A Non-Classical Parameterization for Density Estimation Using Sample Moments

Guangyu Wu, *Student Member, IEEE*, and Anders Lindquist, *Life Fellow, IEEE*

***Abstract*—Probability density estimation is a core problem of statistics and signal processing. Moment methods are an important means of density estimation, but they are generally strongly dependent on the choice of feasible functions, which severely affects the performance. In this paper, we propose a non-classical parametrization for density estimation using sample moments, which does not require the choice of such functions. The parametrization is induced by the squared Hellinger distance, and the solution of it, which is proved to exist and be unique subject to a simple prior that does not depend on data, and can be obtained by convex optimization. Statistical properties of the density estimator, together with an asymptotic error upper bound are proposed for the estimator by power moments. Applications of the proposed density estimator in signal processing tasks are given. Simulation results validate the performance of the estimator by a comparison to several prevailing methods. To the best of our knowledge, the proposed estimator is the first one in the literature for which the power moments up to an arbitrary even order exactly match the sample moments, while the true density is not assumed to fall within specific function classes.**

***Index Terms*—Density estimation, squared Hellinger distance, parametric model, moment problem.**

I. Introduction

DENSITY estimation is an important problem of statistics and signal processing, which also lies in the core of numerous machine learning tasks, e.g. clustering and generative modeling. It can be formulated as follows. Given a set of independent and identically distributed (i.i.d.) samples from an unknown true distribution, find a density estimate that best describes the true one.

Since no prior information about the density function is given other than the data samples, it has been considered infeasible to treat the density estimation problem unless assuming the densities to fall within specific classes of functions, which we call a parametrization of the density. The mixture models, such as Parzen windows [1], [2] or mixtures of Gaussians or other basis functions [3], [4] are parameterized as mixtures of kernel functions, of which the type and the bandwidth need to be chosen carefully. However the performance of nonparametric algorithms is quite limited when the sample size is small.

On the other hand, power moments have been used to characterize the data samples. Methods matching the moments of the estimators to those of the data have been proposed in several papers [5]–[7]. However, these density estimators employ exponential family models, and the feasible density classes of these methods are very limited. The moment matching method for nonparametric mixture models proposed in [8] brings flexibility to the conventional moment methods, but a good knowledge of the function class is still required. Moreover, the existence of solution has not been proved in the previous papers. Neither are the statistical properties and error upper bounds proved, which severely lowers the value of those algorithms in application.

In conclusion, how to parameterize the density estimates given the samples is one of most significant problems in density estimation. In a long series of contributions, the parametrization has been separated into several small tasks. For example, mode estimation is about estimating the modes of a distribution, e.g. [1], [2], [9]–[15], with modes viewed as the central tendencies of a distribution. Class probability estimation involves estimating the probability distribution over a set of classes for a given input [16], etc. These results made significant contributions to the parametrization problem. However, since all of the tasks will bring individual biases to the parametrization, a parametrization of densities with minimum requirement of individual prior constraints, for example, the modes of the distribution and the function classes, is of great interest.

A parametrization for spectral density estimation using sample moments by Kullback-Leibler distance has been proposed in [17]. It only requires a prior spectral density independent of the samples. However in this problem, the number of data samples is limited. It makes the Kullback-Leibler divergence no longer the most satisfactory criterion to estimate the probability density functions, since especially it depends sensitively on events that are very rare in the reference distribution, which may induce sharp peaks in the density estimates. We naturally consider other metrics for density estimation using sample moments.

This paper treats the problem of density estimation by sample moments. The density estimation problem is formulated as a truncated Hamburger moment problem, and a solution to the proposed problem is proved to exist. The problem is then formulated in a matrix equation form and the squared Hellinger distance are used to form a convex optimization problem, and a parametrization of a rational form is proved to be the unique solution of it by proving the map from parameters of the parametrization to the sample moments being homeomorphic, which also makes it possible to apply gradient-based algorithms to treat the convex optimization problem. Then we prove the statistical properties of the proposed estimator. An asymptotic error upper bound of

Guangyu Wu is with Department of Automation, Shanghai Jiao Tong University, Shanghai, China. (e-mail: chinarustin@sjtu.edu.cn).

Anders Lindquist is with Department of Automation and School of Mathematical Sciences, Shanghai Jiao Tong University, Shanghai, China. (e-mail: alq@kth.se).

the estimator is also derived. Then we propose to use the density estimator in two signal processing tasks, including the observation noise estimation for filtering, and the recursive Bayesian estimator with non-Gaussian density functions. We also explain why the proposed estimator is indispensable in these tasks. Last but not the least, the simulation results of density estimation on mixtures of Gaussians, Laplacians and Gumbels are given, which validate the proposed density estimator. We emphasize that our density estimator can treat multi-modal densities without estimation/prior knowledge of modes or feasible classes.

II. Problem formulation

In this section, we will formulate our problem of density estimation using sample moments. We assume the density estimate to be supported on $\mathbb{R}$. Then the problem falls within the Hamburger moment problem [18], [19].

In the conventional Hamburger moment problem, one investigates whether a sequence is a feasible moment sequence. A sequence $\mu = (\mu_0, \mu_1, \ldots, \mu_\nu)$ is a feasible ν-sequence, if there exists a random variable X with a probability density function $p(x)$ supported on $\mathbb{R}$, of which the power moments are given by

$$\mu_k = \mathbb{E}\{X^k\} = \int_{\mathbb{R}} x^k p(x)dx, \quad k = 0, 1, \ldots, \nu.$$

Then any such random variable X is said to have a μ-feasible distribution and is denoted as $X \sim \sigma$.

However, in density estimation, we need an estimate of the probability density $p(x)$, a problem which may have infinitely many solutions. In this paper, we shall deal with a moment estimation problem to distinguish it from the conventional Hamburger moment problem. We should always remember that order ν moment estimation problem is ill-posed. Only if proper constraints are given, an analytic solution to the Hamburger moment problem can be obtained. Moreover, rather than the true moment sequence, we treat the Hamburger moment problem with a sample power moment sequence.

Definition II.1 (Order $2n$ moment density estimation problem)**.** Given a sequence $\mu = (\mu_0, \cdots, \mu_{2n})$ with

$$\mu_k = \frac{1}{m}\sum_{j=1}^{m} X_j^k, \quad k = 0, \ldots, 2n, \tag{1}$$

where $X_1, X_2, \ldots, X_m$ are independent and identically distributed samples. μ is the sample moment sequence. The estimation problem is then to find a density estimate $p(x)$ corresponding to a random variable $X \sim \mu$.

Thus density estimation using the truncated moment sequence obtained from the samples has been formulated as a Hamburger moment problem. Before treating this problem, we first need to prove the existence of solutions.

III. Existence of solutions

Since we are using sample moments, which due to sampling errors differ from the true population moments of the density function to be estimated, we need to prove that there exists a solution to Definition 1. To this end, we review some facts about the solvability of the power moment problem.

Theorem III.1 (Solution of the Hamburger Moment Problem [18])**.** *Denote the nonnegative integers as $\mathbb{N}_0$ and the positive Radon measures on the real numbers as $M_+(\mathbb{R})$. For a real sequence $s = (s_n)_{n\in\mathbb{N}_0}$ the following are equivalent:*
(i) s is a Hamburger moment sequence, that is, there is a Radon measure $\eta \in M_+(\mathbb{R})$ such that $x^n \in \mathcal{L}^1(\mathbb{R}, \eta)$ and

$$s_n = \int_{\mathbb{R}} x^n d\eta(x) \text{ for } n \in \mathbb{N}_0$$

(ii) The sequence s is positive semidefinite.
(iii) All Hankel matrices

$$H_n(s) = \begin{bmatrix} s_0 & s_1 & \ldots & s_n \\ s_1 & s_2 & \ldots & s_{n+1} \\ \vdots & \vdots & \ddots & \vdots \\ s_n & s_{n+1} & \ldots & s_{2n} \end{bmatrix}, \quad n \in \mathbb{N}_0$$

are positive semidefinite.

Next we shall prove that the truncated Hamburger moment problem in Definition II.1 is solvable.

Theorem III.2. *The problem in Definition 1 for μ with the moments given by (II.1) is solvable, if and only if $X_1, X_2, \ldots, X_m$ are not all equal. Moreover, the sequence μ is positive definite.*

Proof. We note that the empirical distribution function

$$\eta(x) = \frac{1}{m}\sum_{i=0}^{m} \mathbb{I}_{[X_i, +\infty)}(x),$$

where $\mathbb{I}$ is the indicator function, is a Radon measure. Then, by Theorem III.1, the sample moment sequence $\mu = (\mu_0, \mu_1, \cdots \mu_{2n})$ is a positive semidefinite sequence (because the full sample moment sequence is positive semidefinite). We note that a positive semidefinite sequence μ is positive definite if and only if $X_1, X_2, \ldots, X_m$ are not all equal, which is an event of probability $1 - \int_{\mathbb{R}} (p(x))^m dx$. Then by Corollary 9.2 in [18], we have that the truncated Hamburger moment problem for μ is solvable given that $X_1, X_2, \ldots, X_m$ are not all equal. □

IV. An analytic solution by squared Hellinger distance

In the previous section, solutions to the order $2n$ moment estimation problem is proved to exist (Theorem III.2). However, the existence of solutions doesn't necessarily guarantee analytic solutions to the moment problem. In this section, we will propose a method to obtain analytic solutions to this problem. In [17], the constraints on the sample moments were the positive definiteness of a Toeplitz matrix, Pick matrix or a similar object. In this paper, the appropriate Hankel matrix needs to be positive definite. Therefore, we write the Hamburger moment problem in a Hankel matrix form following some lines of thoughts in [17], [20].

The power moment conditions up to order $2n$ read

$$\mu_k = \int_{\mathbb{R}} x^k p(x)dx, \quad k = 0, 1, \ldots, 2n.$$

Then we can write them as a matrix equation

$$\int_{\mathbb{R}} F(x)p(x)F^T(x)dx = \mathcal{M}_{2n}, \tag{2}$$

where

$$F(x) = \begin{bmatrix} 1 \\ x \\ \vdots \\ x^{n-1} \\ x^n \end{bmatrix}$$

and

$$\mathcal{M}_{2n} = \begin{bmatrix} \mu_0 & \mu_1 & \cdots & \mu_n \\ \mu_1 & \mu_2 & \cdots & \mu_{n+1} \\ \vdots & \vdots & \ddots & \vdots \\ \mu_n & \mu_{n+1} & \cdots & \mu_{2n} \end{bmatrix}$$

with the power moments $\mu_k, k = 0, \cdots, 2n$, calculated by (1). $\mathcal{M}_{2n}$ is the Hankel matrix. By doing this, we give a formulation of Definition II.1.

We denote by $\mathcal{P}$ be space of all probability density functions supported on $\mathbb{R}$. Let $\mathcal{P}_{2n}$ be the subspace of all $p \in \mathcal{P}$ of which at least $2n$ orders of power moments (in addition to μ_0, which is 1) exist and are finite. By Theorem III.2, the set of all feasible $p \in \mathcal{P}$ which satisfy (2) is nonempty and that $\mathcal{M}_{2n} \succ 0$. Moreover, $\mathcal{M}_{2n}$ falls within the range of a linear operator

$$\Xi : p \mapsto \mathcal{M}_{2n} = \int_{\mathbb{R}} F(x)p(x)F^T(x)dx, \tag{3}$$

defined on $\mathcal{P}_{2n}$. Since $\mathcal{P}_{2n}$ is convex, $\operatorname{range}(\Xi) = \Xi\mathcal{P}_{2n}$ is also convex.

In the literature, the Kullback-Leibler (KL) distance is a commonly used measure of the difference between probability density functions [20]–[22]. However it doesn't satisfy the symmetric condition for being a metric. Moreover, the Kullback–Leibler divergence depends especially sensitively on events that are very rare in the reference distribution. Always formulated as minimizing the distance between a prior density and a proposal density [17], [23], to use KL divergence as the bona-fide distance measure for density estimation doesn't always yield satisfactory estimates.

Let r be an arbitrary probability density in $\mathcal{P}$. In this paper, following some lines of thoughts of [20], we propose to use the squared Hellinger distance instead of the KL divergence, which is written as

$$\mathbb{H}^2(r,p) = \int_{\mathbb{R}} \left(\sqrt{r(x)} - \sqrt{p(x)}\right)^2 dx \tag{4}$$

to consider the distance between r and p. There are several advantages to use the squared Hellinger distance. First it is jointly convex, and is a real distance metric. Second, it penalizes the estimation error in the sense of L2 norm. It may ameliorate the sharp peaks in the estimates, which is very common when the KL divergence is chosen as the distance measure. We emphasize that the choice of distance metric is not arbitrary. We observe that in the widely used measures, only the KL and the Hellinger distances yield an analytic parametrization for the order $2n$ moment density estimation problem.

Hellinger distance is also a widely used metric. However in the previous results, density estimation by Hellinger distance always needs a prescribed model, and the estimation is performed by estimating the parameters of the model [24], [25]. In this section, we introduce a parametrization of $p \in \mathcal{P}_{2n}$, which is induced by the squared Hellinger distance, but without any other estimation or prior knowledge of the modes and feasible density classes.

Theorem IV.1. *Given Ξ which is defined by (3), denote*

$$\mathcal{R}_+ := \left\{\Omega \in \operatorname{range}(\Xi) \mid F(x)^T \Omega F(x) > 0, x \in \mathbb{R}\right\}.$$

Minimizing (4) subject to $\Xi(p) = \mathcal{M}_{2n}$, i.e., subject to (2), yields a unique $p \in \mathcal{P}_{2n}$ which has the form

$$\hat{p} = \frac{r}{(1 + F^T\hat{\Omega}F)^2}, \tag{5}$$

for any choice of $r \in \mathcal{P}$ and $\mathcal{M}_{2n} \succ 0$. Moreover, minimizing the functional

$$\mathbb{J}_r(\Omega) := \operatorname{tr}(\Omega\mathcal{M}_{2n}) + \int_{\mathbb{R}} \frac{r}{1 + F^T\Omega F}dx \tag{6}$$

over all $\Omega \in \mathcal{R}_+$ yields a unique $\hat{\Omega}$.

Proof. We write the dual functional

$$L(p,\Omega) = \mathbb{H}^2(r,p) + \operatorname{tr}(\Omega(\Xi(p) - \mathcal{M}_{2n})),$$

where $\Omega \in \operatorname{range}(\Xi)$ is the Lagrange multiplier, then it amounts to maximizing the dual functional

$$\Omega \mapsto \inf_{p\in\mathcal{P}_{2n}} L(p,\Omega). \tag{7}$$

It is clear that $p \mapsto L(p,\Omega)$ is a strictly convex map. Therefore, to obtain the maximum of the right member of (7), a $p \in \mathcal{P}_{2n}$ satisfying $\delta L(p,\Omega;\delta p) = 0$ for all δp needs to be obtained. This will further constrain the choice of Ω. By denoting

$$\omega(x) := F(x)^T\Omega F(x) + 1, \tag{8}$$

the dual functional can be written as

$$\begin{aligned} &L(p,\Omega) \\ &= \int_{\mathbb{R}} \left(\sqrt{r(x)} - \sqrt{p(x)}\right)^2 dx + \int_{\mathbb{R}} (\omega(x) - 1)\, p(x)dx \\ &- \operatorname{tr}(\Omega\mathcal{M}_{2n}). \end{aligned}$$

The directional derivative reads

$$\delta L(p,\Omega;\delta p) = \int_{\mathbb{R}} \delta p(x) \left(\omega(x) - 1 + 1 - \frac{\sqrt{r(x)}}{\sqrt{p(x)}}\right) dx,$$

For a p to be a minimum point, all variations δp need to be zero, which is achieved only if

$$\omega(x) = \frac{\sqrt{r(x)}}{\sqrt{p(x)}}, \quad \text{i.e.,} \quad p(x) = \frac{r(x)}{\omega^2(x)}$$

for all $x \in \mathbb{R}$. □

Since $r(x)$ and $p(x)$ are supported on $\mathbb{R}$, which are both strictly positive, we have that $\omega(x)$ is also strictly positive. By (2) and (8), we shall further constrain $\Omega \in \mathcal{R}_+$.

Lemma IV.2. *$\Omega \in \mathcal{R}_+$ only if $\omega(x) > 0$.*

Proof. Given $\Omega \in \mathcal{R}_+$, Ω can be written as

$$\int_{\mathbb{R}} F(x)\psi(x)F^T(x)dx = \Omega,$$

where $\psi \in \mathcal{P}_{2n}$. Therefore we have

$$F^T \int_{\mathbb{R}} F\psi F^T dx F = F^T \Omega F = \omega - 1.$$

Since $\omega(x)$ is a scalar, it can be written as

$$\begin{aligned}\omega(x) &= \operatorname{tr}\left(F^T\Omega F\right) + 1\\ &= \operatorname{tr}\left(F^T \int_{\mathbb{R}} F\psi F^T dx \cdot F\right) + 1\\ &= \operatorname{tr}\left(F^T F \int_{\mathbb{R}} F\psi F^T dx\right) + 1\\ &= F^T F \operatorname{tr}\left(\int_{\mathbb{R}} F\psi F^T dx\right) + 1\\ &= F^T F \operatorname{tr}\left(\int_{\mathbb{R}} \sum_{i=0}^{n} x^{2i}\psi(x)dx\right) + 1\end{aligned}$$

where F^TF is a scalar. Since x^{2i}, ψ and F^TF are all positive, we have $\omega(x) > 0$. □

Then the dual functional reads

$$L\left(\frac{r}{q}, \Omega\right) = -\mathbb{J}_r(\Omega) + \int_{\mathbb{R}} r(x)dx,$$

where $\mathbb{J}_r$ is given by (6). The dual problem is now to minimize $\mathbb{J}_r(\Omega)$ over $\mathcal{R}_+$. To complete the proof we still need to prove the following theorem.

Theorem IV.3. *There exists a unique minimum $\hat{\Omega} \in \mathcal{R}_+$ for the functional $\mathbb{J}_r(\Omega)$ satisfying*

$$\Xi\left(\frac{r}{(1 + F^T\hat{\Omega}F)^2}\right) = \mathcal{M}_{2n}.$$

By this theorem, to be proved below,

$$\hat{p} = \frac{r}{\hat{q}^2}$$

where $\hat{q} = 1 + F^T\hat{\Omega}F$ belongs to $\mathcal{P}_{2n}$. Moreover, $\mathcal{P}_{2n}$ is a stationary point of the strictly convex dual functional $p \mapsto L(p, \hat{\Omega})$. Therefore, we have

$$L(\hat{p}, \hat{\Omega}) \leq L(p, \hat{\Omega}), \quad \forall p \in \mathcal{P}_{2n}.$$

Since $\Xi(\hat{p}) = \mathcal{M}_{2n}$,

$$\mathbb{H}^2(r, \hat{p}) \leq \mathbb{H}^2(r, p) \tag{9}$$

for all $p \in \mathcal{P}_{2n}$ satisfying the constraint (2). The necessary and sufficient condition of the equality in (9) is $p = \hat{p}$. Theorem IV.1 is proved.

Next, we shall finish the proof of Theorem IV.3. We first investigate the dual problem of minimizing $\mathbb{J}_r(\Omega)$ over $\mathcal{R}_+$.

Lemma IV.4. *Define the map $\zeta : \mathcal{R}_+ \to \mathcal{X}_+$ between $\mathcal{R}_+$ and $\mathcal{X}_+ := \{\mathcal{M}_{2n} \in \operatorname{range}(\Xi) \mid \mathcal{M}_{2n} \succ 0\}$ as*

$$\zeta : \ \Omega \mapsto \int_{\mathbb{R}} F(x)\frac{r(x)}{\omega^2(x)}F(x)^T dx$$

where $\omega(x)$ is defined by (8). An Ω is a stationary point of $\mathbb{J}_r(\Omega)$ only if

$$\zeta(\Omega) = \mathcal{M}_{2n}. \tag{10}$$

Proof. By (6) and (8) we have

$$\mathbb{J}_r(\Omega) := \operatorname{tr}(\Omega\mathcal{M}_{2n}) + \int_{\mathbb{R}} \frac{r}{1 + F^T\Omega F}dx$$

and therefore by

$$\delta q(\Omega; \delta\Omega) = F^T\delta\Omega F = \operatorname{tr}\{\delta\Omega F F^T\},$$

the directional derivative reads

$$\begin{aligned}&\delta\mathbb{J}_r(\Omega; \delta\Omega)\\ &= \operatorname{tr}\left(\delta\Omega\left[\mathcal{M}_{2n} - \int_{\mathbb{R}} F(x)\frac{r(x)}{\omega^2(x)}F(x)^T dx\right]\right).\end{aligned} \tag{11}$$

The necessary and sufficient condition for (11) being zero over all $\delta\Omega \in \operatorname{range}(\Xi)$ is (10). The proof is complete. □

We also need to establish the injectivity and surjectivity of the map $\zeta : \mathcal{R}_+ \mapsto \mathcal{X}_+$ for the proof of Theorem IV.3. By doing this, we will prove that there exists a unique solution of (10), and then equivalently the dual functional $\mathbb{J}_r$ has a unique minimum. We begin with injectivity.

Lemma IV.5. *The dual functional $\mathbb{J}_r(\Omega)$ is strictly convex.*

Proof. It requires to prove $\delta^2\mathbb{J}_r > 0$ where

$$\delta^2\mathbb{J}_r(\Omega; \delta\Omega) = \int_{\mathbb{R}} \frac{2r(x)}{\omega(x)^3}\left(F(x)^T\delta\Omega F(x)\right)^2 dx \tag{12}$$

By (12), we have $\delta^2\mathbb{J}_r \geq 0$. Now we still need to prove that

$$\delta^2\mathbb{J}_r > 0, \quad \text{for all} \quad \delta\Omega \neq \mathbf{0}.$$

By Lemma 3.6 in [20], it is proved that $\Omega \mapsto F^T\Omega F$ is injective. By replacing Ω with $\delta\Omega$, this lemma is proved. □

It follows from Lemma IV.5 that there is only one stationary point satisfying (10), i.e., the map $\zeta : \mathcal{R}_+ \to \mathcal{X}_+$ is injective.

Next, we shall prove the surjectivity of $\zeta : \mathcal{R}_+ \to \mathcal{X}_+$. A similar proof can be found in [20] and we omit it here. The injectivity and surjectivity proofs complete the proof of Theorem IV.3. Therefore $\zeta : \mathcal{R}_+ \to \mathcal{X}_+$ is a proper and injective continuous map between connected spaces of the same dimension, which is a homeomorphism. In conclusion, a unique $\hat{p}$ minimizing $\mathbb{H}^2(r, p)$ subject to $\Xi(p) = \mathcal{M}_{2n}$ can be obtained by solving the dual problem.

V. Statistical properties of the proposed density estimator

In the previous sections, we proposed a novel parametrization of density function using power moments by the squared Hellinger distance. In this section, we analyze the statistical properties of the proposed estimator. By paraphrasing Theorem 4.5.5 in [26], we conclude the following theorem.

Theorem V.1. *Denote the true density as p and the corresponding random variable as X. Suppose there is a unique distribution function F_p with the moments $\{\mu_k, k \geq 1\}$, all finite. Denote the estimated density by power moments up to order $2n$ as $\hat{p}_{2n}$, and the corresponding random variable* *as $\hat{X}_{2n}$. Suppose that $(F_{\hat{p}_{2n}})$ is a sequence of distribution functions, each of which has all its moments finite:*

$$\hat{\mu}_{2n,k} = \int_{-\infty}^{\infty} x^k dF_{\hat{p}_{2n}}.$$

Then we have

$$\mathbb{E}_p\left[\hat{\mu}_{2n,k}\right] = \mathbb{E}_p\left[\frac{1}{m}\sum_{j=1}^{m} X_j^k\right] = \frac{1}{m}\sum_{j=1}^{m}\mathbb{E}_p\left[X_j^k\right] = \mu_k.$$

With $n \to +\infty$, the following equation holds for every $k \geq 1$:

$$\lim_{n\to\infty}\mathbb{E}_p\left[\hat{\mu}_{2n,k}\right] = \mu_k.$$

Then $\hat{X}_{2n}$ converges to X in distribution.

Convergence in distribution is a relatively weak type of convergence, which requires the density estimate to be equal to the true density almost everywhere. Therefore Theorem V.1 is indeed a weaker version of asymptotic unbiasedness, with $n \to +\infty$, where asymptotic unbiasedness is the convergence in probability. Here we emphasize that "asymptotic" refers to the number of moment terms used $2n \to +\infty$ rather than the number of samples $m \to +\infty$. Next we prove the consistency of the proposed estimator. Denote the estimation error as $\Delta p = \hat{p}_{2n} - p$ and write the Taylor expansion of it at $x = 0$ as

$$\Delta p = \sum_{k=0}^{+\infty} \frac{x^k}{k!}\Delta p^{(k)}(0).$$

Then we write the estimation error in the L2 norm as

$$\begin{aligned} & L_2\left(\hat{p}_{2n}, p\right) \\ =& \int_{\mathbb{R}} (\Delta p)^2 dx \\ =& \int_{\mathbb{R}} \sum_{k=0}^{+\infty} \frac{x^k}{k!}\Delta p^{(k)}(0)\left(\hat{p}(x) - p(x)\right) dx \\ =& \sum_{k=0}^{+\infty} \frac{\Delta p^{(k)}(0)}{k!}\int_{\mathbb{R}} x^k\left(\hat{p}(x) - p(x)\right) dx. \end{aligned}$$

As assumed in Theorem V.1, all power moments of both the density estimate and the true density exist and are finite. By denoting the k_{th} order moment of $\hat{p}_{2n}, p_{2n}$ correspondingly as $\hat{\mu}_k, \mu_k, k \in \mathbb{N}_0$, we can write

$$L_2\left(\hat{p}_{2n}, p\right) = \sum_{k=0}^{+\infty} \frac{\Delta p^{(k)}(0)}{k!}\left(\hat{\mu}_k - \mu_k\right).$$

By our proposed density surrogates, the first $2n+1$ power moments of $\hat{p}$ are identical to those of p, i.e. $\hat{\mu}_k = \mu_k$ for $k = 0, 1, \cdots, 2n$. Therefore we have

$$L_2\left(\hat{p}_{2n}, p\right) = \sum_{k=2n+1}^{+\infty} \frac{\Delta p^{(k)}(0)}{k!}\left(\hat{\mu}_k - \mu_k\right).$$

Moreover, by the strong law of large numbers, we have

$$\lim_{m\to\infty}\hat{\mu}_k = \lim_{m\to\infty}\frac{1}{m}\sum_{i=1}^{m} X^k \xrightarrow{a.s.} \mu_k, \quad k = 0, 1, \cdots 2n. \tag{13}$$

Therefore, we have

$$L_2\left(\hat{p}_{2n}, p\right) \xrightarrow{a.s.} 0, \quad \text{with } n, m \to +\infty$$

which shows that the proposed estimator is almost surely consistent in the sense of L2 norm [27], [28], given $n \to +\infty$.

VI. An asymptotic error upper bound of the estimator

We will introduce an asymptotic upper bound of error for $\hat{p}(x)$, by the metric of total variation (TV) distance, which is a metric commonly adopted in the research on the moment problem.

The asymptotic TV distance reads

$$\begin{aligned} & \lim_{m\to\infty} TV(\hat{p}, p) \\ =& \lim_{m\to\infty}\sup_x\left|\int_{(-\infty,x]}(\hat{p} - p)dx\right| \\ =& \lim_{m\to\infty}\sup_x\left|F_{\hat{p}} - F_p\right| \end{aligned}$$

where $F_{\hat{p}}$ and F_p are the two distribution functions of the density estimate $\hat{p}$ and the true density p.

Denote $\hat{p}_t$ as the density estimate using the true population moments of p, instead of the sample moments. Then by Theorem V.1, we have $\lim_{m\to\infty}\hat{p} = \hat{p}_t$ almost surely. Finally we have

$$\lim_{m\to\infty} TV(\hat{p}, p) \xrightarrow{a.s.} TV(\hat{p}_t, p).$$

Shannon-entropy is adopted to derive an upper bound of the TV distance [29]. It reads

$$H[p] = -\int_{\mathbb{R}} p(x)\log p(x)dx.$$

We first introduce the Shannon-entropy maximizing distribution $F_{\breve{p}}$, of which the moments are the population moments of the true density. It has the following density function [30],

$$\breve{p}(x) = \exp\left(-\sum_{i=0}^{2n} m_i x^i\right)$$

where $m_0, \cdots, m_{2n}$ are determined by the following constraints,

$$\int_{\mathbb{R}} x^k \exp\left(-\sum_{i=0}^{2n} m_i x^i\right) dx = \mu_j^p, \quad k = 0, 1, \cdots, 2n.$$

By referring to [29], the KL distance between p and $\breve{p}$ then reads

$$\begin{aligned} & KL\left(p \| \breve{p}\right) \\ =& \int_{\mathbb{R}} p(x)\log\frac{p(x)}{\breve{p}(x)}dx \\ =& -H\left[p\right] + \sum_{i=0}^{2n} m_i \mu_j^p \\ =& H\left[\breve{p}\right] - H\left[p\right]. \end{aligned} \tag{14}$$

The KL distance between $\hat{p}_t$ and $\breve{p}$ reads

$$KL\left(\hat{p}_t \| \breve{p}\right) = H\left[\breve{p}\right] - H\left[\hat{p}_t\right]. \tag{15}$$

By [29], [31], together with (14), (15) , we obtain

$$\begin{aligned}&TV\left(\breve{p},\hat{p}_t\right)\\ \leq&3\left[\left(1+\frac{4}{9}KL\left(\hat{p}_t\|\breve{p}\right)\right)^{1/2}-1\right]^{1/2}\\ =&3\left[\left(1+\frac{4}{9}\left(H\left[\breve{p}\right]-H\left[\hat{p}_t\right]\right)\right)^{1/2}-1\right]^{1/2}\end{aligned}$$

and

$$TV\left(\breve{p},p\right)\leq 3\left[\left(1+\frac{4}{9}\left(H\left[\breve{p}\right]-H\left[p\right]\right)\right)^{1/2}-1\right]^{1/2}.$$

Then an asymptotic upper error bound is given by

$$\begin{aligned}&TV\left(\hat{p}_t,p\right)\\ =&\sup_x\left|F_{\hat{p}_t}\left(x\right)-F_p\left(x\right)\right|\\ \leq&\sup_x\left(\left|F_{\hat{p}_t}\left(x\right)-F_{\breve{p}}\left(x\right)\right|+\left|F_{\breve{p}}(x)-F_{p(x)}\right|\right)\\ \leq&\sup_x\left|F_{\hat{p}_t}\left(x\right)-F_{\breve{p}}\left(x\right)\right|+\sup_x\left|F_{\breve{p}}(x)-F_p(x)\right|\\ \leq&3\left[\left(1+\frac{4}{9}\left(H\left[\breve{p}\right]-H\left[\hat{p}_t\right]\right)\right)^{1/2}-1\right]^{1/2}\\ +&3\left[\left(1+\frac{4}{9}\left(H\left[\breve{p}\right]-H\left[p\right]\right)\right)^{1/2}-1\right]^{1/2}.\end{aligned}$$

If we are only given samples from the true density without knowing p, it is not possible for us to obtain the true $H[p]$. Under this circumstance, we approximate $H[p]$ by the empirical distribution function, which is $P\left(X=x_i\right)=r_i$. Then the Shannon entropy can be approximated as $H\left[p\right]=-\sum r_i\log r_i$.

VII. Applications in signal processing tasks

In the previous sections, we proposed an algorithm for density estimation using sample moments. However, the detailed mathematical treatments may have concealed the significance of the proposed density estimator. In [32], a probability density estimator was proposed and applied to subset/feature selection. In this section, we will also introduce several applications of the proposed estimator in signal processing, and explain why they are indispensable in these applications.

A. Observation noise estimation for Bayesian filters

In the conventional Kalman filter [33], [34] and its variants, including the extended Kalman filter (EKF) [35], the central-difference Kalman filter (CDKF) [36], [37], the unscented Kalman filter (UKF) [38], and the quadrature Kalman filter (QKF) [39], the probability density function of the additive observation noise is assumed to be Gaussian. Then estimating the density of the noise amounts to estimating the mean and variance of the noise. Given the samples of the noise, it is straightforward to obtain the sample mean and variance.

However, in real applications, the noises are always not Gaussian. In our recent papers [20], [40], we proposed a type of Bayesian filter based on power moments, where both the probability density functions of the system states and the observation noises can be non-Gaussian. Since the density function of the additive noise can be a non-Gaussian analytic one, the problem comes to estimating the observation noise given the samples from the noise. Since we are not provided with the model for estimation, conventional density estimation algorithms don't apply to this task. Furthermore, we desire the estimate to have a simple and analytic form of function for the ease of calculation, traditional nonparametric estimators are not proper options for this task either, due to the massive parameters. For example, given the i.i.d. samples $\left(X_1,X_2,\cdots,X_m\right)$ from the distribution of the observation noise, a typical kernel density estimator has the form

$$\hat{p}(x)=\frac{1}{m}\sum_{i=1}^{m}K_h\left(x-X_i\right)=\frac{1}{mh}\sum_{i=1}^{m}K\left(\frac{x-X_i}{h}\right)$$

where $K(\cdot)$ is the kernel function (a non-negative function), and $h>0$ is a smoothing parameter called the bandwidth. Since we are always provided with hundreds even thousands of samples, the kernel density estimator needs to store the same amount of parameters, together with the parameters of the kernel function and the bandwidth.

Compared with the kernel density estimators, our estimator has a much more compact form, where there are much less parameters to store. Assume that we use the first $2n$ orders of sample moments for density estimation. Our proposed estimator of the form (5) has only $2n$ parameters in $\hat{\Omega}$ and the parameters in the reference density $r(x)$. If we choose $r(x)$ to be a Gaussian density function, which is a pretty common choice, the number of parameters in the model is $2n+4$. The much fewer parameters greatly decreases the computation load for each filtering step, which reveals the significance of our proposed estimator.

B. Recursive Bayesian estimator with non-Gaussian density functions

We now consider a recursive Bayesian estimator with non-Gaussian density functions. Let the observation at time step k be u_k. Denote the sequence of observations as

$$\mathbf{u}_{1:k}\triangleq\left\{u_i,i=1,\ldots,k\right\},$$

with each $u_i\in\mathbb{R}$. The goal of the recursive Bayesian estimator is to estimate the posterior density function of the system state y_k, i.e.,

$$p\left(y_k\mid\mathbf{u}_{1:k}\right).$$

By Bayes formula, we have the following equation

$$\begin{aligned}&p\left(y_k\mid\mathbf{u}_{1:k}\right)\\ =&\frac{1}{p\left(u_k\mid\mathbf{u}_{1:k-1}\right)}p\left(y_k\mid\mathbf{u}_{1:k-1}\right)p\left(u_k\mid y_k\right)\\ \propto&p\left(y_k\mid\mathbf{u}_{1:k-1}\right)p\left(u_k\mid y_k\right)\end{aligned}\qquad(16)$$

Now the problem amounts to calculating the right-hand-side of (16) [41]. By the Chapman-Kolmogorov equation, we have

$$\begin{aligned}&p\left(y_k\mid\mathbf{u}_{1:k-1}\right)\\ =&\int p\left(y_k\mid y_{k-1}\right)p\left(y_{k-1}\mid\mathbf{u}_{1:k-1}\right)dy_{k-1}.\end{aligned}\qquad(17)$$

We note that $p\left(y_k \mid y_{k-1}\right)$ and $p\left(u_k \mid y_k\right)$ can be directly determined by the system equation and the observation equation respectively. In the Bayesian estimation problem we treat, the probability densities are not necessarily Gaussian. Hence we are not always able to obtain an analytic $p\left(y_k \mid \mathbf{u}_{1:k-1}\right)$ by (17). We are then confronted with the problem of how to treat the possibly intractable integral (17).

A common solution is to use the Monte-Carlo integration technique. The integral in (17) can then be approximated by

$$\begin{aligned} & p\left(y_k \mid \mathbf{u}_{1:k-1}\right) \\ \approx & \frac{\sum_{i=1}^{N} p\left(y_k \mid y_{k-1,i}\right) p\left(y_{k-1,i} \mid \mathbf{u}_{1:k-1}\right)}{\int \sum_{i=1}^{N} p\left(y_k \mid y_{k-1,i}\right) p\left(y_{k-1,i} \mid \mathbf{u}_{1:k-1}\right) dy_k} \end{aligned} \tag{18}$$

where $y_{k-1,i}$ for $i = 1, \cdots, N$ are N i.i.d. samples uniformly drawn from its domain. By doing this, we note that $p\left(y_{k-1,i} \mid \mathbf{u}_{1:k-1}\right)$ is a probability value and the r.h.s. of (18) is now a weighted sum of $p\left(y_k \mid y_{k-1,i}\right)$.

For a better estimation, N needs to be selected as a large positive integer. However, it will cause the number of parameters in the estimated $p\left(y_k \mid \mathbf{u}_{1:k-1}\right)$ to be quite great, which makes the density estimate complicated. With the estimator proposed in the previous sections, we consider using the power moments to parameterize the density function to give a more compact representation of it. We note that it is not a difficult task to calculate the power moments of the density estimate, namely the r.h.s. of (18). Then by properly selecting the number of power moments, we shall obtain a density estimate by power moments, which has the form (5), with the algorithm proposed in the previous sections.

In conclusion, density estimation is closely related to signal processing tasks, as proposed in this paper and in other previous ones. In the previous sections, we proposed the statistical properties and an asymptotic error upper bound of the density estimator. However, in the real applications of signal processing, people would be more interested in its real performance in different estimation tasks. As to better validate the performance of our proposed estimator, we perform three numerical simulations in the following section with a comparison to three prevailing estimators.

VIII. Monte Carlo simulations

This section reports the results of a Monte Carlo study designed to evaluate the performance of the proposed density estimator. We simulate mixtures of probability density functions, including Gaussian and non-Gaussian, smooth and non-smooth. These simulations validate the ability of the proposed density estimator as applied to much wider classes of functions.

We give performance comparisons of the following algorithms. First is the estimate by the density parametrization using moments by squared Hellinger distance (DPMSH), of which the curves are colored blue in all the subsequent figures. The orange curves are those of estimates by the density parametrization using moments by Kullback-Leibler distance (DPMKL), which was proposed in [20]. The green curves represent the estimates by a typical kernel density estimator (KDE), of which the kernel function is chosen as Gaussian and the corresponding bandwidth is chosen by Silverman's bandwidth selection. The red curves are the ones by the Gaussian mixture model (GMM) where the number of modes is set to be two for the five examples. We note that since previous methods of moments are not able to treat the density estimation problem without knowledge of the number of modes or feasible function class, we don't compare them to our proposed algorithm in this paper. In each of the following examples, a figure showing the true density and the average density estimates by the four algorithms, a figure showing the average TV distances between the true densities and the estimates over different number of samples, and one showing the Kullback-Leibler distance between the true density and the estimates shall be given for a complete comparison between the four algorithms.

The reference density r can usually be chosen as a Gaussian $r(x) = \mathcal{N}\left(m, \sigma^2\right)$, of which the parameters can be selected as $m = \mu_1$ and $\sigma^2 > \mu_2$. μ_1, μ_2 can be calculated by (1). Here we note that a relatively large variance σ^2 always yields better estimation performance for the density functions which have multiple peaks (modes).

The first example is a mixture of two Gaussians

$$p(x) = \frac{0.5}{\sqrt{2\pi}} \exp\left(\frac{(x-2)^2}{2}\right) + \frac{0.5}{\sqrt{2\pi}} \exp\left(\frac{(x+2)^2}{2}\right).$$

The prior r is chosen as a Gaussian distribution $\mathcal{N}(0, 6.7^2)$. The simulation results are given in Figure 1-3. Figure 1 shows the average density estimates of 50 Monte Carlo simulations with 100 data samples, i.e. $\mathbb{E}_p[\hat{p}(x)]$, which is used in density estimation to show the unbiasedness [27]. Figure 2 shows the TV distances between the density estimates and the true density by the four methods with different number of data samples. Figure 3 shows the Kullback-Leibler distances with different number of data samples. We observe in the left image that the average estimate by GMM is closest to the true density. However it is partly due to the prior knowledge that there are two Gaussians in the true density. We also note that the estimates by KDE suffer from the lack of data samples. The density estimate by DPMSH in this example uses the sample moments up to order 4. It has the second best performance, in the senses of both the TV distance and the Kullback-Leibler distance. We emphasize that unlike GMM, our proposed density estimator doesn't have prior knowledge of the true density to be estimated, e.g. the number of modes or the feasible function classes. As we mentioned in the previous sections, DPMKL has sharp peaks due to using the Kullback-Leibler distance.

The second example is another mixture of Gaussians

$$p(x) = \frac{0.7}{\sqrt{2\pi}} \exp\left(\frac{(x-2)^2}{2}\right) + \frac{0.3}{\sqrt{2\pi}} \exp\left(\frac{(x+2)^2}{2}\right).$$

We design this example to test the ability of the proposed estimator in estimating modes with small values of probability. The prior r is chosen as a Gaussian distribution $\mathcal{N}(-0.7, 6.2^2)$. The simulation results are given in Figure 4-6. Figure 4 shows the average density estimates of 50 Monte Carlo simulations with 100 data samples. GMM has the best performance. KDE and DPMSH have comparable

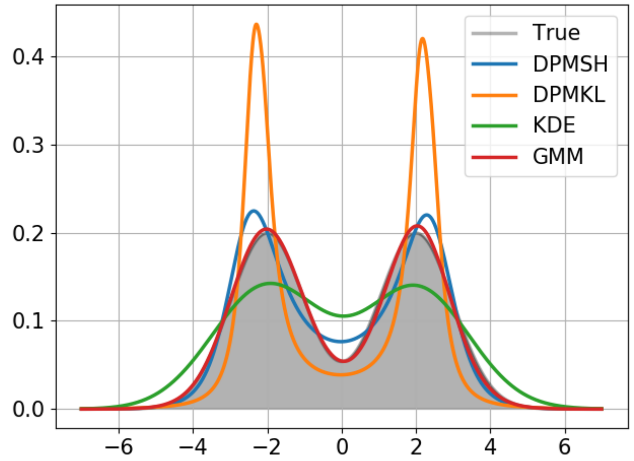

Fig. 1. Simulation results of Example 1. Average density estimates of 50 Monte Carlo simulations with 100 data samples.

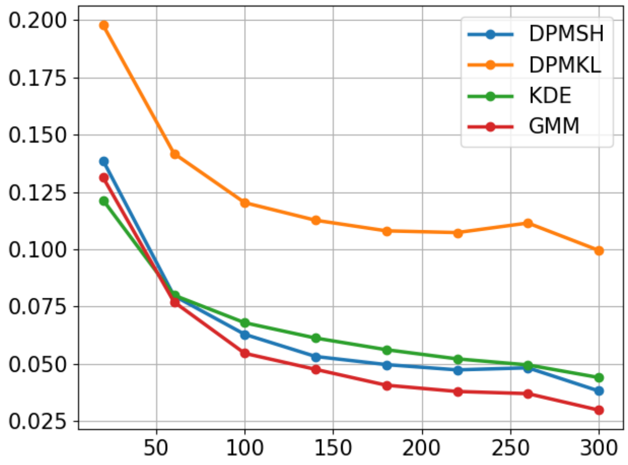

Fig. 2. Simulation results of Example 1. The total variation distances between the density estimates and the true density over different number of samples.

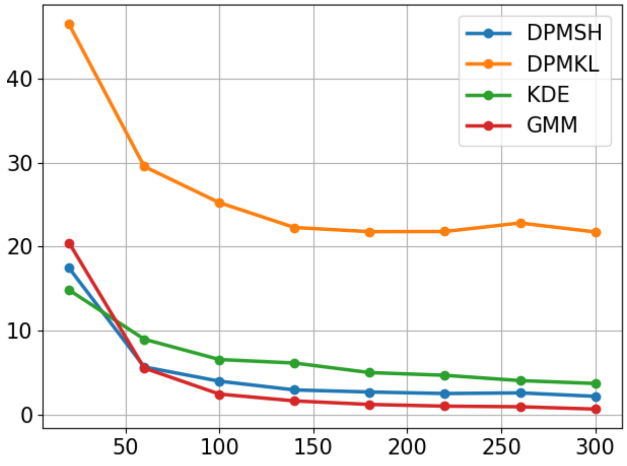

Fig. 3. Simulation results of Example 1. The Kullback-Leibler distances between the density estimates and the true density over different number of samples.

performances in the senses of both the TV distance and the KL distance. KDE model stores the same number of the parameters as the data samples. However there are only 5 parameters in our proposed DPMSH model, where $2n = 4$ in this example. It reveals the advantage of our proposed DPMSH over other methods.

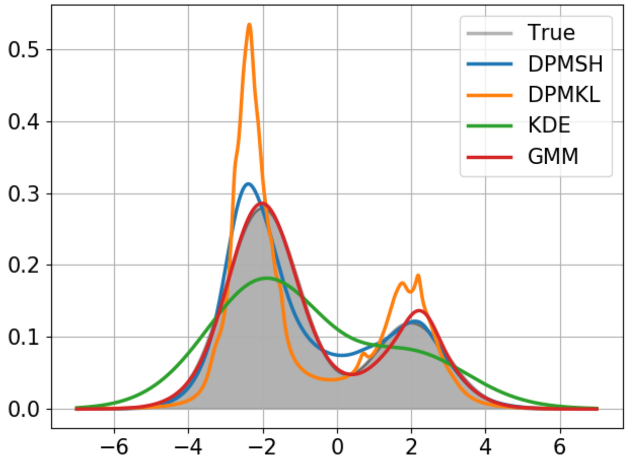

Fig. 4. Simulation results of Example 2. Average density estimates of 50 Monte Carlo simulations with 100 data samples.

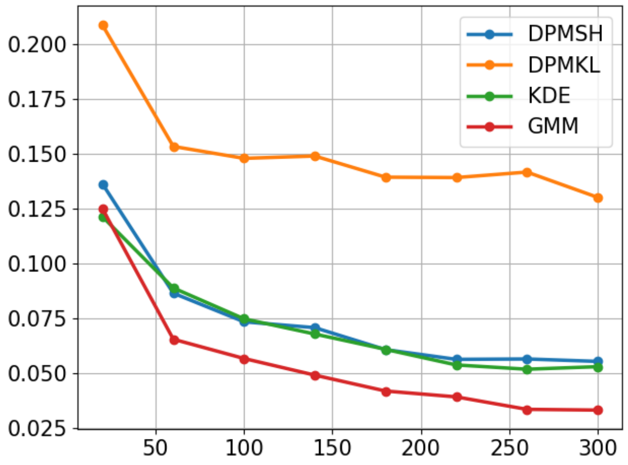

Fig. 5. Simulation results of Example 2. The TV distances between the estimators and the true density.

In the following two examples, we simulate on mixtures of non-Gaussian densities. Example 3 simulates a mixture of two Laplace distributions

$$p(x) = 0.5\exp\left(-2\left|x-2\right|\right) + 0.5\exp\left(-2\left|x+2\right|\right).$$

The prior r is chosen as a Gaussian distribution $\mathcal{N}(0, 6.5^2)$. The simulation results are given in Figure 7-9. Figure 7 shows the average density estimate of 50 Monte Carlo simulations with 200 data samples. We note that the performance of the density estimate by DPMSH using sample moments up to order 4 is better than KDE without prior knowledge of the number of modes.

Example 4 is a mixture of two Gumbel distributions

$$\begin{aligned} p(x) =& 0.5\exp\left(-\left(x-1+\exp\left(-(x-1)\right)\right)\right) \\ &+ 0.5\exp\left(-\left(x+1+\exp\left(-(x+1)\right)\right)\right) \end{aligned}$$

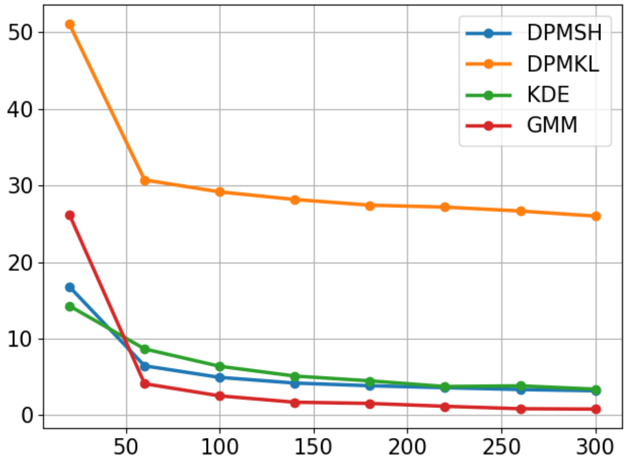

Fig. 6. Simulation results of Example 2. The KL distances between the estimators and the true density.

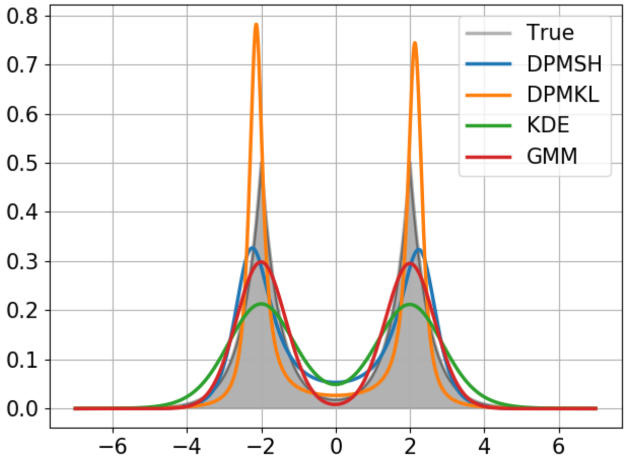

Fig. 7. Simulation results of Example 3. Average density estimates of 50 Monte Carlo simulations with 200 data samples.

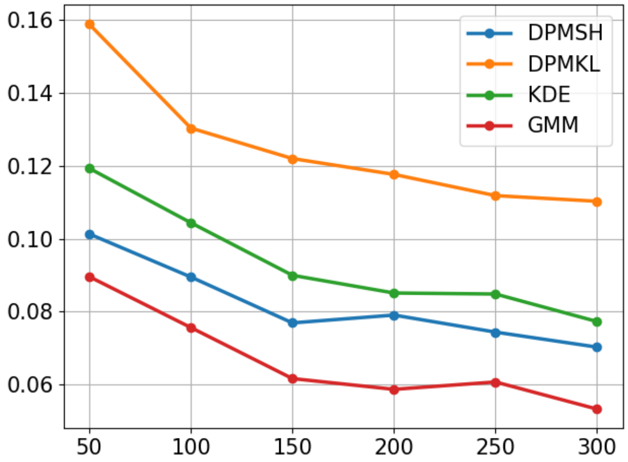

Fig. 8. Simulation result of Example 3. The TV distances between the estimators and the true density.

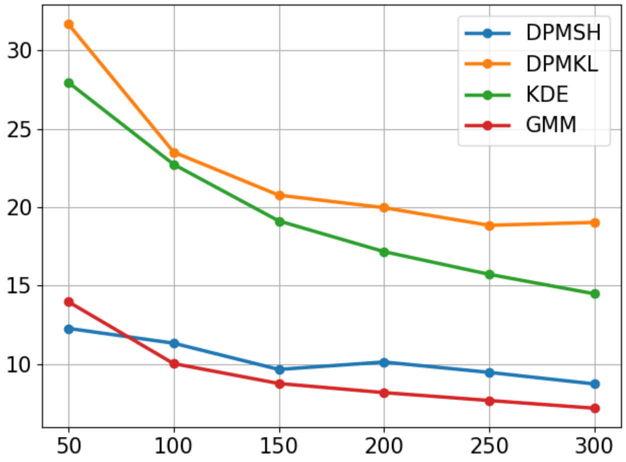

Fig. 9. Simulation result of Example 3. The KL distances between the estimators and the true density.

The prior r is chosen as a Gaussian distribution $\mathcal{N}(0.5, 3.5^2)$. The simulation results are given in Figure 10-12, which are the average of 50 Monte Carlo simulations with 200 data samples. In this example, the two modes are not easy to distinguish. Our proposed DPMSH, which uses sample moments up to order 6, obtains the best performance comparable to KDE. Since in this example, the prior constraint of the densities being Gaussian is no longer valid for GMM, the estimation performance of it is not as good as that of DPMSH. Moreover, except for the DPMKL estimate which has two distinct modes but is not close to the true density, only DPMSH approximates the two modes in the remaining three methods.

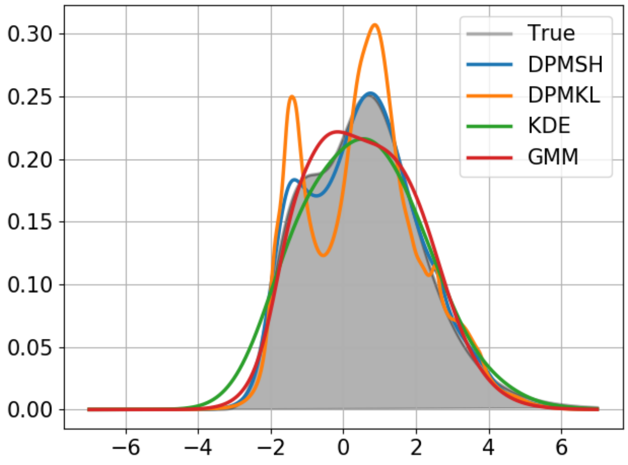

Fig. 10. Simulation result of Example 4. Average density estimates of 50 Monte Carlo simulations with 200 data samples.

Last we simulate the case where the number of densities in the mixture is larger than the number of modes. Example 5 is a mixture of 3 Gaussians, however there are only 2 modes,

$$p(x) = \frac{0.3}{\sqrt{2\pi}}\exp\left(\frac{(x-3)^2}{2}\right) + \frac{0.3}{\sqrt{2\pi}}\exp\left(\frac{(x+3)^2}{2}\right) + \frac{0.4}{\sqrt{2\pi}\cdot 2}\exp\left(\frac{(x-1)^2}{2\cdot 4}\right).$$

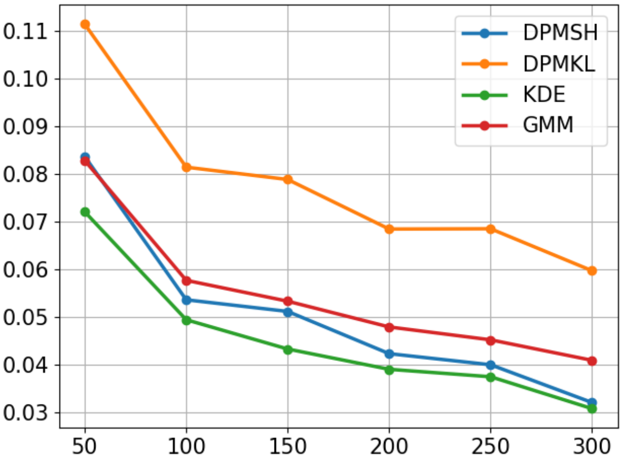

Fig. 11. Simulation result of Example 4. The TV distances between the estimators and the true density.

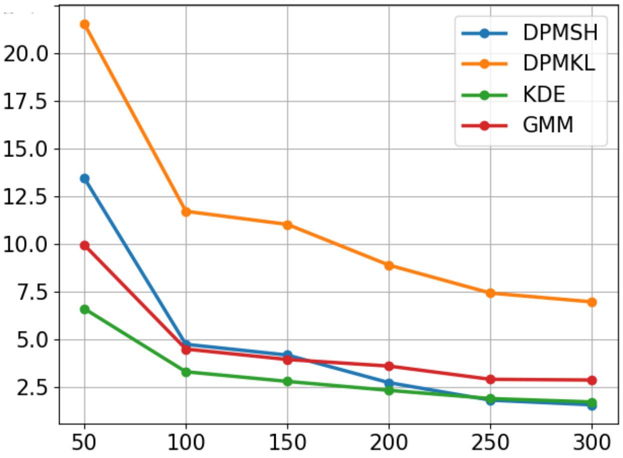

Fig. 12. Simulation result of Example 4. The KL distances between the estimators and the true density.

The prior r is chosen as a Gaussian distribution $\mathcal{N}(0.3, 5.0^2)$. The simulation results are given in Figure 13-15, which are the average of 50 Monte Carlo simulations with 200 data samples. In this example, we use sample moments up to order 6. We note that the performance of our proposed DPMSH estimate achieves the best performance. This example reveals the ability of our proposed parameterizaiton in estimating the modes which are a mixture of densities.

IX. Conclusion

We have developed an algorithm to parameterize and estimate probability density $p(x)$ on the real line from sample power moments by the squared Hellinger distance, leading to feasible solutions of the form (5). No prior constraints are imposed on the density to be estimated, such as a prescribed mixture of densities. The parametrization is in terms of a general prior density $r(x)$ with no particular connection to the data, generally chosen to be Gaussian. For each choice of prior $r(x)$ we obtain an analytic form the density estimate which is closest to $r(x)$ in the squared Hellinger distance. The map $\zeta : \mathcal{R}_+ \rightarrow \mathcal{X}_+$ is proved to be homeomorphic,

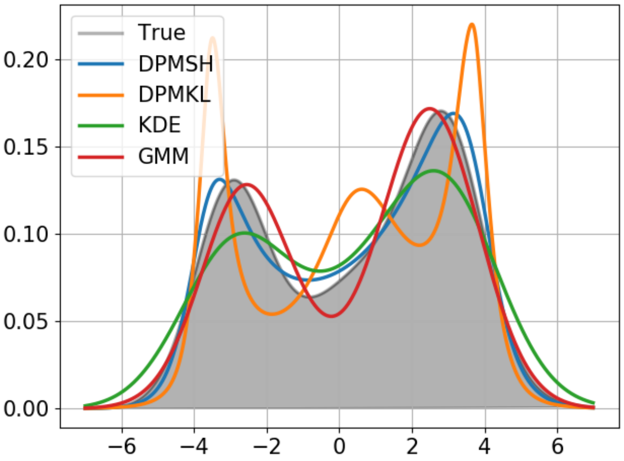

Fig. 13. Simulation result of Example 5. Average density estimates of 50 Monte Carlo simulations with 200 data samples.

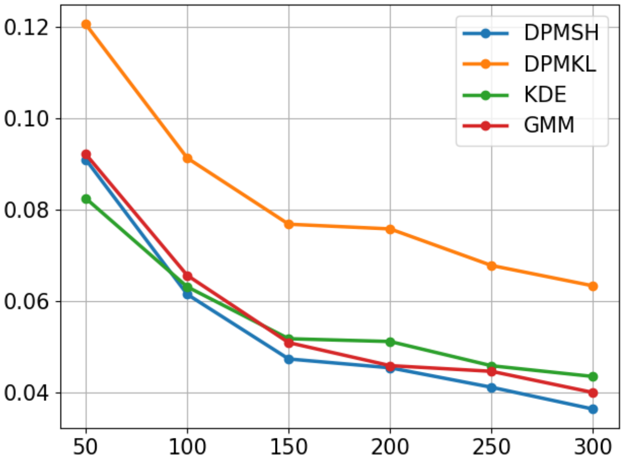

Fig. 14. Simulation result of Example 5. The TV distances between the estimators and the true density.

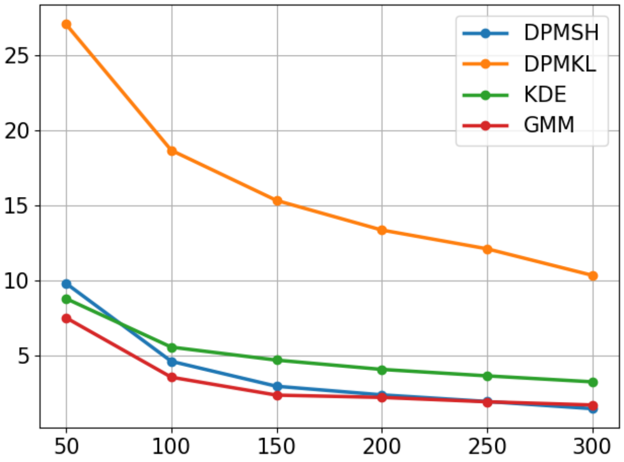

Fig. 15. Simulation result of Example 5. The KL distances between the estimators and the true density.

which establishes the existence and uniqueness of the solution. This also provides a convex optimization problem with the cost functional (6). Moreover, we propose statistical properties and an asymptotic error upper bound for the density estimator using power moments. Two of its important applications in the signal processing tasks are given, together with an explanation of its indispensability in these tasks. The simulation results on multi-modal density estimation also show the performance of the proposed estimator without prior information or estimation of the number of modes or the feasible classes of the density. The theoretical proofs and the simulation results both reveal the significance of the non-classical parametrization by power moments.

References

[1] E. Parzen, "On estimation of a probability density function and mode," *The annals of mathematical statistics*, vol. 33, no. 3, pp. 1065–1076, 1962.

[2] B. W. Silverman, *Density estimation for statistics and data analysis*. Routledge, 2018.

[3] G. J. McLachlan and K. E. Basford, *Mixture models: Inference and applications to clustering*. M. Dekker New York, 1988, vol. 38.

[4] F. Bunea, A. B. Tsybakov, and M. H. Wegkamp, "Sparse density estimation with $l1$ penalties," in *International Conference on Computational Learning Theory*. Springer, 2007, pp. 530–543.

[5] O. Barndorff-Nielsen, *Information and exponential families: in statistical theory*. John Wiley & Sons, 2014.

[6] M. Dudik, S. J. Phillips, and R. E. Schapire, "Performance guarantees for regularized maximum entropy density estimation," in *International Conference on Computational Learning Theory*. Springer, 2004, pp. 472–486.

[7] Y. Altun and A. Smola, "Unifying divergence minimization and statistical inference via convex duality," in *International Conference on Computational Learning Theory*. Springer, 2006, pp. 139–153.

[8] L. Song, X. Zhang, A. Smola, A. Gretton, and B. Schölkopf, "Tailoring density estimation via reproducing kernel moment matching," in *Proceedings of the 25th international conference on Machine learning*, 2008, pp. 992–999.

[9] H. Chernoff, "Estimation of the mode," *Annals of the Institute of Statistical Mathematics*, vol. 16, no. 1, pp. 31–41, 1964.

[10] W. F. Eddy, "Optimum kernel estimators of the mode," *The Annals of Statistics*, vol. 8, no. 4, pp. 870–882, 1980.

[11] Y. Cheng, "Mean shift, mode seeking, and clustering," *IEEE transactions on pattern analysis and machine intelligence*, vol. 17, no. 8, pp. 790–799, 1995.

[12] C. Abraham, G. Biau, and B. Cadre, "On the asymptotic properties of a simple estimate of the mode," *ESAIM: Probability and Statistics*, vol. 8, pp. 1–11, 2004.

[13] S. Dasgupta and S. Kpotufe, "Optimal rates for k-nn density and mode estimation," *Advances in Neural Information Processing Systems*, vol. 27, pp. 2555–2563, 2014.

[14] C. R. Genovese, M. P. Pacifico, I. Verdinelli, L. Wasserman *et al.*, "Minimax manifold estimation," *Journal of machine learning research*, vol. 13, pp. 1263–1291, 2012.

[15] H. Jiang and S. Kpotufe, "Modal-set estimation with an application to clustering," in *Artificial Intelligence and Statistics*. PMLR, 2017, pp. 1197–1206.

[16] P. Rigollet, "Generalization error bounds in semi-supervised classification under the cluster assumption." *Journal of Machine Learning Research*, vol. 8, no. 7, 2007.

[17] T. T. Georgiou and A. Lindquist, "Kullback-Leibler approximation of spectral density functions," *IEEE Transactions on Information Theory*, vol. 49, no. 11, pp. 2910–2917, 2003.

[18] K. Schmüdgen, *The moment problem*. Graduate Texts in Mathematics, 2017, vol. 277.

[19] D. Bertsimas and I. Popescu, "Optimal inequalities in probability theory: A convex optimization approach," *SIAM Journal on Optimization*, vol. 15, no. 3, pp. 780–804, 2005.

[20] G. Wu and A. Lindquist, "Non-Gaussian Bayesian filtering by density parametrization using power moments," *Automatica*, vol. 153, p. 111061, 2023.

[21] P. Hall, "On kullback-Leibler loss and density estimation," *The Annals of Statistics*, pp. 1491–1519, 1987.

[22] J. Q. Li and A. R. Barron, "Mixture Density Estimation." in *NIPS*, vol. 12, 1999, pp. 279–285.

[23] V. Vapnik, *The nature of statistical learning theory*. Springer science & business media, 1999.

[24] A. Cutler and O. I. Cordero-Brana, "Minimum Hellinger distance estimation for finite mixture models," *Journal of the American Statistical association*, vol. 91, no. 436, pp. 1716–1723, 1996.

[25] Z. Lu, Y. V. Hui, and A. H. Lee, "Minimum Hellinger distance estimation for finite mixtures of poisson regression models and its applications," *Biometrics*, vol. 59, no. 4, pp. 1016–1026, 2003.

[26] K. L. Chung, *A course in probability theory*. Academic press, 2001.

[27] A. J. Izenman, "Review papers: Recent developments in nonparametric density estimation," *Journal of the american statistical association*, vol. 86, no. 413, pp. 205–224, 1991.

[28] L. Gordon and R. A. Olshen, "Almost surely consistent nonparametric regression from recursive partitioning schemes," *Journal of Multivariate Analysis*, vol. 15, no. 2, pp. 147–163, 1984.

[29] A. Tagliani, "A note on proximity of distributions in terms of coinciding moments," *Applied Mathematics and Computation*, vol. 145, no. 2-3, pp. 195–203, 2003.

[30] J. N. Kapur and H. K. Kesavan, "Entropy optimization principles and their applications," in *Entropy and energy dissipation in water resources*. Springer, 1992, pp. 3–20.

[31] S. Kullback, "Correction to a lower bound for discrimination information in terms of variation," *IEEE Transactions on Information Theory*, vol. 16, no. 5, pp. 652–652, 1970.

[32] S. Kay, Q. Ding, B. Tang, and H. He, "Probability density function estimation using the eef with application to subset/feature selection," *IEEE Transactions on Signal Processing*, vol. 64, no. 3, pp. 641–651, 2015.

[33] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Journal of Basic Engineering*, vol. 82, no. 1, pp. 35–45, 1960.

[34] R. E. Kalman and R. S. Bucy, "New results in linear filtering and prediction theory," *Journal of Basic Engineering*, vol. 83, no. 1, pp. 95–108, 1961.

[35] B. D. Anderson and J. B. Moore, *Optimal filtering*. Courier Corporation, 2012.

[36] T. S. Schei, "A finite-difference method for linearization in nonlinear estimation algorithms," *Automatica*, vol. 33, no. 11, pp. 2053–2058, 1997.

[37] M. Norgaard, N. K. Poulsen, and O. Ravn, "New developments in state estimation for nonlinear systems," *Automatica*, vol. 36, no. 11, pp. 1627–1638, 2000.

[38] S. Julier, J. Uhlmann, and H. F. Durrant-Whyte, "A new method for the nonlinear transformation of means and covariances in filters and estimators," *IEEE Transactions on automatic control*, vol. 45, no. 3, pp. 477–482, 2000.

[39] K. Ito and K. Xiong, "Gaussian filters for nonlinear filtering problems," *IEEE transactions on automatic control*, vol. 45, no. 5, pp. 910–927, 2000.

[40] G. Wu and A. Lindquist, "A multivariate non-Gaussian Bayesian filter using power moments," *arXiv preprint arXiv:2211.13374*, 2022.

[41] H. A. Blom and E. A. Bloem, "Exact bayesian and particle filtering of stochastic hybrid systems," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 43, no. 1, pp. 55–70, 2007.

Guangyu Wu (S'22) received the B.E. degree from Northwestern Polytechnical University, Xi'an, China, in 2013, and two M.S. degrees, one in control science and engineering from Shanghai Jiao Tong University, Shanghai, China, in 2016, and the other in electrical engineering from the University of Notre Dame, South Bend, USA, in 2018. He is currently pursuing the Ph.D. degree at Shanghai Jiao Tong University.

Anders Lindquist (M'77–SM'86–F'89–LF'10) received the Ph.D. degree in optimization and systems theory from the Royal Institute of Technology (KTH), Stockholm, Sweden, in 1972, an honorary doctorate (Doctor Scientiarum Honoris Causa) from Technion (Israel Institute of Technology) in 2010, and Doctor Jubilaris from KTH in 2022.

He is currently a Zhiyuan Chair Professor at Shanghai Jiao Tong University, China, and Professor Emeritus at KTH, Stockholm, Sweden. Before that he had a full academic career in the United States, after which he was appointed to the Chair of Optimization and Systems at KTH.

Dr. Lindquist is a Member of the Royal Swedish Academy of Engineering Sciences, a Foreign Member of the Chinese Academy of Sciences, a Member of Academia Europaea (Academy of Europe), an Honorary Member the Hungarian Operations Research Society, a Life Fellow of IEEE, a Fellow of SIAM, and a Fellow of IFAC. He received the 2003 George S. Axelby Outstanding Paper Award, the 2009 Reid Prize in Mathematics from SIAM, and the 2020 IEEE Control Systems Award, the IEEE field award in Systems and Control.