EFFICIENT ESTIMATION OF MULTIVARIATE PROBABILITY DENSITY FUNCTION USING THE MAXIMUM ENTROPY PRINCIPLE

Lucas P. Damasceno[1], *Charles C. Cavalcante*[1], *and Zois Boukouvalas*[2]

[1]Dept. of Teleinformatics Eng., Federal University of Ceará, Fortaleza - CE, 60455-760, Brazil
[2]Dept. of Mathematics and Statistics, American University, Washington, DC 20016, USA

ABSTRACT

Due to the rapid growth of modern machine and deep learning techniques the multivariate probability density function estimation problem has received considerable attention in the broader area of signal processing. The main goal regarding this estimation problem is to achieve a desirable balance between flexibility while maintaining a simple form that would enable generalization, and efficient implementation. In this paper, we use the maximum entropy principle to achieve this goal and present a computational framework that is based on the Monte Carlo multidimensional integration in order to guarantee efficiency. We demonstrate the superior performance of the proposed technique over classical probability density estimation approaches using simulated data.

***Index Terms*—Multivariate probability density estimation, maximum entropy distributions, Gaussian Kernel, Monte Carlo methods.**

I. INTRODUCTION

The estimation of a multivariate probability density function (PDF) has been a common problem in a variety of classical disciplines such as computational physics and applied statistics [1]. Recently, within the field of artificial intelligence (AI) and its sub-fields machine learning and deep learning, which comprise the area of statistical signal processing, the knowledge of the multidimensional PDF that best matches the underlying properties of the data is necessary either explicitly or implicitly [2]. Examples of AI tasks that require knowledge of the underlying multivariate PDF, include Generative Adversarial Networks (GANs) [3], Variational Auto-encoders (VAEs) [4], among many others. Therefore, effective estimation of the multivariate PDF is vital to the success of these AI tasks.

Multidimensional density estimation approaches can be broadly classified as either parametric and non-parametric. Parametric methods provide a simple form for the PDF and are computationally efficient, however they are limited when the underlying distribution of the data deviates from the assumed parametric form. For instance, the multivariate Gaussian mixture model (GMM) [5], limits the PDFs to these that are symmetric and unimodal model, thus, yielding highly inaccurate results due to the inherent model mismatch. On the other hand, non-parametric methods, such as histogram estimation, k-nearest neighbors (kNN), and multivariate kernel density estimation (KDE) [6], [7], can provide flexible density matching. Although non-parametric techniques are not limited to any specific distribution, they are generally computationally demanding, especially when sample size is large, and they highly depend on the choice of tuning parameters. For instance, histogram, kNN, and KDE, highly depend on the choice of parameters, such as the number of bins, number of samples in a neighborhood, and the bandwidth for histogram, respectively.

Semi-parametric methods, such as those based on the maximum entropy principle (MEP) [8]–[10], combine the simple density form and the flexibility of non-parametric and parametric methods, yielding a global solution provided by the MEP [11]. In this paper, we present a multivariate density estimator algorithm that is based on the MEP and by jointly using global and local measuring functions we provide flexible PDFs while keeping the complexity low. The global measuring functions yield constraints on the statistics of the PDF, such as the mean, variance, and higher-order statistics (HOS) and the local measuring functions provide constraints on the overall statistics and gain insight into the local behavior of the source PDFs. In contrast to the univariate MEP density estimator, a main challenge in the multivariate case is the multi-dimensional integration during the estimation phase. To overcome this difficulty, we integrate into the proposed algorithm a multidimensional Monte-Carlo (MC) integration technique in order to achieve computational efficiency. We call the new algorithm multivariate entropy maximization with kernels (M-EMK) and consider it an extension of the univariate case presented in [9].

The remainder of this paper is organized as follows. In Section II, we provide a brief background on the maximum entropy distributions. Section III provides a framework for M-EMK. Simulation results that illustrate the performance of our proposal are shown in Section IV and, the conclusions and future research directions are stated in Section V.

II. BACKGROUND

The maximum entropy principle states that the probability distribution which best represents the current state of knowledge is the one with the largest entropy [11], [12]. The maximum entropy density, subject to known constraints, can be written as the following optimization problem [13]:

$$\begin{aligned} &\max_{p(\mathbf{x})} \; H(p(\mathbf{x})) = -\int_{\mathbb{R}^K} p(\mathbf{x}) \log p(\mathbf{x}) d\mathbf{x} \\ &\text{s.t.} \;\; \int_{\mathbb{R}^K} r_i(\mathbf{x}) p(\mathbf{x}) d\mathbf{x} = \alpha_i, \;\text{ for } i = 0, ..., M, \end{aligned} \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^K$, $p(\mathbf{x}) \geq 0$, and $r_i(\mathbf{x}) \in C(\mathbb{R}^K, \mathbb{R})$ for $i = 0, ..., M$. The space $C(\mathbb{R}^K, \mathbb{R})$, represents all measuring functions with domain $\mathbb{R}^K$ and co-domain $\mathbb{R}$, and $\alpha_i = \sum_{t=1}^{T} r_i(\mathbf{x}_t)/T$ for $i = 0, ..., M$ represent their corresponding sample averages, given observations $\mathbf{x}(t) \in \mathbb{R}^K$, $t = 1, ..., T$. We note that the first constraint need to be $\int_{\mathbb{R}^K} p(\mathbf{x}) d\mathbf{x} = 1$, equivalently $r_0 = 1$ and $\alpha_0 = 1$, in order $p(\mathbf{x})$ to be a valid PDF. The optimization problem in (1) can be rewritten in a Lagrangian form and is given by

$$\begin{aligned} \mathcal{L}(p(\mathbf{x})) = &-\int_{\mathbb{R}^K} p(\mathbf{x}) \log p(\mathbf{x}) d\mathbf{x} \\ &+ \sum_{i=0}^{M} \lambda_i \int_{\mathbb{R}^K} (r_i(\mathbf{x}) - \alpha_i) p(\mathbf{x}) d\mathbf{x}, \end{aligned} \tag{2}$$

where $\lambda_i, i = 0, ..., M$, are the Lagrangian multipliers. Through the use of functional variation, we can "differentiate" (2) with respect to p. By setting $\partial \mathcal{L}(p\mathbf{x})/\partial p = 0$, we obtain the equation of maximum entropy distribution,

$$p(\mathbf{x}) = \exp\left\{-1 + \sum_{i=0}^{M} \lambda_i r_i\left(\mathbf{x}\right)\right\}, \tag{3}$$

where Lagrangian multipliers are chosen such that p satisfies the constraints in (1). By substituting (3) into the constraints in (1), we generate a nonlinear system of $M+1$ equations for the $M+1$ Lagrangian multipliers.

III. M-EMK

III-A. Mathematical Formulations

We can evaluate the Lagrangian multipliers in (3) by the Newton iteration scheme, given by

$$\boldsymbol{\lambda}_{n+1} = \boldsymbol{\lambda}_n - \mathbf{J}^{-1} E_{p_n}\left\{\boldsymbol{r} - \boldsymbol{\alpha}\right\}, \tag{4}$$

where p_n is the estimated PDF for the nth iteration, E_{p_n} is the expectation over the distribution p_n, and $\mathbf{r} = [r_0, ..., r_M]$, $\boldsymbol{\lambda} = [\lambda_0, ..., \lambda_M]$, $\boldsymbol{\alpha} = [\alpha_0, ..., \alpha_M] \in \mathbb{R}^{M+1}$ denote the vectors of global and local measuring functions, the Lagrangian multipliers and sample averages, respectively. By $\mathbf{J} \in \mathbb{R}^{M \times M}$ we denote the Jacobian matrix where the ij-th entry of $\mathbf{J}$ is given by

$$\begin{aligned} \mathbf{J}_{ij} &= \int_{\mathbb{R}^K} \frac{\partial p(\mathbf{x})}{\partial \lambda_j} d\mathbf{x} \\ &= \int_{\mathbb{R}^K} r_i(\mathbf{x}) r_j(\mathbf{x}) p(\mathbf{x}) d\mathbf{x} = E_{p_n}\left\{r_i r_j\right\}. \end{aligned} \tag{5}$$

The i-th entry of $E_{p_n}\left\{\boldsymbol{r} - \boldsymbol{\alpha}\right\}$ is given by

$$E_{p_n}\left\{\boldsymbol{r} - \boldsymbol{\alpha}\right\} = \int_{\mathbb{R}^K} r_i(\mathbf{x}) p(\mathbf{x}) d\mathbf{x} - \alpha_i. \tag{6}$$

As we can see from (6), the optimization process highly depends on the proper selection of the constraints both in terms of the number of constraints as well as the different types of measuring functions that provide information about the underlying statistical properties of the data. Failure to do so may result in high complexity as well as poor data characterization.

III-B. Selection of Constraints

In this paper, we jointly use global and local constraints to provide flexible density estimation while keeping the complexity low. As in [9], [10] we use $\mathbf{1}, \mathbf{x}, \mathbf{x}^2, \mathbf{x}/(\mathbf{1} + \mathbf{x^2})$ as the global constraints, since they provide computational efficiency and desirable performance for a wide range of distributions. Furthermore, these global constraints provide information on the PDF's overall statistics, such as the mean, variance, and higher order statistics (HOS). For the local constraint we use the following Gaussian kernel, given by

$$q(\mathbf{x}) = \frac{1}{\sqrt{|\boldsymbol{\Sigma}|(2\pi)^K}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}) \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})'\right), \tag{7}$$

where $\boldsymbol{\mu}$ denotes the mean and $\boldsymbol{\Sigma}$ denotes the covariance matrix. The use of Gaussian kernels provides localized information about the PDF. As we will see in Section IV, adding the local constraint provides more effective density estimation as opposed to use only global constraints.

It is important to note that when we add the Gaussian kernel the multidimensional integrals present in (5) and (6) cover the entire space, which makes their computation a challenge due to the fact that the Gaussian kernel has infinite support.

III-C. Multidimensional Integration

As noted in the previous section, the multidimensional integration is one of the main challenges in our estimation problem. To overcome this difficulty, we integrate into the proposed approach an efficient multidimensional integration technique that is based on Quasi-Monte Carlo (QMC) methods. QMC methods are variants of the classical Monte Carlo (MC) methods and have shown to be efficient in terms of the rate of convergence and they provide a simple theoretical formulation making them ideal for our proposed approach [14]. QMC and MC methods both enjoy exactly the same form of approximating integrals. However, the purpose of the QMC methods is to achieve faster convergence than the rate of convergence that is provided by MC. The rate of convergence for MC methods is $O(1/\sqrt{n})$ and is often considered exceedingly slow, mainly when the function is smooth. QMC methods achieve a convergence rate of order $O((\log n)^K/n)$ or faster if sufficient smoothness of the function is assumed [15].

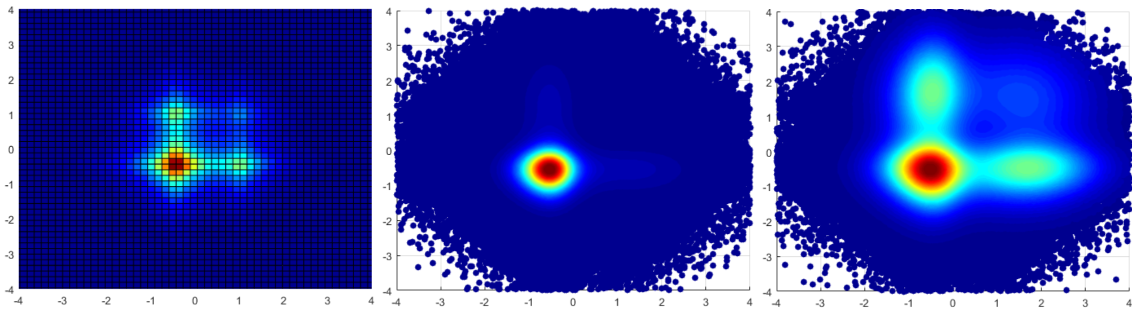

Fig. 1. The histogram of the two-dimensional generated data and its approximated PDFs using only global constraints, and adding the local constraint. Left–histogram of generated data from mixture of multivariate GGD. Middle–approximated PDF only using global constraints computed by the M-EMK. Right–approximated PDF adding the local constraint computed by the M-EMK.

In order to introduce the QMC integration methods in our approach, we need to generate a sequence of quasi-random points which presents an advantage in terms of convergence rate compared to pseudo-random sequence that MC integration methods use. To achieve this, we use the van der Corput sequence, which is an example of a one-dimensional low-discrepancy sequence that uniformly covers the unit hypercube and can be constructed using a computational efficient procedure [16]. The z-th quasi-random number w_z, is constructed in the following way:

Let b_k denote the k-th prime number, for instance, when $k = 1$ then $b_1 = 2$, when $k = 2$ then $b_1 = 3$ and so forth, and $Z_{b_k} = \{0, 1, ..., b_k - 1\}$ denotes the least residue system mod b_k. Every integer $n \geq 0$ has a unique digit expansion given by

$$n = \sum_{i=0}^{N-1} a_i(n) b_k^i \tag{8}$$

in base-b_k, where N is the sample size of quasi-random numbers and $a_i(n) \in Z_{b_k}$. The n-th quasi-random number is given by the following radical-inverse function in base-b_k,

$$w_n = \sum_{i=0}^{N-1} a_i(n) b_k^{-i-1}. \tag{9}$$

Using, the numbers generated by (9), we approximate the multidimensional integrals in (5) and (6) in a similar manner as we do using traditional MC methods. Thus, each of the integrals is evaluated by

$$Q_{N,K}\left(p(\mathbf{x})\right) = (\text{dim. mea.}) \left(\frac{1}{N} \sum_{i=0}^{N-1} p(w_n) \right), \tag{10}$$

where (dim. mea.) denotes the dimensional measure of the region of integration. For instance, length, area and volume for one, two and three-dimensional space, respectively.

IV. EXPERIMENTAL RESULTS

In this section, we demonstrate the performance estimation, computational efficiency, and flexibility of M-EMK by using simulated data. We compare M-EMK, with adaptive kernel density estimator (A-KDE) [17] which is a widely used non-parametric method and Gaussian mixture model (GMM) [5] which is classical parametric method. In all of the following experiments, we set the mean of the Gaussian kernel equal to the zero vector and $\mathbf{\Sigma}$ equal to the identity matrix. We generate data according to a mixture of generalized Gaussian distributions. The PDF is given by

$$p(\mathbf{x}; \beta, \mu, \sigma_i) = \sum_{i=1}^{L} \pi_i g_i(\mathbf{x}; \beta, \mu, \sigma_i), \ \mathbf{x} \in \mathbb{R}^K$$

where each g_i is a multivariate generalized Gaussian distribution defined in [18]. The shape and mean parameters for each of the components and each dimension are chosen to be $\beta = 0.5$ and $\mu = 1$ respectively. The weight parameters π_1 and π_2 are chosen to be equal to 0.3 and 0.7, respectively. For all the experiments we select $K = 2$.

IV-A. Estimation Performance

For the first experiment, we demonstrate how the performance of the density estimation is affected if we include only global measuring functions and how the performance is affected if we jointly include global and local measuring functions into the estimation phase. It should be noted that for this experiment the number of sample size is $T = 10000$. Fig. 1, shows the histogram of the generated data as well as the estimated density by using only global measuring functions and the estimated density by jointly using global and local measuring functions. We can see that by only including global measuring functions we are not able to capture the local behavior of the true underlying PDF and thus we achieve sub-optimal estimation performance in terms of matching with the histogram.

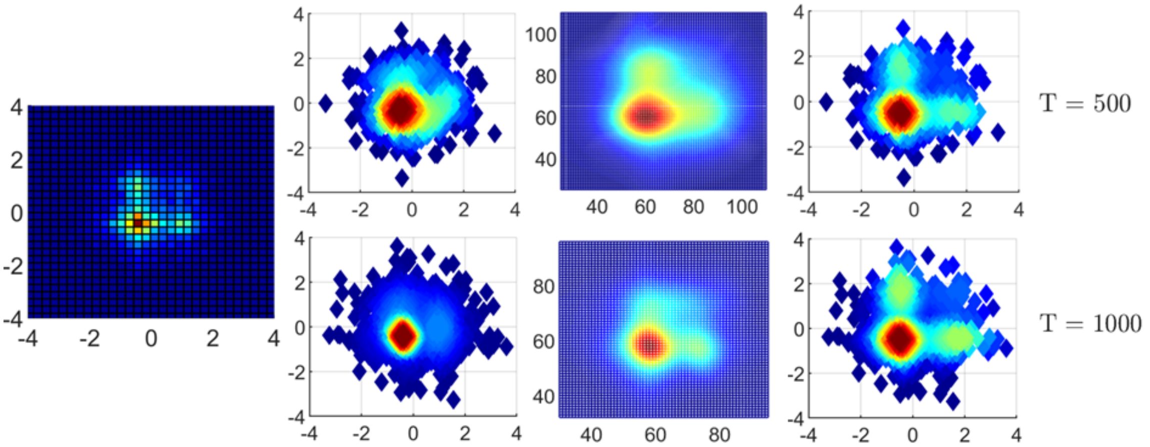

Fig. 2. Comparison in terms of matching with the histogram when $T = 500$ and $T = 1000$ represented in the first and second row, respectively. First column–histogram of the generated data, second column–parametric method (GMM), third column–non-parametric method (A-KDE), fourth column–semi-parametric method (M-EMK).

Furthermore, we show the effectiveness of our approach by comparing M-EMK with A-KDE as well as GMM in terms of matching with the histogram of the generated data for different sample sizes. As we can see in Fig. 2, when the number of sample size is $T = 1000$, M-EMK performs similarly to GMM and A-KDE in terms of matching with the histogram. However, it is worth mentioning that when $T = 500$, GMM and A-KDE are not able to capture the details of the shape of all three peaks while, on the other hand, M-EMK is able to effectively estimate the three highest peaks.

IV-B. Computational Benefits

In addition to the estimation capability, we demonstrate the computational efficiency of M-EMK when compared to A-KDE, GMM, and M-EMK by using only global constraints in terms of the CPU time. From Fig. 3, we can see that M-EMK provides the best performance in terms of CPU time when compared to A-KDE and GMM. As the number of sample size increases, A-KDE becomes computationally demanding making it impractical for high dimensional data applications. The low CPU time for M-EMK, has been observed due to the fast convergence rate introduced by the efficient integration technique. It is worth mentioning that adding local constraints to our approach does not have a significant impact to the average CPU time as the number of sample size increases and verifies the low computational complexity of our proposed method.

V. CONCLUSIONS

In this paper, we introduced the estimator multivariate entropy maximization with kernels (M-EMK), a new multivariate PDF estimation technique using the maximum entropy principle. By jointly using global and local constraints functions, M-EMK enjoys a high level of flexibility while providing a simple exponential form for PDFs. We show that M-EMK yields a very effective density estimation in terms of matching with the histogram while keeping the computational complexity low.

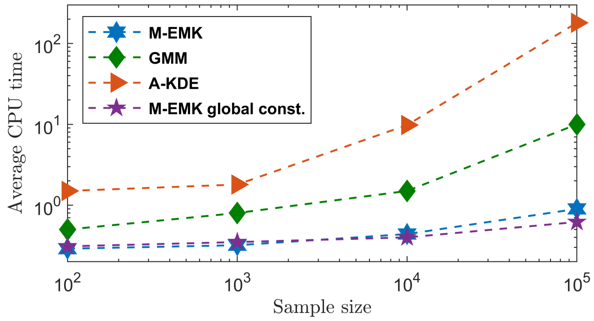

Fig. 3. Performance comparison in terms of CPU time for mixture of GGD data.

The success of the proposed method raises several interesting questions that can be explored in future work. From an algorithmic point of view, given that we have chosen the four global measuring functions we only use one local measuring function. Following the idea in [9], it would be of high interest to choose different number of local measuring functions by information-theoretic criterion. In addition, so far, the position of the local constraint is manually selected and the challenge is to automatize this procedure. Furthermore, we need to evaluate a metric for multidimensional approaches in order to improve our comparisons in terms of statistical distance measures, for instance, multivariate Kullback–Leibler divergence is a candidate to such comparison.

VI. REFERENCES

[1] H. L. V. Trees, K. L. Bell, and Z. Tian, *Detection, Estimation and Modulation Theory*, 2nd ed. Wiley, 2013, vol. 1.

[2] S. Theodoridis, *Machine Learning: A Bayesian and Optimization Perspective*. Academic Press, 2015.

[3] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[4] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[5] G. McLachlan and D. Peel, *Finite Mixture Models*, ser. Wiley Series in Probability and Statistics. Wiley, 2004.

[6] S. Y. Kung, *Kernel Methods and Machine Learning*. Cambridge University Press, 2014.

[7] J. L. Rojo-Álvarez, M. Martínez-Ramon, J. Muñoz-Mari, and G. Camps-Valls, *Digital Signal Processing with Kernel Methods*. Wiley - IEEE Press, 2018.

[8] B. Behmardi, R. Raich, and A. O. Hero, "Entropy estimation using the principle of maximum entropy," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2011, pp. 2008–2011.

[9] G. Fu, Z. Boukouvalas, and T. Adalı, "Density estimation by entropy maximization with kernels," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2015, pp. 1896–1900.

[10] X. Li and T. Adali, "Independent component analysis by entropy bound minimization," *IEEE Transactions on Signal Processing*, vol. 58, no. 10, pp. 5151–5164, Oct 2010.

[11] E. T. Jaynes, "Information Theory and Statistical Mechanics," *Phys. Rev.*, vol. 106, pp. 620–630, May 1957. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRev.106.620

[12] ——, "Information Theory and Statistical Mechanics. II," *Phys. Rev.*, vol. 108, pp. 171–190, Oct 1957. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRev.108.171

[13] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed., ser. Wiley Series in Telecommunications and Signal Processing. Wiley-Interscience, July 2006.

[14] D. O'Leary, *Scientific Computing with Case Studies*, ser. Other Titles in Applied Mathematics. Society for Industrial and Applied Mathematics, 2009.

[15] J. Dick, F. Y. Kuo, and I. H. Sloan, "High-dimensional integration: The quasi-Monte Carlo way," *Acta Numerica*, vol. 22, p. 133–288, 2013.

[16] H. Niederreiter, *Random Number Generation and Quasi-Monte Carlo Methods*. USA: Society for Industrial and Applied Mathematics, 1992.

[17] Z. I. Botev, J. F. Grotowski, and D. P. Kroese, "Kernel density estimation via diffusion," *Annals of Statistics*, vol. 38, no. 5, pp. 2916–2957, 2010.

[18] Z. Boukouvalas, S. Said, L. Bombrun, Y. Berthoumieu, and T. Adalı, "A new riemannian averaged fixed-point algorithm for MGGD parameter estimation," *IEEE Signal Processing Letters*, vol. 22, no. 12, pp. 2314–2318, 2015.