Generalized Probability Density Function Estimation via Convex Optimization

Arian Eamaz*, Farhang Yeganegi, Mojtaba Soltanalian, and Natasha Devroye
University of Illinois Chicago, Chicago, IL, USA

2022 IEEE International Symposium on Information Theory (ISIT) | 978-1-6654-2159-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/ISIT50566.2022.9834583

***Abstract*—A longstanding problem in statistics pertains to the estimation of probability density functions of continuous random variables from a finite set of their samples. In this paper, we propose a new parametric probability density function estimator based on convex programming. Our formulation decomposes the unknown distribution as a Gaussian penalty function plus an error function, which is then expanded by multi-scale wavelet functions (specifically frames) such as B-Spline and Mexican Hat wavelets. To recover the wavelet coefficients in the error function, a convex quadratic program is formulated which takes into account the positivity of the probability density function-through a linear constraint. The proposed decomposition model is shown to facilitate an accurate estimation of the probability density functions of interest.**

I. Introduction

PROBABILITY density function (PDF) estimation has a long history [1]–[8]. Distribution estimation with a finite sample set plays a central role in theoretical and applied statistics, information theory and communication application areas such as data compression, information capacity [9], source coding, time-series prediction [10], mutual information estimation [11], and statistical image processing [12], [13].

Conventional PDF models such as the Gaussian mixture model and K-distribution usually have a limited number of parameters that are calculated based on the first few moments, and thus provide a poor fit for some distributions, including heavy-tailed processes [12]. To address this shortcoming, we propose a new model to estimate the PDF of sample data that also takes advantage of higher moments and the characteristic function, which incorporates more statistical information.

PDFs may be categorized as either (i) *sub-Gaussian* or (ii) *super-Gaussian* [14]. Roughly speaking, the tails of a sub-Gaussian distribution are dominated by the tails of its Gaussian counterpart (signifying a decay *at least as fast as*) [15], [16]. On the other hand, a super-Gaussian distribution has a more spiky peak and a longer tail than a Gaussian distribution permits [17], [18].

We present a generalized model that covers both the sub- and super- Gaussian distributions (including the mixture models, heavy-tailed distributions, and sparse PDFs), while guaranteeing the positivity of the recovered PDF, which is in contrast to some earlier efforts [19]–[21]. The unknown PDF of the input sample data $f_X(x)$ should satisfy two properties: $\int_{-\infty}^{\infty} f_X(x)\,dx = 1$ and $f_X(x) \geq 0$. In our model, the PDF is decomposed into a Gaussian penalty function and an error function which is the difference between the unknown PDF and the penalty function. The penalty function is essentially the PDF of a zero-mean Gaussian distribution whose variance will be estimated using data. The error function, however, is formulated based on wavelet expansion. The choice of wavelet functions for this purpose is grounded in the fact that they are well-localized in both time and frequency, and hence, provide good *local* estimates of the error function [5], [20], [22]. Moreover, the deployment of wavelet functions provides our model with the additional advantage of being able to capture spikes of distributions with high magnitudes (e.g., sparse PDFs). In particular, we will use the *frame wavelets* which have closed-form formulas [23]. This in turn facilitates the formulation of the PDF estimation task as an optimization problem with a positivity constraint as a linear inequality constraint. Such a linear constraint provides an approach to ensure the positivity of the PDF in a much more straightforward manner than the previous efforts [21]. Last but not least, owing to the closed-form formulation of the frame wavelets, we can achieve a closed-form formula for our PDF model that will be key to its application in widely used estimation frameworks such as maximum likelihood (ML) estimation or maximum a posteriori probability (MAP) estimation. The resulting problem is a simple convex quadratic program – whose global minimum is immediately in reach.

Outline: In Section II, our PDF model is presented and the error function expansion using wavelet functions is discussed. Sections III and IV are dedicated to finding the coefficients of the proposed model from the moments and characteristic functions of the input sample data. To this end, a convex optimization problem is formulated which ensures the positivity of the PDF. To showcase the advantages of the proposed model, we utilize B-Spline wavelets and Mexican Hat wavelet functions as two examples of frame wavelets. Several numerical examples are presented to illustrate the effectiveness of the proposed model in Section V. Section VI concludes the paper.

Notation: We use bold lowercase letters for vectors, bold uppercase letters for block matrices, and uppercase letters for matrices. $(\cdot)^\top$ denotes the vector/matrix transpose. $[a_{ij}]^{N_1 \times N_2}$ is an $N_1 \times N_2$ matrix with a_{ij} as its ij-th element. $\mathbf{1}_n$ is a n-dimensional all-one vector. $\mathbb{E}\{.\}$ denotes the expected value operator. The r-th moment of a distribution $p(x)$ is defined as $\mu_p^r = \mathbb{E}\{x^r\}$. Also, a characteristic function (CF) of a random variable X is stated as

The first two authors have contributed equally to this work. The work of N. Devroye is partially supported by NSF under award 1815428. The contents of this article are solely the responsibility of the authors and do not necessarily represent the official views of the NSF.

* Corresponding author: A. Eamaz (e-mail: aeamaz2@uic.edu).

$P_X(\omega) = \mathbb{E}\left\{e^{i\omega x}\right\}$. Finally, the Gamma function is given by $\Gamma(x) = \int_0^\infty z^{x-1}e^{-z}dz$.

II. Problem Formulation

Suppose $\mathbf{x} = [x_i]^{N_\mathbf{x}\times 1}$ is a vector of zero-mean random samples generated from an unknown probability distribution f_X, which is to be estimated. Inspired by the PDF representation in [19], we propose to decompose the unknown PDF into a core penalty term $G(x)$ and an error function $e(x)$. Since we want the model to cover both sub-Gaussian and super-Gaussian distributions, the penalty function is assumed to be the PDF of the Gaussian distribution, while the error function $e(x)$ is tasked to represent the discrepancy between these distributions and the underlying Gaussian core PDF. The proposed decomposition takes the form,

$$f_X(x) \triangleq G(x) + e(x), \tag{1}$$

where $G(x) = \frac{1}{\sqrt{2\pi}\sigma_f}e^{-\frac{x^2}{2\sigma_f^2}}$. The variance σ_f^2 can be estimated from the observation $\mathbf{x}$ based on the maximum-likelihood estimation (MLE). Note that since every distribution consists of a tail and a center, the Gaussian penalty PDF with a variance estimated directly from the data can take the center, and the wavelet expanded error function can cover spikes and the tail of the desired distribution. Interestingly, a special case of (1) is the model previously proposed in [19] which considers $e(x)$ to be

$$e(x) \triangleq \sum_{n=1}^{N} \beta_n g_n\left(\frac{x}{\sigma_f}\right) G(x), \tag{2}$$

where $g_n(x)$ denotes the Hermite polynomial of order n and $\{\beta_n\}$ are given via the orthogonality principle. However, as shown in [12], it appears that the approach in [19] cannot guarantee the positivity of the PDF, which is critical for PDF formation. In contrast to [19], we will form the error function $e(x)$ using multi-scale wavelet functions, or more specifically, by a frame decomposition with non-orthogonal functions. Let $\psi(x)$ denote a *mother wavelet function* [23]. Then, a multi-scale representation of $\psi(x)$ is obtained as [22], [23]:

$$\psi_{jk}(x) = \psi\left(2^j x - k\right), \tag{3}$$

where j and k are the scaling and shifting factors, respectively. Therefore, our decomposition model in (1) can be rewritten as

$$f_X(x) = G(x) + \sum_{j=j_m}^{j_m+N_J-1} \sum_{K_l^j}^{K_u^j} c_{jk}\psi_{jk}(x), \tag{4}$$

where N_J denotes the number of scalings, and $\{j_m\}$, $\{K_l^j\}$, and $\{K_u^j\}$ maybe chosen based on the input data $\mathbf{x}$ as follows.

- Choosing $\{j_m\}$: Assuming that $2b_e$ denotes the effective bandwidth of ψ at $j = 0$, the effective bandwidth of ψ at a generic scale j will be $2^{-j+1}b_e$. Therefore, to find an appropriate scale to start with, one can set j_m to the minimum integer j such that $2^{-j+1}b_e \leq b$, where b is the dynamic range of a zero-mean version of the input data $\mathbf{x}$.
- Choosing $\{K_l^j\}$ and $\{K_u^j\}$: At a scale j, $\{K_l^j\}$ and $\{K_u^j\}$ are the minimum and maximum integer values of k such that the smallest number of shifting factors can be achieved considering the dynamic range (i.e., b) of the zero-mean version of the input data $\mathbf{x}$.

As a result, estimating the unknown PDF f_X boils down to estimating the coefficients $\{c_{jk}\}$. However, since the functions $\{\psi_{jk}(x)\}$ in (4) are non-orthogonal, the orthogonality principle cannot be employed. In the following sections, two distinct approaches are proposed to tackle this estimation problem. Namely, in Section III, a moment matching technique is used, whereas a characteristic function-based approach is considered in Section IV.

III. Moment Matching Technique

We begin our efforts by using a moment matching technique to obtain $\{c_{jk}\}$ in (4) while preserving the positivity of the estimated PDF $\hat{f}_X$; i.e. $\hat{f}_X \geq 0$. A considerable advantage of a moments-based technique is that, in some applications, the moments of the arbitrary distribution might be the only information available for PDF estimation. One such instance is the problem of estimating the probability of false alarm in detection theory [24].

A. PDF Estimation via Moment Matching

Multiplying (4) by x^r (with $r \in \mathbb{R}^+$) and integrating with respect to x yields:

$$\int_{-\infty}^{\infty} x^r f_X(x)\,dx = \int_{-\infty}^{\infty} x^r G(x)\,dx + \sum_{j=j_m}^{j_m+N_J-1} \sum_{k=K_l^j}^{K_u^j} c_{jk} \int_{-\infty}^{\infty} x^r \psi_{jk}(x)\,dx. \tag{5}$$

Let μ_G^r and $\mu_{f_X}^r$ be t he r-th moment of the Gaussian process associated with $G(x)$ and the r-th moment of the desired PDF $f_X(x)$, respectively. We can rewrite (5) as

$$\begin{aligned}\mu_{f_X}^r &= \mu_G^r + \sum_{j=j_m}^{j_m+N_J-1} \sum_{k=K_l^j}^{K_u^j} c_{jk} \int_{-\infty}^{\infty} x^r \psi_{jk}(x)\,dx, \\ &= \mu_G^r + \sum_{j=j_m}^{j_m+N_J-1} \sum_{k=K_l^j}^{K_u^j} c_{jk}\alpha_{jk}^r,\end{aligned} \tag{6}$$

where

$$\alpha_{jk}^r = \int_{-\infty}^{\infty} x^r \psi_{jk}(x)\,dx. \tag{7}$$

Since $r \in \mathbb{R}^+$, to avoid complex coefficients α_{jk}^r, one can use the absolute moment ($|x|^r$) instead of the ordinary moment (x^r). The moment μ_G^r in (6) is obtained as

$$\mu_G^r = \int_{-\infty}^{\infty} x^r G(x)\,dx = \frac{\sigma_f^r 2^{\frac{r}{2}}\Gamma\left(\frac{r+1}{2}\right)}{\sqrt{\pi}}\left(\frac{1+(-1)^r}{2}\right), \tag{8}$$

whereas for the absolute moment one obtains

$$\mu_G^r = \int_{-\infty}^{\infty} |x|^r G(x)\,dx = \frac{\sigma_f^r 2^{\frac{r}{2}}\Gamma\left(\frac{r+1}{2}\right)}{\sqrt{\pi}}. \tag{9}$$

Since f_X is unknown, the value of $\mu^r_{f_X}$ in (6) can be estimated as below [14]:

$$\hat{\mu}^r_{f_X} = \frac{1}{N_{\mathbf{x}}} \sum_{i=1}^{N_{\mathbf{x}}} x_i^r. \tag{10}$$

On the other hand, for the absolute moment counterpart, we have that

$$\hat{\mu}^r_{f_X} = \frac{1}{N_{\mathbf{x}}} \sum_{i=1}^{N_{\mathbf{x}}} |x|_i^r = \frac{\|x\|_r^r}{N_{\mathbf{x}}}. \tag{11}$$

Moreover, α^r_{jk} in (7) can be evaluated either by analytical or numerical integration approaches, depending on the choice of the mother wavelet function $\psi(x)$. Suppose $\mathcal{J}$ and $\mathcal{R}$ are the sets containing the scaling and moment numbers utilized in (6), respectively. Also, suppose $\mathcal{K}$ is the set containing the number of shifts in each scaling, with $k_i \in \mathcal{K}$ denoting the number of shifts in the ith scaling. To recast (6) as a linear vector equation, we define a matrix $\boldsymbol{A}$ of the form

$$\boldsymbol{A} = \left[A_{j_1} \vdots A_{j_2} \vdots \cdots \vdots A_{j_{N_J}} \right]^{l_r \times l_k}, \quad j_i \in \mathcal{J}, \tag{12}$$

where

$$A_{j_i} = \left[\alpha^r_{j_i k} \right]^{l_r \times k_i}, \quad j_i \in \mathcal{J}, \quad r \in \mathcal{R}, \quad K_l^{j_i} \le k \le K_u^{j_i}, \tag{13}$$

with l_r denoting the cardinality of the set $\mathcal{R}$ and $l_k = \sum_i k_i$. Furthermore, we define the vector of coefficients $\boldsymbol{c}$ (to be recovered) as

$$\mathbf{c} = \left[\mathbf{c}_{j_1}^\top \vdots \mathbf{c}_{j_2}^\top \vdots \cdots \vdots \mathbf{c}_{j_{N_J}}^\top \right]^\top, \quad j_i \in \mathcal{J}, \tag{14}$$

where $\mathbf{c}_{j_i} = [c_{j_i k}]^{k_i \times 1}$. Using these definitions, one may immediately rewrite (6) as

$$\boldsymbol{A}\mathbf{c} = \boldsymbol{\mu}, \quad \text{with} \quad \boldsymbol{\mu} = [\hat{\mu}^r_{f_X} - \mu^r_G]^{l_r \times 1}, \quad r \in \mathcal{R}. \tag{15}$$

We now include a constraint to guarantee the positivity of the estimated PDF ($\hat{f}_X \ge 0$). Note that, based on (4), we must have

$$G(x) + \sum_{j=j_m}^{j_m+N_J-1} \sum_{K_l^j}^{K_u^j} c_{jk} \psi_{jk}(x) \ge 0. \tag{16}$$

Suppose $\mathcal{B}$ is a set containing l uniformly chosen samples in the interval $[x_{\text{inf}}, x_{\text{sup}}]$, where x_{inf} and x_{sup} denote the infimum and the supremum of the input signal $\mathbf{x}$, respectively. To recast (16) as a linear inequality in matrix form, we define the matrix $\boldsymbol{\Psi}$ as follows:

$$\boldsymbol{\Psi} = \left[\Psi_{j_1} \vdots \Psi_{j_2} \vdots \cdots \vdots \Psi_{j_{N_J}} \right]^{l \times l_k}, \quad j_i \in \mathcal{J}, \tag{17}$$

where

$$\Psi_{j_i} = \left[\psi_{j_i k}(b) \right]^{l \times k_i}, \quad j_i \in \mathcal{J}, \quad b \in \mathcal{B}, \quad K_l^{j_i} \le k \le K_u^{j_i}. \tag{18}$$

Consequently, the positivity constraint on the estimated PDF may be formulated as

$$\boldsymbol{\Psi}\mathbf{c} \succeq -\boldsymbol{g}, \quad \text{with} \quad \boldsymbol{g} = [G(b)]^{l \times 1}, \quad b \in \mathcal{B}. \tag{19}$$

We must also ensure that the PDF estimate integrates to one. Based on our model,

$$\int_{-\infty}^{\infty} f_X(x)\,dx = 1 + \sum_{j=j_m}^{j_m+N_J-1} \sum_{K_l^j}^{K_u^j} c_{jk} \int_{-\infty}^{\infty} \psi_{jk}(x)\,dx. \tag{20}$$

As a result, we must have

$$\sum_{j=j_m}^{j_m+N_J-1} \sum_{K_l^j}^{K_u^j} c_{jk} \int_{-\infty}^{\infty} \psi_{jk}(x)\,dx = 0. \tag{21}$$

Note that for compact support wavelets we usually have $\int_{-\infty}^{\infty} \psi_{jk}(x)\,dx = 0$, which means that the property in (21) holds. However, in cases with $\int_{-\infty}^{\infty} \psi_{jk}(x)\,dx \neq 0$, we should have

$$\int_{-\infty}^{\infty} \psi_s(x)\,dx = \epsilon \Rightarrow \int_{-\infty}^{\infty} \psi_s(2^j x - k)\,dx = 2^{-j}\epsilon, \tag{22}$$

where ϵ is an arbitrary constant. Based on (21) and (22), the following relation is obtained:

$$\sum_{j=j_m}^{j_m+N_J-1} \sum_{K_l^j}^{K_u^j} c_{jk} 2^{-j} = 0. \tag{23}$$

To formulate the linear equality (23) in matrix form, the following definition may be considered:

$$\mathbf{d} = \left[\cdots \quad \underbrace{2^{-j_i} \cdots 2^{-j_i}}_{k_i \text{ times}} \quad \cdots \right]^\top, \quad j_i \in \mathcal{J}, \quad k_i \in \mathcal{K}. \tag{24}$$

Using this definition, (23) can be simply written as $\mathbf{d}^\top \mathbf{c} = 0$. Based on (15), (19) and (23), to find the unknown vector of coefficients $\mathbf{c}$, one should consider the following optimization problem:

$$\begin{aligned} \min_{\mathbf{c}} \quad & \|\boldsymbol{A}\mathbf{c} - \boldsymbol{\mu}\|_2^2 \\ \text{s.t.} \quad & -\boldsymbol{\Psi}\mathbf{c} \preceq \boldsymbol{g}, \\ & \mathbf{d}^\top \mathbf{c} = 0, \end{aligned} \tag{25}$$

which is a linearly constrained quadratic program [25]. Since $\boldsymbol{A}^\top \boldsymbol{A}$ is positive semi-definite for all $\boldsymbol{A}$, the quadratic program in (25) is always *convex* [26].

To show the effectiveness of our wavelet-focused formulation, in Section V, we will consider applying (25) in the case of B-spline wavelets as an example mother wavelet function. We only consider the cubic (ψ_c) B-spline wavelet [27]–[30]. Once the coefficients $\{\alpha^r_{jk}\}$ are obtained, one can formulate the optimization problem in (25) and obtain the estimated PDF using (4).

IV. Characteristic Function Approach

Some distributions lack the theoretical moments (μ^r) as $\mu^r \to \infty$. Therefore, for such distributions, the natural estimation of moments may diverge which in turn makes our problem ill-posed [14]. In order to avoid this issue, the idea of employing the characteristic function is introduced to obtain the coefficients $\{c_{jk}\}$ in (4) instead of moment matching.

A. Characteristic Function-Aided PDF Estimation

Applying the Fourier transform to (4) yields

$$\int_{-\infty}^{\infty} e^{i\omega x} f_X(x)\,dx = \int_{-\infty}^{\infty} e^{i\omega x} G(x)\,dx + \int_{-\infty}^{\infty} e^{i\omega x} e(x)\,dx,$$
$$F_X(\omega) = G_f(\omega) + \sum_{j=j_m}^{j_m+N_J-1} \sum_{K_l^j}^{K_u^j} c_{jk}\tau_{jk}(\omega), \tag{26}$$

where $G_f(\omega) = e^{-\frac{\omega^2\sigma_f^2}{2}}$, and $\tau_{jk}(\omega) = \int_{-\infty}^{\infty} e^{i\omega x}\psi_{jk}(x)\,dx$ is the Fourier transform of $\psi_{jk}(x)$. Note that $F_X(\omega)$ and $G_f(\omega)$ in (26) denote the characteristic functions of the input random variable $\mathbf{x}$ and the Gaussian random variable with distribution $\mathcal{N}(0,\sigma_f)$. Since the desired PDF $f_X(x)$ is unknown, $F_X(\omega)$ can be estimated from the input sample data via the *empirical characteristic function* (ECF) as [31]:

$$\hat{F}_X(\omega) = \frac{1}{N_{\mathbf{x}}} \sum_{n=1}^{N_{\mathbf{x}}} e^{i\omega x_n}. \tag{27}$$

One advantage of using the characteristic function is that it exists for all real-valued random variables, even for distributions that do not have bounded moments, such as the Cauchy distribution, and heavy-tailed processes in general [14]. Similar to Section III, we will take advantage of (26) along with the positivity constraint to estimate the input PDF. Suppose $\mathcal{W}$ contains l_ω frequency values. Define the matrix $\boldsymbol{B}$ as

$$\boldsymbol{B} = \left[B_{j_1} \vdots B_{j_2} \vdots \cdots \vdots B_{j_{N_J}}\right]^{2l_\omega \times l_k}, \quad j_i \in \mathcal{J}, \tag{28}$$

where

$$B_{j_i} = \left[\tau_{j_i k}(\omega)\right]^{2l_\omega \times k_i}, \quad j_i \in \mathcal{J}, \quad \omega \in \mathcal{W}, \quad K_l^{j_i} \le k \le K_u^{j_i}. \tag{29}$$

For each $\omega \in \mathcal{W}$, we have two equations in (26) emerging from its real and imaginary parts. Therefore, the total number of equations will be $2l_\omega$, as can be seen in our definition of the matrix $\boldsymbol{B}$ in (28). For a generic frequency value ω_0, and $E(\omega) = \int_{-\infty}^{\infty} e^{i\omega x} e(x)\,dx$, denoting the Fourier transform of $e(x)$, the real part of (26) can be written as:

$$\mathrm{Re}\left\{\hat{F}_X(\omega_0)\right\} = \mathrm{Re}\left\{G_f(\omega_0) + E(\omega_0)\right\}, \tag{30}$$

or equivalently,

$$\frac{1}{N_{\mathbf{x}}} \sum_{n=1}^{N_{\mathbf{x}}} \cos(\omega_0 x_n) = e^{-\frac{\omega^2\sigma_f^2}{2}} + \sum_{j=j_m}^{j_m+N_J-1} \sum_{K_l^j}^{K_u^j} c_{jk}\,\mathrm{Re}\left\{\tau_{jk}(\omega_0)\right\}, \tag{31}$$

while for its imaginary counterpart, we have

$$\mathrm{Im}\left\{\hat{F}_X(\omega_0)\right\} = \mathrm{Im}\left\{G_f(\omega_0) + E(\omega_0)\right\},$$
$$\frac{1}{N_{\mathbf{x}}} \sum_{n=1}^{N_{\mathbf{x}}} \sin(\omega_0 x_n) = \sum_{j=j_m}^{j_m+N_J-1} \sum_{K_l^j}^{K_u^j} c_{jk}\,\mathrm{Im}\left\{\tau_{jk}(\omega_0)\right\}. \tag{32}$$

Therefore, (26) can be represented as a linear equation in matrix form as follows:

$$\boldsymbol{B}\mathbf{c} = \boldsymbol{\gamma},$$
$$\boldsymbol{\gamma}^\top = \left[\mathrm{Re}\{\gamma^\star\} \vdots \mathrm{Im}\{\gamma^\star\}\right]^{1\times 2l_\omega}, \quad \omega \in \mathcal{W}, \tag{33}$$
$$\gamma^\star = \hat{F}_X(\omega) - e^{-\frac{\omega^2\sigma_f^2}{2}}.$$

Based on (33), (19) and (23), we consider the following optimization problem to recover the coefficient vector $\mathbf{c}$:

$$\begin{aligned} \min_{\mathbf{c}} \quad & \|\boldsymbol{B}\mathbf{c} - \boldsymbol{\gamma}\|_2^2 \\ \text{s.t.} \quad & -\boldsymbol{\Psi}\mathbf{c} \preceq \boldsymbol{g}, \\ & \mathbf{d}^\top\mathbf{c} = 0. \end{aligned} \tag{34}$$

As in (25), this is a *convex* quadratic program.

In the following, the Mexican Hat wavelet function will be utilized as an illustrative example for the significant potential of the proposed approach. The *Mexican Hat* wavelet, also known as the *Ricker* wavelet, is defined as [32]–[34],

$$\psi_h(x) = \frac{2}{\sqrt{3\sigma}\pi^{\frac{1}{4}}}\left(1 - \left(\frac{x}{\sigma}\right)^2\right) e^{-\frac{x^2}{2\sigma^2}}. \tag{35}$$

The Fourier transform of the Mexican Hat wavelet function is given by

$$\tau(\omega) = \frac{2\sqrt{2}}{\sqrt{3}}\pi^{\frac{1}{4}}\sigma^{\frac{5}{2}}\omega^2 e^{-\frac{\sigma^2\omega^2}{2}}. \tag{36}$$

The Fourier transform of the multi-scale Mexican Hat wavelet can be obtained as

$$\begin{aligned} \tau_{jk}(\omega) &= 2^{-j} e^{i\frac{\omega}{2^j}k}\,\tau\left(\frac{\omega}{2^j}\right), \\ &= \frac{2\sqrt{2}\pi^{\frac{1}{4}}\sigma^{\frac{5}{2}}}{\sqrt{3}} 2^{-3j} e^{i\frac{\omega}{2^j}k}\omega^2 e^{-\frac{\sigma^2\omega^2}{2^{2j+1}}}. \end{aligned} \tag{37}$$

These closed-form expressions can be used to form and solve (34), which will lead to an estimate of the input PDF.

V. Numerical Illustrations

In this section, the efficacy of the proposed PDF estimator is evaluated using the moment matching technique with the B-spline wavelet (cubic) introduced in Section III and the proposed characteristic function-aided method deploying the Mexican Hat wavelet introduced in Section IV. To examine our approach, we define the mixture model for the PDF of interest as below:

$$f_X(x;\theta) = \sum_{i=1}^{N_M} s_i f_i(x;\theta_i), \tag{38}$$

where $\{\theta_i\}$ are the parameters of the sub-PDFs $\{f_i\}$. For the B-spline wavelet example, we utilize the Gaussian Mixture Model (GMM) for which all sub-PDFs in (38) are Gaussian. For the Mexican Hat wavelet example, we use the (i) Gaussian mixture, the (ii) Cauchy-Exponential-Gaussian mixture: f_1 is a Cauchy distribution with the parameters x_0 as the location parameter and γ as the scale parameter, f_2 is the exponential distribution with the parameter λ as the rate parameter, f_3 is a Gaussian distribution $\mathcal{N}(\mu,\sigma)$, and

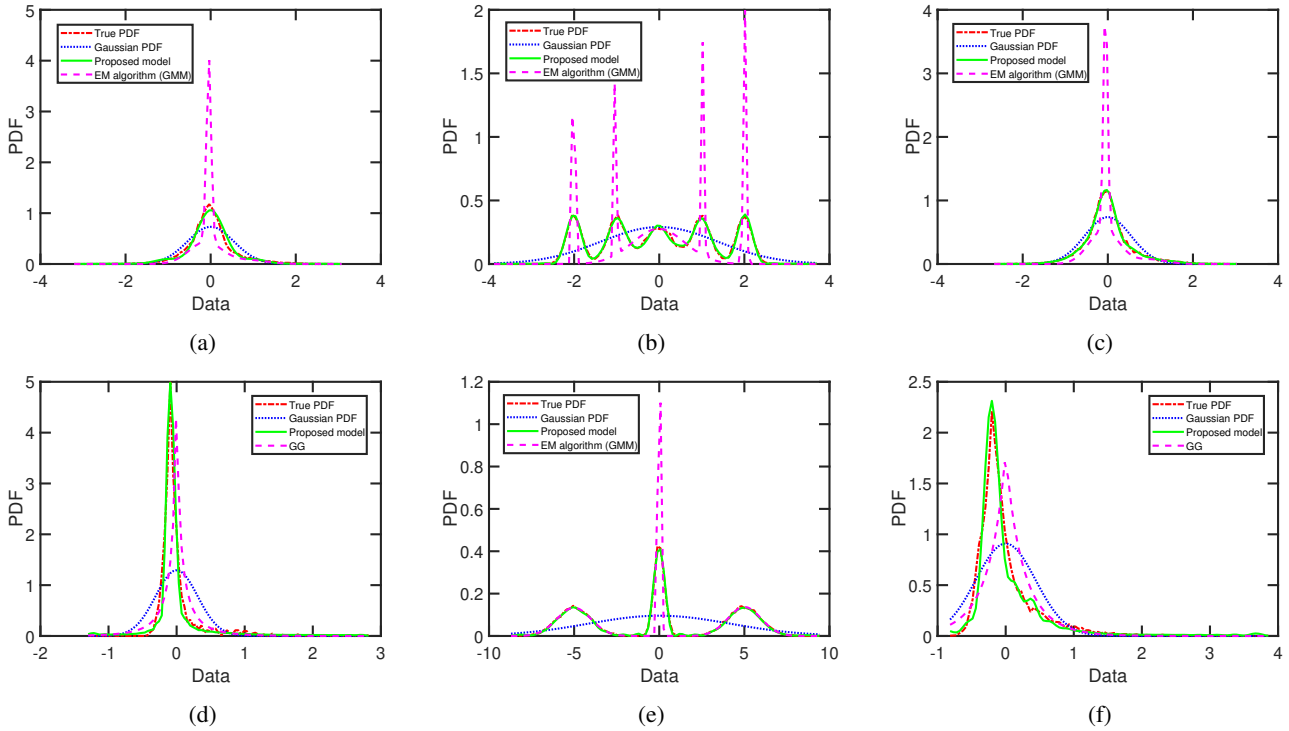

Figure 1: PDF estimation of the input data following the model (38) using the proposed model with the B-spline wavelet: (a) presents the Gaussian mixture model with parameters $\boldsymbol{\mu} = [0, -0.1, 0.2]$, $\boldsymbol{\sigma} = [0.2, 0.4, 0.8]$ and $\mathbf{s} = 1/3 \times \mathbf{1}_3$ and the Mexican Hat wavelet; (b), (c) and (e) show the Gaussian mixture models with the parameters $\boldsymbol{\mu} = [-2, -1, 0, 0, 1, 2]$, $\boldsymbol{\sigma} = [0.18, 0.2, 0.3, 1, 0.2, 0.18]$, $\mathbf{s} = 1/6 \times \mathbf{1}_6$, $\boldsymbol{\mu} = [0, -0.1, 0.2]$, $\boldsymbol{\sigma} = [0.2, 0.4, 0.8]$, $\mathbf{s} = 1/3 \times \mathbf{1}_3$, as well as $\boldsymbol{\mu} = [0, -5, 5]$, $\boldsymbol{\sigma} = [0.3, 1, 1]$, $\mathbf{s} = 1/3 \times \mathbf{1}_3$, respectively; (d) illustrates the Cauchy-Exponential-Gaussian mixture with parameters $\mu = 0$, $\sigma = 0.1$, $\lambda = 0.5$, $x_0 = 0$, $\gamma = 0.01$, $s_1 = 2/3$, $s_2 = 1/5$ and $s_3 = 2/15$; (f) shows the Exponential-Gaussian mixture with the parameters $\lambda = 0.5$, $\mu = 0$, $\sigma = 0.15$, $s_1 = 0.5$ and $s_2 = 0.5$.

the (iii) Exponential-Gaussian mixture: f_1 is the exponential distribution with the parameter λ as the rate parameter and f_2 is a Gaussian distribution.

The performance of PDF estimators are visually evaluated in Fig. 1, suggesting the satisfactory performance of the proposed approach in estimating the input PDFs in (38). In these examples, we cover both sub- and super- Gaussian distributions; namely, (b) and (e) are sub-Gaussian, whereas (a), (c), (d) and (f) are super-Gaussian.

To numerically scrutinize the proposed method, we utilize the mean integrated squared error (MISE) metric defined as $\mathbb{E}\left\{\int_{-\infty}^{\infty}\left(f_X(x) - \hat{f}_X(x)\right)^2 dx\right\}$ and the Hellinger distance which is given as $d_H^2\left(f_X, \hat{f}_X\right) = \int_{-\infty}^{\infty}\left(\sqrt{f_X(x)} - \sqrt{\hat{f}_X(x)}\right)^2 dx$ where f_X and $\hat{f}_X$ are the true PDF and the estimated PDF, respectively.

As can be seen in Table I, our model applied with the moment matching technique and characteristic function-based method is able to estimate the unknown PDF. We compare the proposed model with GMM (EM algorithm [35], [36]) for input data generated from the Gaussian mixture models (in Fig. 1 (a), (b), (c) and (e)) and with Generalized Gaussian Distribution (GG), which is a widely used parametric PDF model in the literature and a strong tool to estimate centralized PDFs with high magnitudes [12], [37], for input data generated from Cauchy-Exponential-Gaussian mixture and the Exponential-Gaussian mixture (in Fig. 1 (d) and (f)). As can be observed, our approach appears to have a better performance based on these metrics.

Table I: Performance Comparison for the Proposed PDF Estimator Based on the Results in Fig. 1

	Proposed Model		Compared Model		
Figure	MISE	d_H^2	Model	MISE	d_H^2
a	1.84e-03	2.01e-03	GMM (EM)	1.12e-01	2.02e-02
b	1.31e-04	6.21e-04	GMM (EM)	7.61e-02	4.96e-02
c	2.27e-04	3.96e-04	GMM (EM)	1.25e-01	2.32e-02
d	2.62e-02	2.00e-02	GG	2.81e-01	5.17e-02
e	3.09e-05	7.14e-04	GMM (EM)	7.16e-03	9.21e-03
f	1.02e-02	6.56e-03	GG	9.85e-02	2.59e-02

VI. Conclusion

We proposed a parametric PDF estimation method with a Gaussian penalty and wavelet expanded error function. The coefficients of our expansion model are determined via a convex program ensuring the positivity of density functions. Numerical results showcase the effectiveness of the proposed approach in obtaining the unknown PDF of the input data.

REFERENCES

[1] Luc Devroye and Lazlo Gyorfi, *Non-parametric density estimation*, Wiley, 1985.

[2] Luc Devroye and Gábor Lugosi, *Combinatorial methods in density estimation*, Springer Science & Business Media, 2001.

[3] Andrew B Nobel, Gusztáv Morvai, and Sanjeev R Kulkarni, "Density estimation from an individual numerical sequence," *IEEE Transactions on Information Theory*, vol. 44, no. 2, pp. 537–541, 1998.

[4] Aaron B Wagner, Pramod Viswanath, and Sanjeev R Kulkarni, "Probability estimation in the rare-events regime," *IEEE Transactions on Information Theory*, vol. 57, no. 6, pp. 3207–3229, 2011.

[5] Rebecca M Willett and Robert D Nowak, "Multiscale Poisson intensity and density estimation," *IEEE Transactions on Information Theory*, vol. 53, no. 9, pp. 3171–3187, 2007.

[6] Elias Masry, "Probability density estimation from sampled data," *IEEE Transactions on Information Theory*, vol. 29, no. 5, pp. 696–709, 1983.

[7] Amir Aboubacar and Mohamed El Machkouri, "Recursive kernel density estimation for time series," *IEEE Transactions on Information Theory*, vol. 66, no. 10, pp. 6378–6388, 2020.

[8] Yanjun Han, Jiantao Jiao, and Tsachy Weissman, "Minimax estimation of discrete distributions," in *2015 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2015, pp. 2291–2295.

[9] Feng Liang and Andrew Barron, "Exact minimax strategies for predictive density estimation, data compression, and model selection," *IEEE Transactions on Information Theory*, vol. 50, no. 11, pp. 2708–2726, 2004.

[10] Boris Ryabko, "Compression-based methods for nonparametric prediction and estimation of some characteristics of time series," *IEEE Transactions on Information Theory*, vol. 55, no. 9, pp. 4309–4315, 2009.

[11] Jorge Silva and Shrikanth Narayanan, "Nonproduct data-dependent partitions for mutual information estimation: strong consistency and applications," *IEEE Transactions on Signal Processing*, vol. 58, no. 7, pp. 3497–3511, 2010.

[12] SM Mahbubur Rahman, M Omair Ahmad, and MNS Swamy, "Bayesian wavelet-based image denoising using the Gauss–Hermite expansion," *IEEE Transactions on Image Processing*, vol. 17, no. 10, pp. 1755–1771, 2008.

[13] Aleksandra Pizurica and Wilfried Philips, "Estimating the probability of the presence of a signal of interest in multiresolution single-and multiband image denoising," *IEEE Transactions on Image Processing*, vol. 15, no. 3, pp. 654–665, 2006.

[14] Maurice George Kendall, Alan Stuart, and J Keith Ord, *Kendall's advanced theory of statistics*, Oxford University Press, Inc., 1987.

[15] Ronald W Cornew, Donald E Town, and Lawrence D Crowson, "Stable distributions, futures prices, and the measurement of trading performance," *The Journal of Futures Markets (pre-1986)*, vol. 4, no. 4, pp. 531, 1984.

[16] Jacky C So, "The sub-Gaussian distribution of currency futures: stable Paretian or nonstationary?," *The Review of Economics and Statistics*, pp. 100–107, 1987.

[17] Mike Novey, Tülay Adali, and Anindya Roy, "A complex generalized Gaussian distribution—characterization, generation, and estimation," *IEEE Transactions on Signal Processing*, vol. 58, no. 3, pp. 1427–1433, 2009.

[18] Fangzhou Yao, Jeff Coquery, and Kim-Anh Lê Cao, "Independent principal component analysis for biologically meaningful dimension reduction of large biological data sets," *BMC Bioinformatics*, vol. 13, no. 1, pp. 1–15, 2012.

[19] Alexander Zayezdny, Daniel Tabak, Dov Wulich, and Peter Smith, *Engineering applications of stochastic processes: theory, problems and solutions*, vol. 5, Taylor & Francis Group, 1989.

[20] David L Donoho, Iain M Johnstone, Gérard Kerkyacharian, and Dominique Picard, "Density estimation by wavelet thresholding," *The Annals of Statistics*, pp. 508–539, 1996.

[21] Adrian M Peter and Anand Rangarajan, "Maximum likelihood wavelet density estimation with applications to image and shape matching," *IEEE Transactions on Image Processing*, vol. 17, no. 4, pp. 458–468, 2008.

[22] Ingrid Daubechies, *Ten lectures on wavelets*, SIAM, 1992.

[23] Stéphane Mallat, *A wavelet tour of signal processing*, Elsevier, 1999.

[24] Hamidreza Amindavar and James A Ritcey, "Padé approximations of probability density functions," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 30, no. 2, pp. 416–424, 1994.

[25] Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe, *Convex optimization*, Cambridge university press, 2004.

[26] Mokhtar S Bazaraa, Hanif D Sherali, and Chitharanjan M Shetty, *Nonlinear programming: theory and algorithms*, John Wiley & Sons, 2013.

[27] Michael Unser, "Ten good reasons for using spline wavelets," in *Wavelet Applications in Signal and Image Processing V*. International Society for Optics and Photonics, 1997, vol. 3169, pp. 422–431.

[28] Michael Unser, Akram Aldroubi, and Murray Eden, "B-spline signal processing. i. theory," *IEEE Transactions on Signal Processing*, vol. 41, no. 2, pp. 821–833, 1993.

[29] Charles K Chui and Jian-zhong Wang, "A cardinal spline approach to wavelets," *Proceedings of the American Mathematical Society*, vol. 113, no. 3, pp. 785–793, 1991.

[30] Akram Aldroubi, Murray Eden, and Michael Unser, "Discrete spline filters for multiresolutions and wavelets of l_2," *SIAM Journal on Mathematical Analysis*, vol. 25, no. 5, pp. 1412–1432, 1994.

[31] Andrey Feuerverger and Roman A Mureika, "The empirical characteristic function and its applications," *The Annals of Statistics*, pp. 88–97, 1977.

[32] Ralph Brinks, "On the convergence of derivatives of b-splines to derivatives of the Gaussian function," *Computational & Applied Mathematics*, vol. 27, pp. 79–92, 2008.

[33] J-M Bardet, "Statistical study of the wavelet analysis of fractional brownian motion," *IEEE Transactions on Information Theory*, vol. 48, no. 4, pp. 991–999, 2002.

[34] Xiangcheng Mi, Haibao Ren, Zisheng Ouyang, Wei Wei, and Keping Ma, "The use of the Mexican Hat and the Morlet wavelets for detection of ecological patterns," *Plant Ecology*, vol. 179, no. 1, pp. 1–19, 2005.

[35] Douglas A Reynolds, "Gaussian mixture models.," *Encyclopedia of Biometrics*, vol. 741, pp. 659–663, 2009.

[36] Jeff A Bilmes et al., "A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models," *International Computer Science Institute*, vol. 4, no. 510, pp. 126, 1998.

[37] Minh N Do and Martin Vetterli, "Wavelet-based texture retrieval using generalized Gaussian density and Kullback-Leibler distance," *IEEE Transactions on Image Processing*, vol. 11, no. 2, pp. 146–158, 2002.