PDF Estimation via Characteristic Function and an Orthonormal Basis Set

ROY M. HOWARD
School of Electrical Engineering and Computing
Curtin University of Technology
GPO Box U1987, Perth 6845.
AUSTRALIA
r.howard@exchange.curtin.edu.au

Abstract: - The efficacy of estimating the probability density function through, first, reconstruction of the characteristic function and, second, through use of a series approximation based on a Hermite orthonormal basis set, is shown. Both approaches yield a lower integrated error than a benchmark uniform kernel density function for the case of a probability density function with a significant tail. For a Hermite series it is shown that a scaling parameter can be set according to the root mean square value of the data to obtain optimum error performance. The Hermite series estimator is applied to approximating the probability density function evolution of a generalized shot noise process and to estimation of a bimodal probability density function.

Key-Words: - Probability density function, PDF, estimation, kernel density estimate, characteristic function, Hermite, orthonormal, shot noise.

1 Introduction

Randomness is inherent, and its effects important, in many technological, biological and non-biological systems. Randomness is characterized in many ways including the use of spectral analysis and for parameters defining random variables, the probability density function. Unfortunately, in many instances it is not possible to derive an analytical expression for such a function although new results continue to be published, e.g. [1]. Due to its importance, the estimation of the probability density function of a random variable has a long history, e.g. [2], [3].

There are many approaches for estimating the probability density function including the model based approach proposed by Kay, [4], (only suitable for symmetric probability density functions), kernel density approaches (a non-parametric approach), e.g. [3], and [5] where windowing of the characteristic function was used to define new kernels. The first contribution of this paper is an approach for estimating the probability density function based on reconstruction of the characteristic function.

Orthogonal series can also be used to facilitate probability density function estimation, either directly, e.g. [6], or through a Gram-Charlier series, e.g. [7], [8]. The problem with the latter approach is that it requires the estimation of high order moments. In general, an orthogonal series approach has not received much attention in the literature. A second contribution of this paper is demonstrate the usefulness of a orthogonal series, based on weighted Hermite polynomials, for approximating a probability density function defined on the infinite interval. A third contribution is to provide results for determining the optimum scaling factor to use with such a series. The series is applied to approximating a bimodal probability density function and the probability density function evolution of a generalized shot noise process.

2 Theory

It is assumed that N independent samples, $x_1, \ldots, x_N$, from a random variable X, are available. The probability density function of X, and an estimator for this function, are denoted, respectively, $f_X(x)$ and $\hat{f}_X(x)$. The mean and variance of X are denoted, respectively, μ_X and σ_X^2. $\boldsymbol{E}$ is the expectation operator.

2.1 Kernel Density Estimation

A kernel density estimator for the probability density function, e.g. [3], is

$$\hat{f}_X(x) = \frac{1}{Nh} \cdot \sum_{i=1}^{N} K\left(\frac{x - x_i}{h}\right) \tag{1}$$

where K is the weighing function and h is a scaling parameter which needs to be optimally chosen. The simplest weighting function is the rectangular function:

$$K(x) = \begin{cases} 1/2 & -1 \le x < 1 \\ 0 & \text{elsewhere} \end{cases} \tag{2}$$

and this defines the uniform kernel estimator.

Xie has shown (see table III in [5]) for the case considered below, where the probability density function has a significant tail, that this estimator provides overall performance which is similar to, or better than, more sophisticated estimators. Accordingly, this estimator is suitable to be used as reference estimator.

2.2 Characteristic Function Estimator

By definition the characteristic function of a random variable X is

$$\Phi_X(f) = \boldsymbol{E}[e^{j2\pi fX}] = \int_{-\infty}^{\infty} e^{j2\pi fx} f_X(x)dx \qquad (3)$$

where, for convenience, $\boldsymbol{E}[e^{j2\pi fX}]$ is used rather than $\boldsymbol{E}[e^{jwX}]$. An estimator for the characteristic function is

$$\hat{\Phi}_X(f) = \frac{1}{N}\sum_{i=1}^{N} e^{j2\pi fx_i} \qquad (4)$$

As the characteristic function is the inverse Fourier transform of the probability density function an estimate of $f_X(x)$ can be obtained from $\hat{\Phi}_X(f)$ by taking its Fourier transform. However, the Fourier transform of (4) does not exist (or, using distribution theory, consists of impulses - see equation 3 in [5]) and windowing is required according to

$$\hat{f}_X(x) = \int_{-\infty}^{\infty} \hat{\Phi}_X(f)W(f)e^{-j2\pi fx}df \qquad (5)$$

where W is the windowing function. This leads to the kernel estimate of the probability density function according to

$$\hat{f}_X(x) = \frac{1}{N}\sum_{i=1}^{N} w(x-x_i) \qquad (6)$$

where w is the Fourier transform of W. In [5] results from windowing Fourier transform theory have been used to propose new kernels.

An alternative approach is to use $2N_f+1$ samples of the characteristic function, with a spacing of Δf, and a reconstruction function, to first estimate the characteristic function according to

$$\hat{\Phi}_X(f) = \sum_{k=-N_f}^{N_f} \hat{\Phi}_X(k\Delta f)\mathrm{sinc}\left(\frac{f-k\Delta f}{\Delta f}\right) \qquad (7)$$

over the range $-N_f\Delta f \le f \le N_f\Delta f$ and where $\mathrm{sinc}(x) = \sin(\pi x)/\pi x$. It then follows that an estimator for the probability density function, valid over the range $-1/2\Delta f < x < 1/2\Delta f$, is

$$\hat{f}_X(x) = \Delta f \sum_{k=-N_f}^{N_f} \hat{\Phi}_X(k\Delta f)\exp[-j2\pi k\Delta fx] \qquad (8)$$

2.3 Orthonormal Series Approach

Consider an orthonormal basis set $\{b_1, \ldots, b_i, \ldots\}$ defined on the interval where the probability density function is non-zero. With such a basis set, the probability density function of a random variable X can be written as [6]

$$f_X(x) = c_1b_1(x) + c_2b_2(x) + \ldots \qquad (9)$$

where the ith coefficient is

$$c_i = \boldsymbol{E}[b_i(X)] = \int_{-\infty}^{\infty} b_i(x)f_X(x)dx \qquad (10)$$

This result follows as

$$\begin{aligned}\boldsymbol{E}[b_i(X)] &= \int_{-\infty}^{\infty} b_i(x)f_X(x)dx \\ &= \int_{-\infty}^{\infty} b_i(x)[c_1b_1(x) + c_2b_2(x) + \ldots]dx \qquad (11)\\ &= c_i\int_{-\infty}^{\infty} b_i^2(x)dx = c_i\end{aligned}$$

The coefficient c_i can be estimated according to

$$\hat{c}_i = \frac{1}{N}\sum_{i=1}^{N} b_i(x_i) \qquad (12)$$

and, consequently, an estimator for $f_X(x)$ is

$$\hat{f}_X(x) = \hat{c}_1b_1(x) + \hat{c}_2b_2(x) + \ldots \qquad (13)$$

It is the case, [6], that the estimate of the ith coefficient, as given by (12), is unbiased and has a variance which decreases according to $1/N$.

2.4 Hermite Basis Set

For the case of a probability density function, which is defined on $(-\infty, \infty)$, a Hermite orthonormal basis set is useful. Such a basis set is defined according to:

ISSN: 1792-4235

ISBN: 978-960-474-199-1

$$\left\{ b_i^N(x) = \frac{H_i(x) \cdot \exp[-x^2/2]}{2^{i/2}\sqrt{i!}\pi^{1/4}} : \ i \in \{0, 1, \ldots\} \right\} \quad (14)$$

where

$$H_0(x) = 1 \qquad H_1(x) = 2x \qquad (15)$$

and the recurrence relationship is:

$$H_i(x) = 2xH_{i-1}(x) - 2(i-1)H_{i-2}(x) \qquad (16)$$

With an expanding scale factor k_S, the orthonormal basis set becomes

$$\left\{ b_i(x) = \frac{1}{\sqrt{k_S}} b_i^N\left(\frac{x}{k_S}\right) : \ i \in \{0, 1, \ldots\} \right\} \qquad (17)$$

The first seven basis functions are graphed in Fig. 1.

2.5 Measures of Performance

Two useful error measures for a probability density function estimator, over the range $-M\Delta x < x < M\Delta x$, are the integrated error

$$\varepsilon_1 = \Delta x \sum_{i=-M}^{M} \left| f_X(i\Delta x) - \hat{f}_X(i\Delta x) \right| \qquad (18)$$

and the root integrated square error

$$\varepsilon_2 = \sqrt{\Delta x \sum_{i=-M}^{M} [f_X(i\Delta x) - \hat{f}_X(i\Delta x)]^2} \qquad (19)$$

The former is preferable when approximation of the tail of a probability density function is important as the latter de-emphasizes errors when the probability function takes on small values.

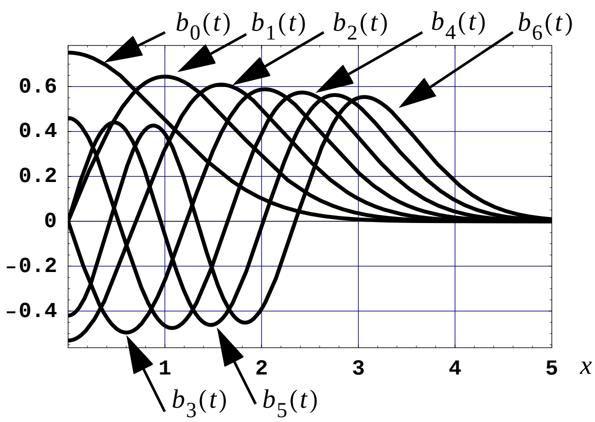

Fig. 1 Zeroth to sixth order basis functions defined by Hermite polynomials for the case of $k_S = 1$.

3 Results

In [4] and [5] a random variable X with a probability density function

$$f_X(x) = \sum_{k=1}^{2} p_k \cdot \frac{e^{-(x-\text{mean}_k)^2/2\text{var}_k}}{\sqrt{2\pi\text{var}_k}} \qquad (20)$$

was used as a basis for comparison of different probability density function estimators. The parameters used were: $p_1 = 1 - \varepsilon$, $p_2 = \varepsilon$, $\varepsilon = 0.1$, $\text{mean}_1 = 0$, $\text{mean}_2 = 0$, $\text{var}_1 = 1$ and $\text{var}_2 = 100$. In the following the value for mean_1 is set to 2 to generate an asymmetrical probability density function. For this case $\mu_X = 1.8$, $\sigma_X^2 = 11.26$ and $\sigma_X = 3.23$. This probability density function is consistent with a random sum of independent random variables:

$$X = \sum_{i=1}^{M} X_i \qquad (21)$$

where, for example, $M \in \{1, 2\}$, $P[M = 1] = 1 - \varepsilon$, $P[M = 2] = \varepsilon$, $\mu_{X_1} = 2$, $\mu_{X_2} = -2$, $\sigma_{X_1}^2 = 1$ and $\sigma_{X_2}^2 = 99$.

The following results are based on 10000 independent samples of the random variable X. To obtain accurate values for ε_1 and ε_2, M is set to 500 and Δx is chosen such that all significant error values are included.

The true probability density function, and the histogram estimate, are shown in Fig. 2. The uniform kernel density estimator, with an optimum parameter $h = 0.4$, is shown in Fig. 3. The characteristic function estimator, as specified by (8), is shown in Fig. 4. The estimator obtained by using a 100 term series expansion based on a Hermite basis set, as given by (13), is shown in Fig. 5. An optimum value for the scaling parameter of $k_S = 4.5$ has been used.

3.1 Comparison

In Table 1 the integrated errors for the three probability density function estimators shown in Fig. 3, Fig. 4 and Fig. 5 are tabulated.

To compare the estimators for the case of a probability density function with no tail, a Gaussian probability density function with zero mean and unit variance was used and the results in Table 2 were obtained using optimum parameters of $h = 3$, $N_f = 50$, $\Delta f = 0.01$ and $k_S = 3.5$.

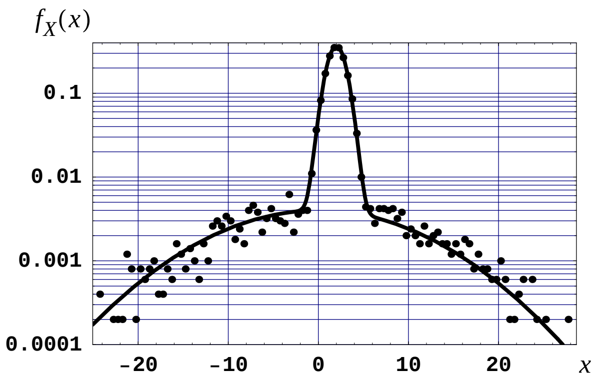

Fig. 2 True probability density function and an estimate arising from a normalized histogram with a bin width of 0.5.

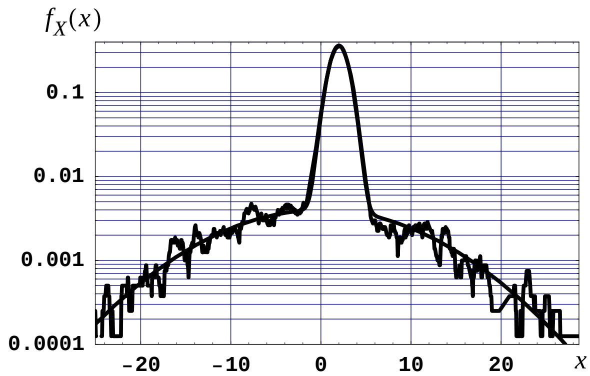

Fig. 3 True and approximate probability density function based on a uniform kernel density with a parameter $h = 0.4$.

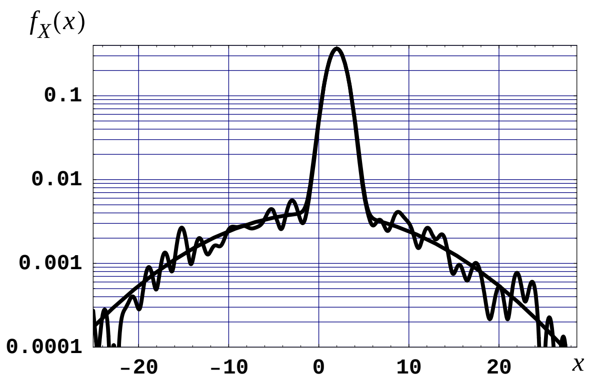

Fig. 4 True and approximate probability density function obtained via a characteristic function approach. 100 samples of the characteristic function were used, over the range $[-2/\sigma_X, 2/\sigma_X] \approx [-0.62, 0.62]$ with a frequency resolution of 0.0124.

These results illustrate the potential of lower integrated errors with use of either the proposed characteristic function, or Hermite based orthonormal series, estimator.

Table 1 Integrated errors: PDF of (20).

PDF Estimator	ε_1	ε_2
Uniform Kernel	0.0485	0.0168
Characteristic Function	0.0305	0.00825
Hermite Basis Set	0.0330	0.00891

Table 2 Integrated errors: Gaussian case.

PDF Estimator	ε_1	ε_2
Uniform Kernel	0.0226	0.0124
Characteristic Function	0.0119	0.0054
Hermite Basis Set	0.0161	0.0072

3.2 Discussion: Characteristic Function

The characteristic function estimator is computationally efficient and the relative error measures are relatively insensitive to the number of samples taken above a threshold number. Naturally, it is important to sample the characteristic function over an appropriate range and with an appropriate resolution. The range depends on the nature of the probability density function and optimum ranges for the Gaussian case are close to $(-1/2\sigma_X, 1/2\sigma_X)$ and $(-2/\sigma_X, 2/\sigma_X)$ for the probability density function specified by (20).

The probability density estimator shown in Fig. 4 is based on the approximation to the real and imaginary parts of the characteristic function as shown in Fig. 6. For a probability density function, which is an even function, the imaginary part of the characteristic function is zero.

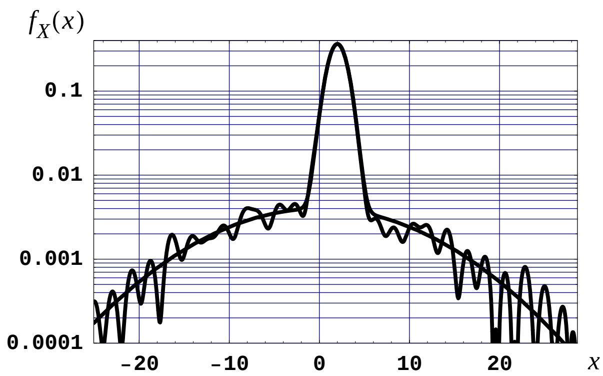

Fig. 5 True and approximate probability density function obtained from a 100 term Hermite basis function expansion with a scaling factor of $k_S = 4.5$.

3.3 Discussion: Hermite Based Estimator

3.3.1 Optimum Scaling Factor

In Table 3 the integrated errors ε_1 and ε_2 are tabulated, for the case of 100 terms, as the scaling factor k_S is varied. Note the relative insensitivity of the integrated errors to variations in k_S with an optimum level around $k_S = 4.5$ which is $1.34\sigma_X$.

Table 3 Integrated errors as k_S is varied.

k_S	ε_1	ε_2
2	0.0447	0.0121
2.5	0.0395	0.0108
3	0.0356	0.0101
3.5	0.0333	0.00953
4	0.0334	0.00928
4.5	0.0330	0.00891
5	0.0367	0.00931
5.5	0.0488	0.0114

3.3.2 Number of Terms in Expansion

The performance of the Hermite series estimator improves, in general, as the number of terms increases and exhibits a slow deterioration in integrated error for very large number of terms. For example, with $k_S = 4.5$ the integrated error improves to 0.032 with 150 terms and then decreases to 0.035 for 200 terms. Such good behaviour is due, in part, to the nature of the Hermite basis functions, which, as illustrated in Fig. 1, are bounded, decrease rapidly to zero outside a given interval and lead to coefficients in the decomposition (13) which are bounded [6].

Good estimation can be achieved with a much lower number of terms, but, in general, the performance is more sensitive to the scaling parameter k_S.

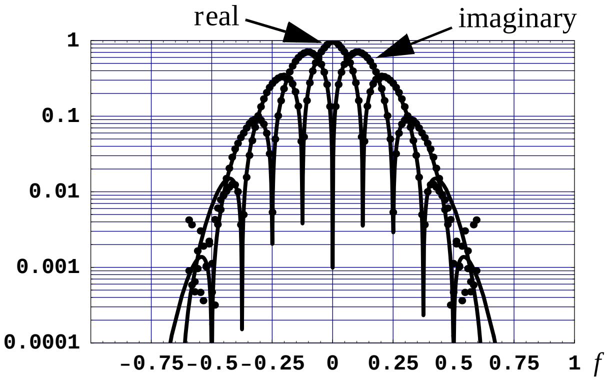

Fig. 6 Approximations (dots) to the real and imaginary parts of the characteristic function which underpin the approximation shown in Fig. 4.

For example with 20 terms the integrated error ε_1 equals 0.04 for $k_S = 2$ but 0.12 when $k_S = 3$.

Consistent with the data in Table 3, and as a guide, the use of 100 terms, and with k_S set between σ_X and $1.5\sigma_X$, leads to low integrated errors. For the case of a random variable whose probability density function does not have a significant tail, e.g. a Gaussian random variable, simulation results suggests that a higher value for k_S is warranted with a value between $2\sigma_X$ to $4\sigma_X$ being optimum.

Naturally, when the mean varies significantly from zero it is appropriate to use basis waveforms centered at the mean of the random variable.

4 Applications

4.1 Approximation of a Bimodal PDF

Consider the modelling of a bimodal probability density function, as defined by (21), where independent Gaussian random variables are assumed and $M \in \{1, 2\}$, $P[M = 1] = P[M = 2] = 0.5$, $\mu_{X_1} = -4$, $\mu_{X_2} = 8$, $\sigma^2_{X_1} = 1$ and $\sigma^2_{X_2} = 1$. The equal probable outcomes are: X_1 with a mean of -4 and a variance of 1; $X_1 + X_2$ with a mean of 4 and a variance of 2. Overall $\mu_X = 0$ and $\sigma^2_X = 17.5$. The true, and approximate, probability density functions are shown in Fig. 7. $\hat{f}_X(x)$ is based on a 100 term Hermite basis series with a scaling parameter given by $k_S = \sigma_X$. The integrated error is $\varepsilon_1 = 0.026$. For reference, $\varepsilon_1 = 0.034$ for the uniform kernel with $h = 0.35$ (optimal) and $\varepsilon_1 = 0.038$ for the characteristic function estimator with sampling over $[-2.5/\sigma_X, 2.5/\sigma_X]$ (optimal).

4.2 Generalized Shot Noise

The evolution of the probability density function of a random process conveys important information but, in general, finding an analytic expression is difficult. To demonstrate the usefulness of the Hermite basis estimator consider the generic, and ubiquitious, random process of generalized shot noise defined according to

$$X(t) = \sum_{k=1}^{\infty} A_k h(t - \Gamma_k) \qquad (22)$$

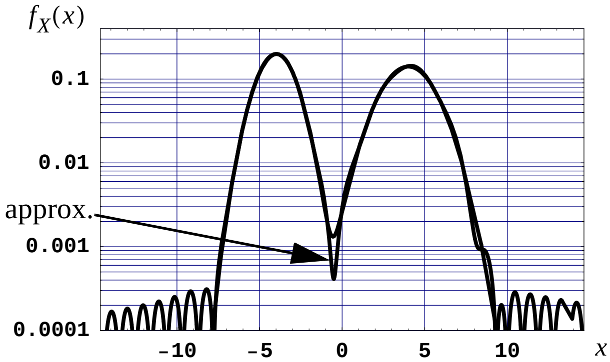

Fig. 7 True and approximate probability density function. The approximate probability density function is based on a 100 term Hermite based series with a scaling parameter $k_S = \sigma_X$.

where h is a pulse function which is assumed to be causal, continuous and with $h(0) = 0$. $A_1, \ldots, A_i, \ldots$ are independent random variables with respective probability density function, $f_{A_1}, \ldots, f_{A_i}, \ldots$. The random variables $\Gamma_1, \ldots, \Gamma_i, \ldots$ are assumed to be independent with a uniform distribution on all finite intervals and collectively define a set of Poisson points with a rate of λ per second.

The probability density function, as approximated by a Hermite basis set according to (13), with the scaling parameter k_S adaptively set according to $\sigma_{X(t)}$, is shown in Fig. 8 for the case of $\lambda = 2$, $h(t) = t\exp[-t]$ and $A_1, \ldots, A_i, \ldots$ being zero mean, unit variance, Gaussian random variables. The probability of a zero outcome is shown in Fig. 9.

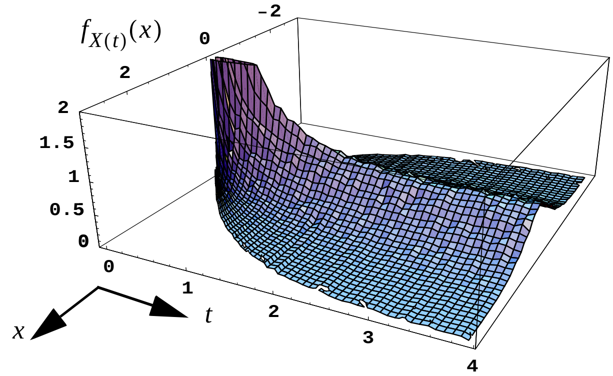

Fig. 8 Probability density function of a generalized shot noise process as approximated by a 100 term Hermite basis set with an adaptive scaling factor.

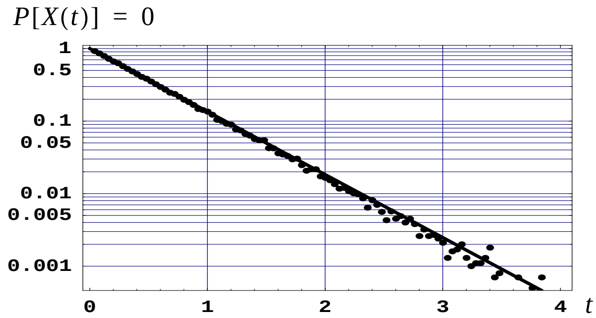

Fig. 9 Probability of a zero outcome case for the generalized shot noise process. The true probability is $\exp[-\lambda t]$.

5 Conclusion

In this paper, estimation of the probability density function, via reconstruction of the characteristic function, was proposed and the usefulness of a Hermite based orthonormal series estimator was shown. Presented results justified the two approaches with a lower integrated error than a benchmark uniform kernel density estimator for the case of a probability density function with a significant tail. It was shown that the scaling parameter required for the Hermite basis set estimator should be set close to the root mean square value of the data and the convergence of the series generally improves with the number of terms in the series. The Hermite series was applied to estimating the probability density function evolution associated with a generalized shot noise process and to estimation of a bimodal probability density function.

References:

[1] H. V. Khuong & H. Y. Kong, 'General expression for pdf of a sum of independent exponential random variables', *IEEE Communication Letters*, vol. 10, 2006, pp. 159-161.

[2] M. Rosenblatt, 'Remarks on some nonparametric estimates of a density function', *The Annals of Mathematical Statistics*, vol. 27, 1956, pp. 832-837.

[3] E. Parzen, 'On the estimation of a probability density function and mode', *The Annals of Mathematical Statistics*, vol. 33, 1962, pp. 1065–1076.

[4] S. Kay, 'Model-based probability density function estimation', *IEEE Signal Processing Letters*, vol. 5, 1998, pp. 318-320.

[5] J. Xie & Z. Wang, 'Probability density function estimation based on windowed Fourier transform of characteristic function', *Proc. of 2nd International Congress on Image and Signal Processing*, Oct. 17-19, 2009, Tianjin, China, 4 pages.

[6] S. C. Schwartz, 'Estimation of probability density by an orthogonal series', *The Annals of Mathematical Statistics*, vol. 38, 1967, pp. 1261-1265.

[7] J. R. Thompson & R. A. Tapia, '*Nonparametric Function Estimation, Modelling, and Simulation*', SIAM, 1990, p. 37.

[8] L. N. Bowers, 'Expansion of probability density functions as a sum of gamma densities with applications in risk theory', *Transactions of Society of Actuaries*, vol. 18, 1966, pp. 125-147.