576} - [\frac{p''(x)}{p(x)}]^2 \qquad (6)$$

The derivation of the above equation is based upon observing N independent sample values. If the sample are correlated then this expression is not appropriate and it is sometimes suggested that, in this case, the expression is replaced by:

$$\varepsilon^2_m \cong \frac{c2}{NWp(x)} + \frac{W4}{576}\,[\frac{p''(x)}{p(x)}]^2 \qquad (7)$$

Where N is the sample size and the constant c is dependent upon the autocorrelation function of the data and the sampling period [2]. However, when N represents the number of independent samples then equation (7) will reduce to (6). with c being equal to unity.

Appliction to the Uniform Probability Density Function

Consider a uniform probability density function in the range (0,1), for which:

$$p(x) = 1,\ 0 \leq x \leq 1 \qquad (8)$$
$$= 0, \text{ otherwise}$$

Since $p''(x) = 0$, the bias term is zero and the normalized root mean square error becomes the same as the normalized standard error. Hence, equation (7) will not include the bias term and will reduce to:

$$\varepsilon_m = \varepsilon_r \cong \frac{c}{\sqrt{NW}} \qquad (9)$$

Where ε_r is the normalized standard error or the relative error.

Application to the Gaussian Probability Density Function

Consider a Gaussian density funciton in the standard from, i.e. with zero mean and unit varicance, for which:

$$p(x) = (\sqrt{2\pi})^{-1} \exp(-x\frac{2}{2}) \qquad (10)$$

This, when differentiated twice with respect to x gives:

$$[p''(x)]^2 = [(1 - x^2)\ p(x)]^2 \qquad (11)$$

The Maximum value of which occurs at x = 0 and is:

$$\text{Max } [p''(x)]^2 = \frac{1}{2\pi} \tag{12}$$

Hance, from equation (2) the maximum value of the square of the bias term is :

$$\text{Max } \{b^2 [\hat{p}(x)]\} = \frac{W^4}{1152\pi} \tag{13}$$

For a Gaussian distribution, about 99.9% of the data fall in the range (- 30 , 30), where o′ is the standard deviation [4]; thus, for data with unit standard deviation, dividing this practical range into M_s slots gives the window size as:

$$W = \frac{6}{M_s} \tag{14}$$

Then, equation (13) gives:

$$\text{Max } \{b^2 [\hat{p}(x)]\} = \frac{1296}{1152\,\pi M_s^4} \tag{15}$$

If the range has been divided into, say, ten slots, then:

$$\text{Max } \{b^2 [\hat{p}(x)]\} \approx (10^{-5}) \tag{16}$$

Therefore, the bias term appean to be negligible. The

normalized root mean square error than becomes the same as the normalized standard error. Hance, equation (7) may be reduced to:

$$\varepsilon_m \cong \varepsilon_r \cong \frac{c}{\sqrt{NWp(x)}} \tag{17}$$

or equivalently:

$$\frac{[\hat{p}(x) - p(x)]^2}{p(x)} = \frac{c_2}{NW} \tag{18}$$

Minimum and Maximum Errors

The minimum and maximum values of the estimate errors for a standard Gaussian density, may be evaluated from equation (17), as follows. This equation shows that the relative error is a minimum when p(x) is a maximum and vice-versa. The maximum value of p(x), from equation (10), occurs at x=0 and gives:

$$\varepsilon_{min} \cong \frac{1.58c}{\sqrt{NW}} \tag{19}$$

If $x = \pm 30$ is taken as an upper bound of practical interest then, using equations (10) and (17):

$$\varepsilon_{max} \cong \frac{15.02c}{\sqrt{NW}} \tag{20}$$

Hence, for the same sample and window size:

$$\frac{\varepsilon_{max}}{\varepsilon_{min}} \cong 9.5 \qquad (21)$$

Simulation Studies

The digital procedure for probability density estimation was programmed on a digital computer. Independent random variables with a uniform distribution in the range (0,1), independent random variable with a standard Gaussian distribution and also correlated Gaussian processes were simulated on a digital computer, according to the methods described in [5], [6], [7]. The probability density functions were estimated, in the cases of uncorrelated data with various window and sample sizes, and in the cases of correlated data for different sampling periods.

For the uniform density case, sample sizes of, 1000, 5000 to 100000 (in steps of 5000) were chosen. Different window sizes were used and, in each case, the estimation error was calculated according to:

$$\varepsilon_r = \sqrt{\frac{1}{M_s} \{\sum_{n=1}^{M} [p(n) - 1]^2} \qquad (22)$$

M_s is the number of slots, p(n) is the estimate of the probability density in the n^{th} slot and also 1 inside the brackets represents the true value of the uniform density function. For each window size, the relative error ε_r was plotted against the sample size and compared with the error expected theoretically, as given by equation (9); in this equation the constant c was put

equal to unity, since independent random numbers were being used.

Furthermore, in each case, the quantity $\varepsilon_r \sqrt{NM}$ was also calculated, where N is the sample size and $W = \frac{1}{M_s}$ is the window size (it is noted that the density function is in the range 0 to 1). This would give the proportionality constant, c, which is expected to be about unity.

The simulation results, with the range divided into 20 slots, are shown in Figures 1 and 2. Figure 1 shows ε_r plotted against the sample size and Figure 2 shows $\varepsilon_r \sqrt{NW}$ plotted versus the sample size. Using 30,40 and 50 slots, the corresponding results are shown in Figures 3, 4, 5, 6 and 7, 8 respectively.

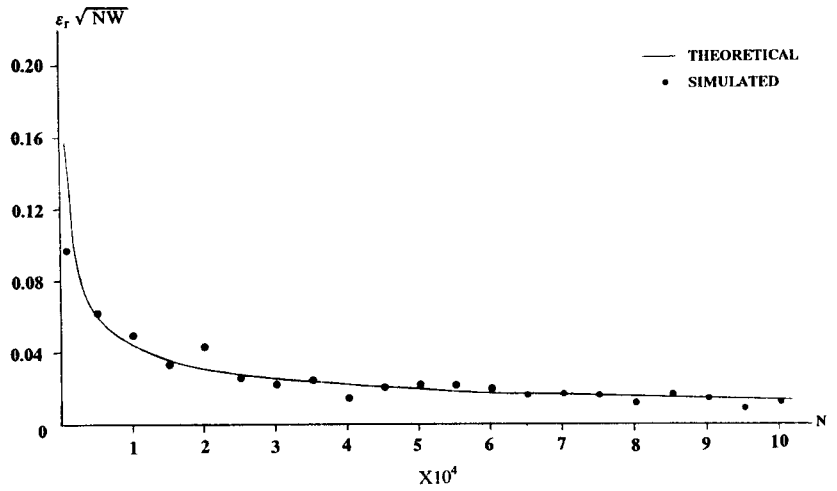

Fig. 1 - Relative error versus sample size, for uniform probability density estimates.

Number of slots = 20

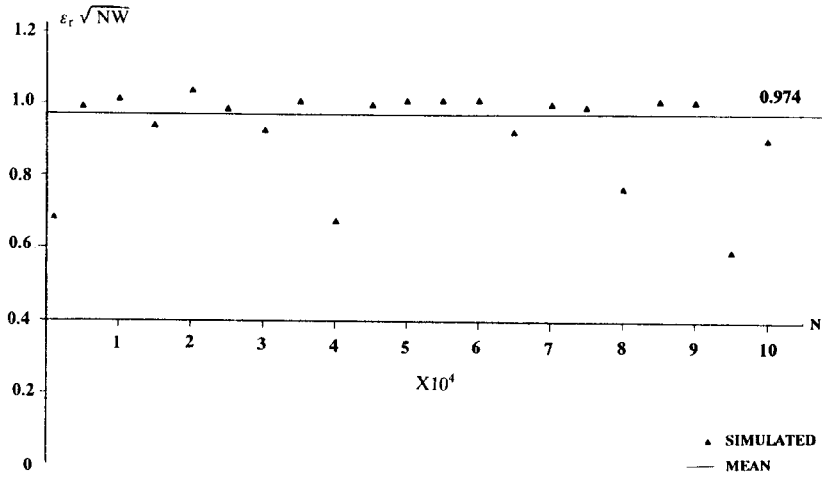

Fig. 2 - Relative error proportionality constant versus sample size, for uniform probability density estimates. Number of slots = 20

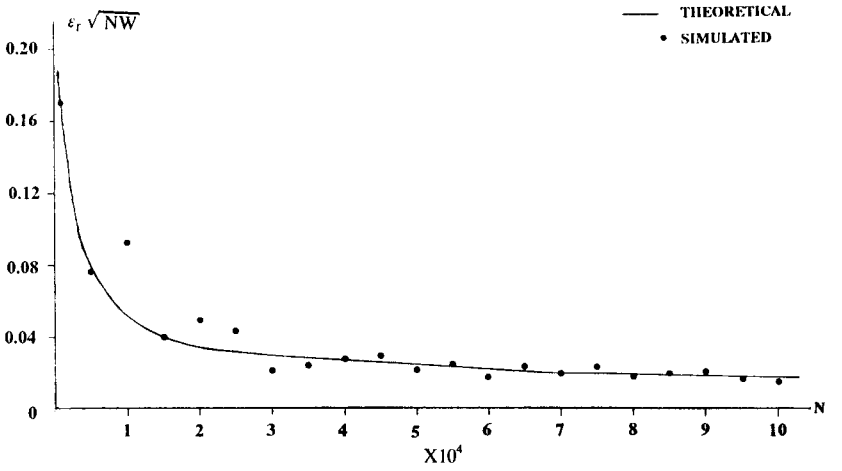

Fig. 3- Relative error versus sample size, for uniform probability density estimates. Number of slots = 30

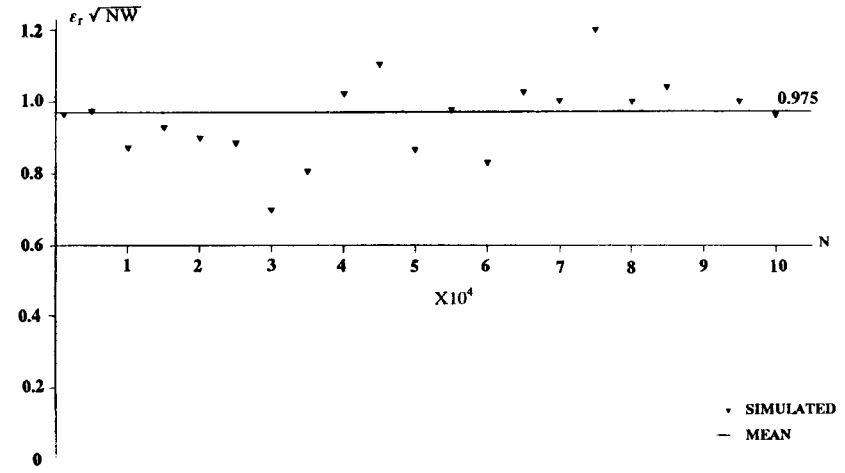

Fig. 4- Relative error proportionality constant versus sample size, for uniform probability density estimates. Number of slots = 30

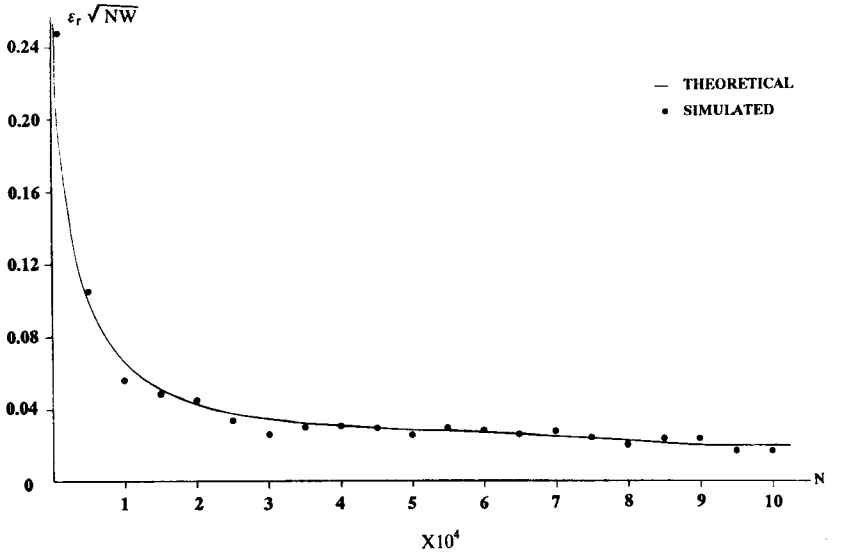

Fig. 5- Relative error versus sample size, for uniform probability density estimates. Number of slots = 40

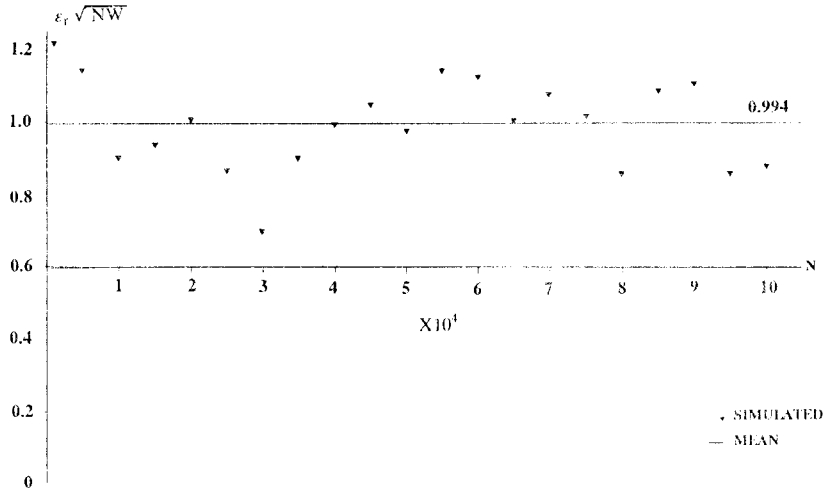

Fig. 6- Relative error proportionality constant versus sample size, for uniform probability density estimates. Number of slots = 40

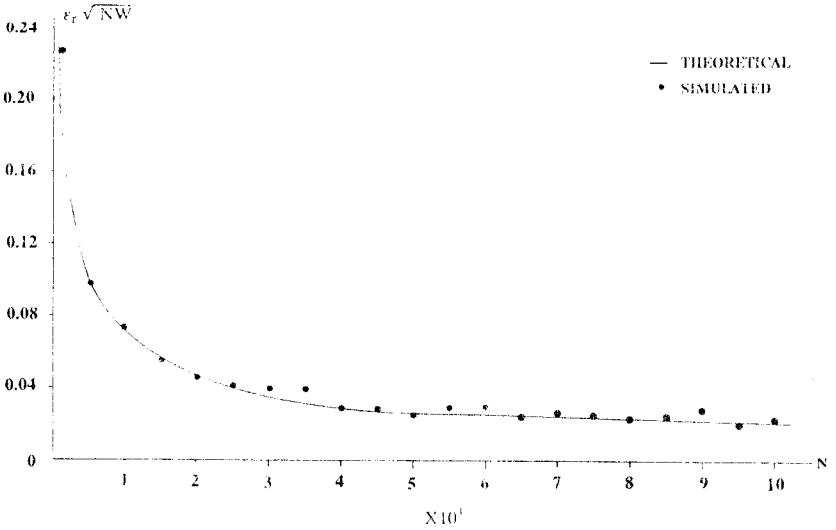

Fig. 7- Relative error versus sample size, for uniform probability density estimates. Number of slots = 50

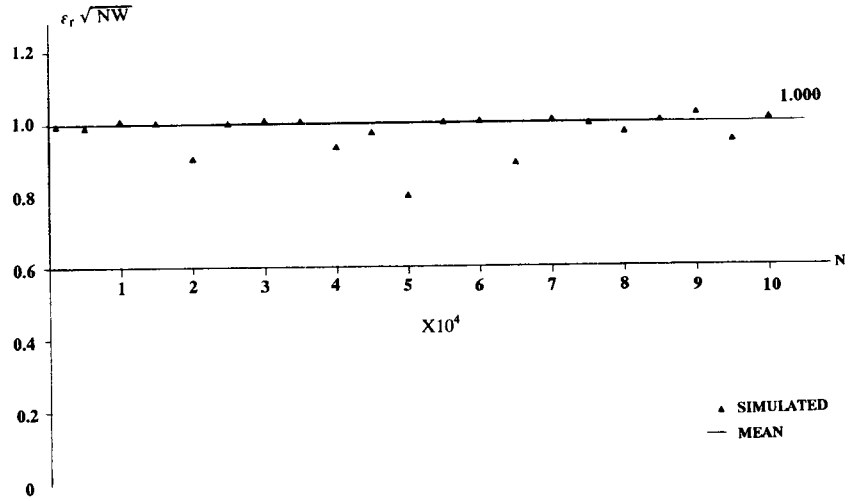

Fig. 8- Relative error proportionality constant versus sample size, for uniform probability density estimates. Number of slots = 50

From the plots of ε_r and their comparisons with the theoretical curves, it is seen that the empirical results, obtained by simulation, show a good agreement with the theoretical curves. For the plots of $\varepsilon_r \sqrt{NW}$, in each case, the mean value of the points was also calculated and drawn as a straight line parallel to the abscissa. This would be an average value for the proportionality constant c. It is seen that in every case the mean value, obtained for c, is very close to unity. It is also seen that the relatively small scatters of ε_r about the theoretical curves are magnified in the $\varepsilon_r \sqrt{NW}$ plots; this is due to multiplication by $\sqrt{N}$ which is relatively large.

Next, independent random variables from a standard Gaussian

distribution were simulated, with a sample size of 100000 and different slot sizes. In each case, the following quantity:

$$c^2 = \frac{[p(n) - p(n)^2]}{p(n)} NW \qquad (23)$$

Which is obtainable from equation (18), was calculated. As before, p(n) in the estimate of probability desnsity in the n^{th} slot and p(n) is the corresponding true value. The quandtity c is expexted to have an average value of unity for independent random variables.

The mean square error constant of proportionality , c^2, was plotted against the slot number, n, and its average value was also calculated and drawn as a straight line parallel to the abscissa. This was repeated for different window sizes given by $W = 6/M_s$, as found by equation (14), where M_s is the number of slots. The simulation results for M_s equal to 20, 30, 40 and 50 are shown in Figures 9, 10, 11 and 12 respectively. It can be seen from these figures that the average values of c^2 are around unity. The scatter of points about the mean have been magnified due to multiplication by a large value N.

The average values of c^2 were calculated for M_s equal to 20 to 100 in steps of 10. A plot of these against M_s is shown in Figure 13. The mean of these values, was also calculated which gives $c \cong 0.98$.

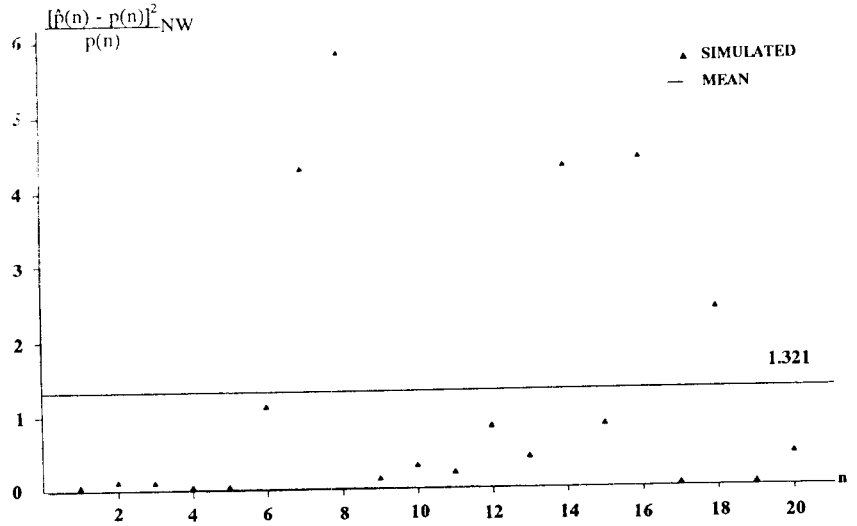

Fig. 9- Mean square error proportionality constant versus slot number, for standard normal density estimates. Number of slots = 20 Sample size = 100 000

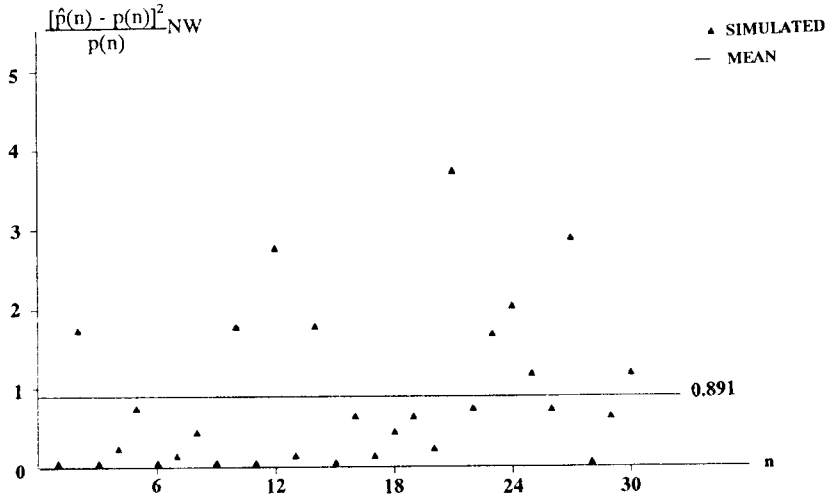

Fig. 10- Mean square error proportionality constant versus slot number, for standard normal density estimates. Number of slots=30 Sample size=100 000

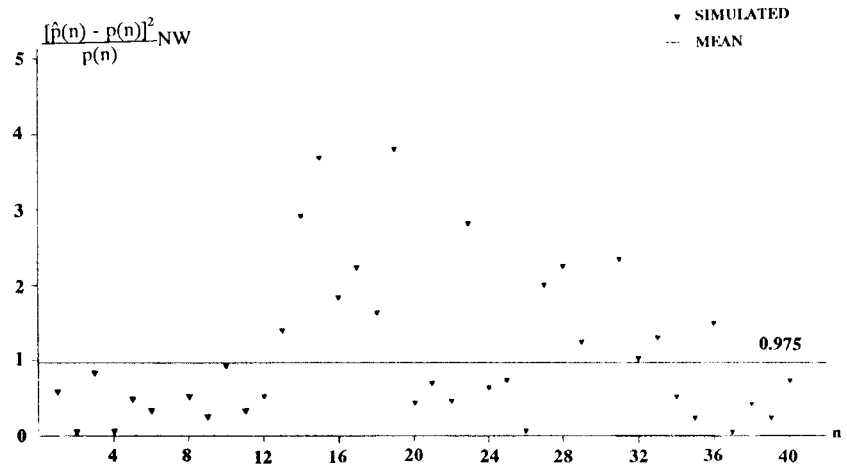

Fig. 11- Mean square error proportionality constant versus slot number, for standard normal density estimates. Number of slots=40 Sample size=100 000

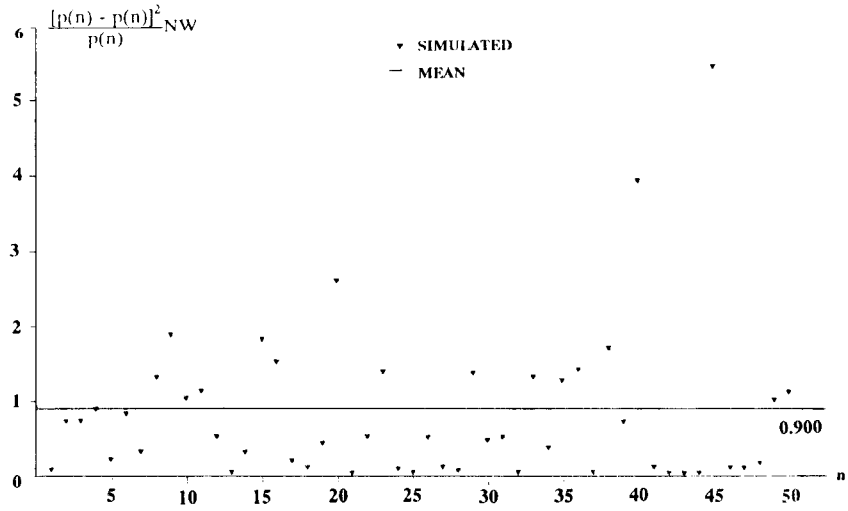

Fig. 12- Mean square error proportionality constant versus slot number, for standard normal density estimates. Number of slots=50 Sample size=100 000

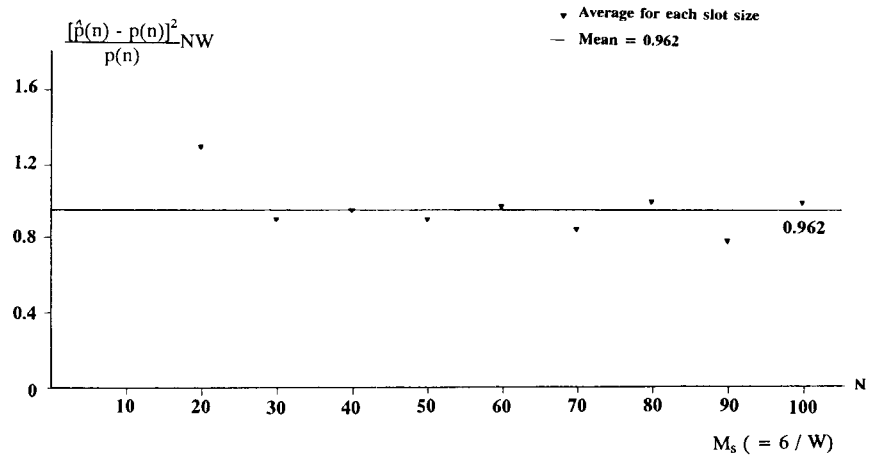

Fig. 13- Average of mean square error proportionality constant versus the number of slots, for standard normal density estimates. Sample size=100 000

Moreover, a different sample size of 50000 was also simulated with $M_s = 50$. The plot of c^2 versus the slot number is shown in Figure 14. Similar observations, as before, can be made from this figure and the mean value is also seen to be very close to unity.

Finally, in order to demonstrate that c depends on the autocorrelation function of the data and the sampling period, Gaussian processes with zero mean, unit variance and known autocorrelation functions were also considered. The time series were simulated with different sampling periods, a relatively large sample size (100000) and, 50 slots were used for the probability density function estimates. In each case the quantity c^2 was

plotted against the slot number and its mean value was computed from contributions of different slots.

The random process with autocorrelation function:

$$R(\tau) = \exp(-\tau) \qquad (24)$$

Was simulated with time intervals of 0.05 and 0.1. The results for $\Delta\tau = 0.05$ are shown in Figure 15; the average value of c^2 gives $c \cong 2.95$. The results corresponding to $\Delta\tau = 0.1$ are also displayed in Figure 16; it gives $c \cong 2.18$.

The random process with the autocorrelation function:

$$R(\tau) = \exp(-\tau) \cos = \pi\tau \qquad (25)$$

Was also simulated, using the same time intervals, as before, i.e. 0.05 and 0.1. Figure 17 shows the results for $\Delta\tau = 0.05$; the average value of c^2 gives $c \cong 1.94$. Figure 18 also shows the results for $\Delta\tau = 0.1$; giving $c \cong 0.97$.

The simulations of the correlated data would, therefore, indicate that the constant 0 depends on the associated autocorrelation function and sampling period.

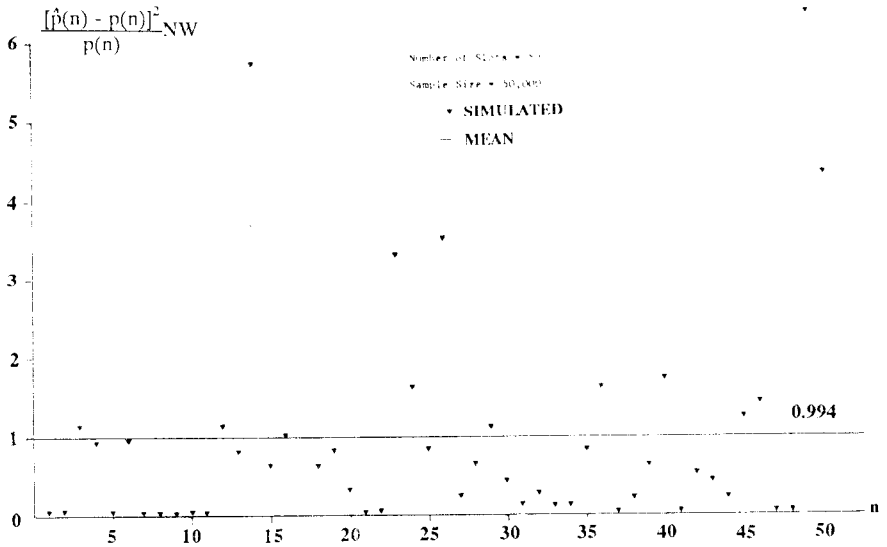

Fig. 14- Mean square error proportionality constant versus slot number, for standard normal density estimates. Number of slots=50 Sample size=50 000

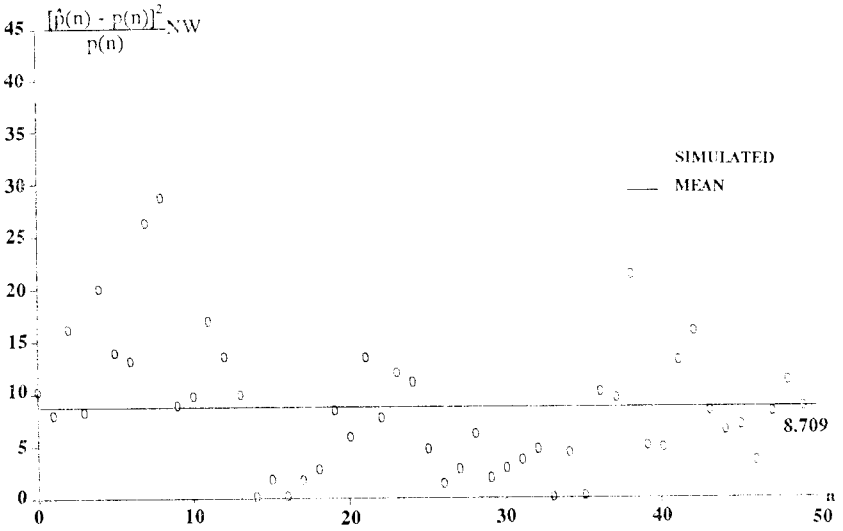

Fig. 15- Mean square error proportionality constant versus slot number, for random data from standard normal density and the autocorrelation function:

$R(\tau) = \exp(-\tau)$, $\Delta\tau = 0.05_s$ Number of slots=50 Sample size=100 000

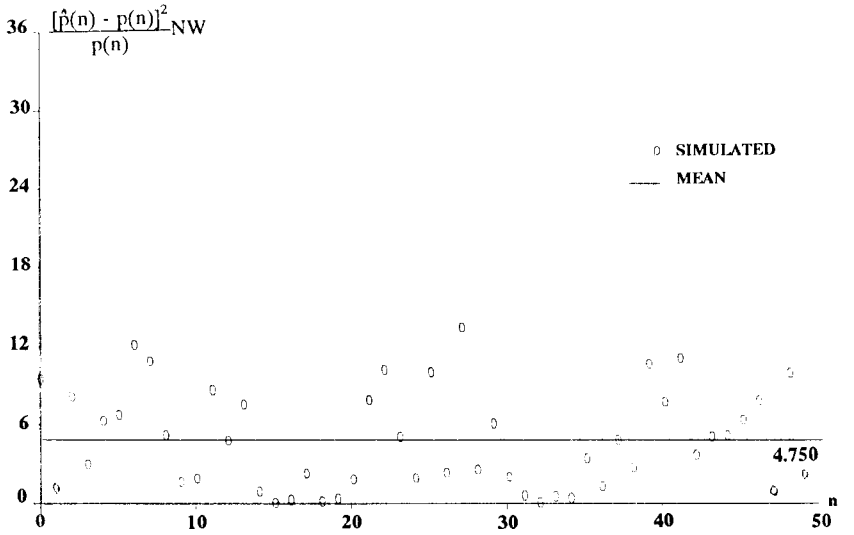

Fig. 16- Mean square error proportionality constant versus slot number, for random data from standard normal density and the autocorrelation function:

$R(\tau) = \exp(-\tau)$, $\Delta\tau = 0.1s$ Number of slots=50 Sample size=100 000

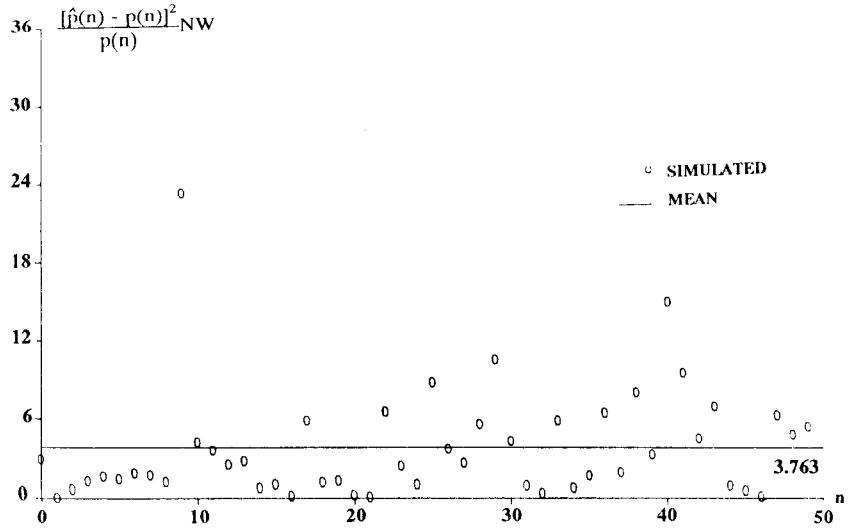

Fig. 17- Mean square error proportionality constant versus slot number, for random data from standard normal density and the autocorrelation function:

$R(\tau) = \exp(-\tau)\cos(\pi\tau)$, $\Delta t = 0.05_s$ Number of slots= 50 Sample size=100 000

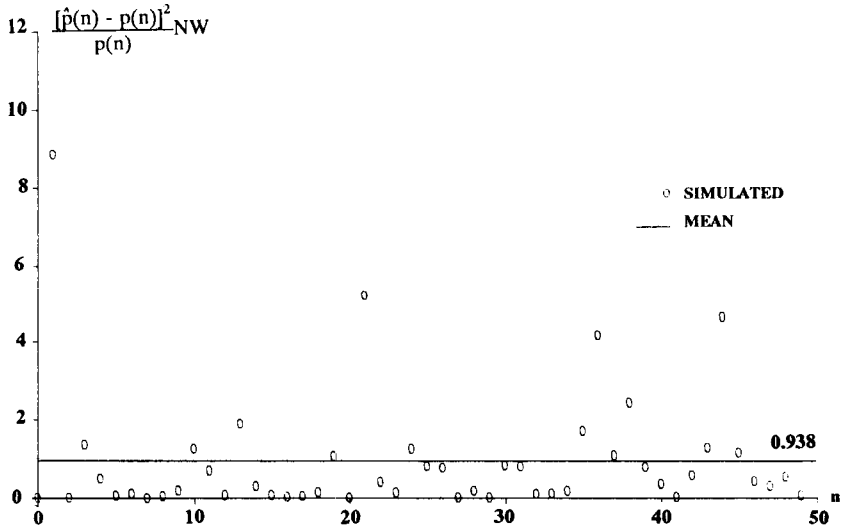

Fig. 18- Mean square error proportionality constant versus slot number, for random data from standard normal density and the autocorrelation function:

$R(\tau) = \exp(-\tau)\cos(\pi t)$, $\Delta t = 0.1_S$ Number of slots= 50 Sample size=100 000

Discussion ane Concluding Remarks

The statistical errors in the digital estimation of probability density functions were considered. The estimato error was :be composed of a random portion and a bias term. The bias term would depend on the window size and the second derivative of the probability density function. The error analysis was applied to the examples of the uniform and standard Gaussioan density functions. The bias term was seen to be zero for the former and negligible for the latter; the relative error could, hence, be represented by the random portion only. That is,

the root mean square error would reduce to the standard error.

The relative error was found to be inversely proportional to the square root of the product of sample size, window size and the probability density function. The constanl of proportionality c. was theoretically expected to be unity tor independent random variables and dependent on the autocorrelation and sampling rate. otherwise.

The inverse proportionality between the error and the window size would oppose the high resolution requirements. In addition. the smaller the window size, the less is the bias error. However, when a high resolution (small window size) is used, the effect on the error may be compensated by increasing the sample size.

The estimate error was further investigated by simulations on a digital computer. Independent random variables, from a uniform distribution, were simulated; this showed a very good agreement between the errors obtained empirically and those expected theoretically. The constant c was also seen to be very close to unity. Independent random variables from a standard Gaussian distribution were also simulated and , again, the overall average value obtained for c was close to unity.

Correlated data, with known autocorrelation functions were also considered; two kinds of Gaussian processes were simulated, with different sampling periods. The indication was that, for correlated data, the constant c would depend on the associated autocorrelation function and sampling period. Consequently, an

empirical value could not be suggested for c, due to the fact that it would depend on the autocorreration function, and for a given autocorrelation, would vary with the sampling period.

APPENDIX A

The Digital Procedure For Probability Density Estimation

In the context of time series analysis, the probability density function of random data describes the probability that the data will assume a value within some defined range at any instant of time. Considering a time series x(t), the probability the x(t) assumes a value within the range x and (x + Δx) may be obtained by taking the ratio T_x / T, where T_x is the total amount of time that x(t) falls inside the range (x, x + Δx) during an observation time T. This ratio will approach an exact probability density as T approaches infinity. The probability density function p (X) can be defined as:

$$p(x)= \lim_{\Delta x \to o} \frac{prob[x \le x(t) \le x + \Delta x]}{\Delta x} = \lim_{\Delta x \to o} [\lim_{T \to \infty} \frac{T_x}{T} \qquad (A.1)$$

The qunatity p(x) is always a real - valued, non - negative function.

However, for the digital estimation of the probability density function, equation (A.l) may be used to obtain an estimate p(x)as:

$$p(x)= \frac{N_x}{NW} \qquad (A.2)$$

Where W is a narrow interval centred at x and Nx is the number of data values (out of N values) which fall within the range $(x-\frac{W}{2}x,$

$X+\frac{W}{2}$). Hence, an estimate p(x) may be evaluated, digitally, by dividing the full range of x into an appropriate number of equal width slots and observing the number of data values falling in each interval. Dividing the observed number by the product of W and the total sample size, N, yields the estimate p(x). Further account of this is given in [2].

References

1- Silverman, B. W. (1990), **"Density Estimation for Statistics and Data Analysis"**, Monographs on Statistics and Applied Probability 26, Chapman and Hall.

2- Bendat, J. S. and Piersol, A. G. (1986), **"Random Data: Analysis and Measurement Procedures"**, Wiley-Interscience.

3- Wertz, W. and Schneider, B. (1979), **"Statistical; Density Estimation: a Bibliography"**, Int. Stat Rev., 47, 155-175.

4- Papoulis A. (1984), **"Probability, Random Variables and Stochastic Processes"**, Mc Graw-Hill.

5- Taha, H. A. (1988), **"Simulation Modelling and Simnet"**, Prentice Hall.

6- Box, G. E. P. and Muller, M. E. (1958), **"A Note on the Generation of Normal Deviates"**, Ann. Math. Stat. 28, pp. 610-611.

7- Burrus, C. S. et al (1994), **"Computer Based Exercises for Signal Processing Using MATLAB"**, Prentice Hall.