# DENSITY ESTIMATION BY ENTROPY MAXIMIZATION WITH KERNELS

*Geng-Shen Fu*[1]*, Zois Boukouvalas*[2]*, and Tülay Adalı*[1]

[1]University of Maryland, Baltimore County, Dept. of CSEE, Baltimore, MD 21250
[2]University of Maryland, Baltimore County, Dept. of Mathematics and Statistics, Baltimore, MD 21250

## ABSTRACT

The estimation of a probability density function is one of the most fundamental problems in statistics. The goal is achieving a desirable balance between flexibility while maintaining as simple a form as possible to allow for generalization, and efficient implementation. In this paper, we use the maximum entropy principle to achieve this goal and present a density estimator that is based on two types of approximation. We employ both global and local measuring functions, where Gaussian kernels are used as local measuring functions. The number of the Gaussian kernels is estimated by the minimum description length criterion, and the parameters are estimated by expectation maximization and a new probability difference measure. Experimental results show the flexibility and desirable performance of this new method.

***Index Terms—*** Probability density estimation, Maximum entropy distributions, Gaussian kernel.

## 1. INTRODUCTION

The estimation of a probability density function (PDF) is a common problem in a wide variety of fields ranging from physics to statistics. Within the field of machine learning, many estimation, detection, and classification problems require knowledge of the data's PDF either explicitly or implicitly, see e.g., [1]. Thus, effective characterization of the density is vital to the success of these machine learning approaches.

There are a number of density estimation methods. The non-parametric methods, such as histogram estimation, k-nearest neighbors (KNN), and kernel density estimation (KDE) [2, 3], can provide flexible density matching. However, these approaches are mostly computationally demanding, especially when sample size is large, and direct histogram estimation provides a discontinuous density function. In addition, the performance of non-parametric methods highly depend on the choice of parameters, such as the number of bins, number of samples in a neighborhood, and the bandwidth for histogram, KNN, and KDE, respectively. In [4], the authors propose KDE via diffusion (KDE-DF), which calculates the bandwidth automatically, and show the improvement in performance for a number of scenarios. Many of the other methods assume a parametric model, such as the generalized Gaussian distribution (GGD) [5], for the density. These parametric methods provide a simple form for the PDF and are computationally efficient, however the parametric form is usually very limited. For example, the GGD limits the PDFs to these that are symmetric and unimodal, and controls the shape through a single parameter. Semi-parametric methods combine the flexibility with the relatively simple density form. Gaussian mixture model (GMM) [3] has been widely used for semi-parametric density estimation. However, GMM is typically time consuming and since the kernel function is limited to only a single type, the tradeoff between flexibility and generalization ability needs to be carefully weighted when selecting the number of mixtures and their parameters.

In this paper, we present a density estimator with smooth characteristics based on the maximum entropy principle. We jointly use global and local measuring functions to provide flexible PDFs with as simple a form as possible. The global measuring functions provide constraints on the statistics of the PDF, such as the mean, variance, and higher-order statistics (HOS). However, methods using only global measuring functions [6–8] ignore local information of the PDF, additionally the use of higher-order polynomials can introduce stability issues. We use Gaussian kernels, which do not have stability issues due to their asymptotic convergence, as local measuring functions to provide local information. The number of Gaussian kernels is chosen using the information-theoretic criterion, the minimum description length (MDL) [9]. We call the new estimator entropy maximization with kernels (EMK). Simulation results show the flexibility of EMK as well as its effectiveness in approximating a range of distributions.

This work was supported by the NSF grant NSF-CIF 1117056.

## 2. MAXIMUM ENTROPY DISTRIBUTIONS

We seek to estimate the PDF, $p(x)$, given observations $x(t) \in \mathbb{R}, t = 1, \dots T$. The seminal papers [10, 11] introduce the principle of maximum entropy, and note that: "the probability distribution which best represents the current state of knowledge is the one with largest entropy subject to precisely stated prior data." The maximum entropy density, subject to known constraints, can be written as the following optimization problem [12]:

$$\max_{p(x)} H(p(x)) = -\int_{-\infty}^{\infty} p(x) \log p(x)\, dx$$
$$\text{s.t.} \int_{-\infty}^{\infty} r_i(x) p(x)\, dx = \alpha_i, \text{ for } i = 0, \dots, M, \quad (1)$$

where $r_i(x)$ and $\alpha_i = \sum_{t=1}^{T} r_i(x_t)/T$ for $i = 0, \dots, M$ are the measuring functions and their corresponding sample averages, respectively. From now on, we simplify the notation for clarity if it can be inferred from context. From (1), $p$ is a density on support set $\mathbb{R}$ meeting $M+1$ constraints $\int r_i p = \alpha_i$. We note that the first constraint must be $\int p = 1$, equivalently $r_0 = 1$ and $\alpha_0 = 1$, to ensure that $p$ is a valid PDF. The optimization problem in (1) can be written as a Lagrangian function:

$$\mathcal{L}(p) = -\int p \log p + \sum_{i=0}^{M} \lambda_i \int (r_i - \alpha_i)\, p, \quad (2)$$

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

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

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 $M+1$ Lagrange multipliers. We can solve the problem by a Newton iteration method,

$$\boldsymbol{\lambda}^{(n+1)} = \boldsymbol{\lambda}^{(n)} - \mathbf{J}^{-1} E_{p^{(n)}}\{\mathbf{r} - \boldsymbol{\alpha}\}, \quad (4)$$

where $p^{(n)}$ is the estimated PDF for the $n$th iteration, and $\mathbf{r} = [r_0, \dots, r_M]^\top$, $\boldsymbol{\lambda} = [\lambda_0, \dots, \lambda_M]^\top \in \mathbb{R}^{M+1}$, and $\boldsymbol{\alpha} = [\alpha_0, \dots, \alpha_M]^\top$ denote the vector of measuring functions, Lagrange multipliers, and sample averages, respectively. The $i$th entry of $E_{p^{(n)}}\{\mathbf{r} - \boldsymbol{\alpha}\}$ is given by

$$E_{p^{(n)}}\{r_i - \alpha_i\} = \int (r_i - \alpha_i)\, p^{(n)}, \quad (5)$$

and the $(i,j)$th entry of $\mathbf{J}$ is given by

$$\mathbf{J}_{ij} = \frac{\partial \int (r_i - \alpha_i)\, p^{(n)}}{\partial \lambda_j} = \int r_i r_j p^{(n)}. \quad (6)$$

## 3. CHOICE OF MEASURING FUNCTIONS

The choice of measuring functions in (1) is critical since they fully determine the maximum entropy distribution. Theoretically, we can model any PDF using maximum entropy distributions if the measuring functions are selected appropriately. To this end, the authors of [7] propose to use a dictionary of polynomials and trigonometric basis functions. However, the use of higher-order polynomials can cause stability issues, and the use of a trigonometric basis can result in noisy estimates. Also, a large number of measuring functions is usually required to accurately model a PDF when polynomials and trigonometric functions are used. Similar to polynomials or trigonometric basis, functions can be approximated by using a radial basis [13], such as Gaussian kernels.

In this paper, we jointly use global and local measuring functions to provide flexible density estimation while keeping the complexity low. As in [6], we use $1$, $x$, $x^2$, $x/(1+x^2)$ as the global measuring functions, since it is computationally efficient to use them and they provide desirable performance for a wide range of distributions, in particular when the true PDF is not complicated. These global measuring functions provide information on the PDF's overall statistics, such as the mean, variance, and certain HOS. For local measuring functions, we use the following Gaussian kernel,

$$r(x) = \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right),$$

where the normalization scalar is ignored since it can be combined with the corresponding Lagrange multiplier. The use of Gaussian kernels provides localized information about the PDF. We note that other kernels can also be used as measuring functions. For example, we can use uniform functions as the measuring functions with the corresponding constraints on the frequency of the observations in the interval, i.e., as $\int_a^b p = T_{ab}/T$. However, the use of uniform measuring functions will lead to discontinuous PDF, and hence are not desirable.

## 4. CHOICE OF NUMBER OF CONSTRAINTS

Given that we have chosen the four global measuring functions, the number of local measuring functions is chosen by information-theoretic criterion, MDL [9, 14]

$$\mathcal{J}(M) = T\mathcal{L}(\mathbf{x}|M) + \varepsilon\eta(\boldsymbol{\theta}_M)\log T, \quad (7)$$

where $\boldsymbol{\theta}_M = \{\lambda_0, \dots, \lambda_3, \lambda_4, \mu_4, \sigma_4, \dots, \lambda_M, \mu_M, \sigma_M\}$, $\mathcal{L}(\mathbf{x}|M) = -\sum_{t=1}^{T} \log p(x(t)|\boldsymbol{\theta}_M)/T$ is the normalized negative log-likelihood of $\mathbf{x}$ given that the number of measuring functions is $M$, $\eta(\boldsymbol{\theta}_M) = 3M - 5$ indicates the number of free parameters, and $\varepsilon$ is $0.5$. The number of measuring functions is chosen to be the one corresponding to the minimum of $\mathcal{J}(M)$, which jointly simplifies the description of the data and the model.

## 5. ESTIMATION OF PARAMETERS FOR GAUSSIAN KERNELS

For each Gaussian kernel, there are two parameters, $\mu$ and $\sigma^2$, that need to be estimated. A straightforward way to estimate these parameters is by finding the greatest deviation between the estimate, $p$, and the true PDF, $p_*$, and placing a kernel there. Since the true PDF is not accessible, it is replaced by the PDF, $\bar{p}$, which is estimated using histogram. In order to match the true PDF well, especially at the parts with high probability, we propose the following difference of probability measure,

$$d_p = \left| (p - \bar{p})\bar{p} \log \frac{\bar{p}}{p} \right|, \tag{8}$$

where the latter term is used as the weight, which emphasizes the parts of the distribution that have higher probability, similar to the Kullback-Leibler (KL) divergence [12]. Then, the two parameters, $\mu$ and $\sigma^2$, are given by the mean and span of the peak with largest area of $d_p$.

The previous approach is affected by the performance of histogram and the robustness of the estimation. In order to improve the performance, we propose to estimate $N$ candidate Gaussian kernels each time and estimate the PDF corresponding to these $N$ candidates, respectively. Then, the parameters, $\mu$ and $\sigma^2$, are chosen among the $N$ candidates using the maximum likelihood principle. In addition, we replace the histogram by KDE-DF to increase accuracy, and $N_1$ of the candidates are selected by the peaks of (8) that have the largest areas. The parameters for the rest of the candidates are estimated through expectation maximization (EM). We assume that data is generated by the mixture of a maximum entropy distribution, $p_0 = p$, and $Q$ Gaussians, $p_q = \mathcal{N}(\mu_q, \sigma_q^2)$, for $1 \leq q \leq Q$. For the expectation step, the posterior probability of the $n$th iteration is given by

$$U_{q,t}^{(n)} = \frac{\tau_q^{(n)} p_q^{(n)}(x(t))}{\sum_{q=0}^{Q} \tau_q^{(n)} p_q^{(n)}(x(t))}.$$

For maximization step, the prior probability and parameters for the Gaussian mixture are given by

$$\tau_q^{(n+1)} = \frac{1}{T} \sum_{t=1}^{T} U_{q,t}^{(n)}, \text{ for } q = 0, \ldots, Q,$$

$$\mu_q^{(n+1)} = \frac{\sum_{t=1}^{T} U_{q,t}^{(n)} x(t)}{\sum_{t=1}^{T} U_{q,t}^{(n)}}, \text{ for } q = 1, \ldots, Q,$$

$$\sigma_q^{2\,(n+1)} = \frac{\sum_{t=1}^{T} U_{q,t}^{(n)} (x(t) - \mu_q^{(n+1)})^2}{\sum_{t=1}^{T} U_{q,t}^{(n)}}, \text{ for } q = 1, \ldots, Q.$$

We select the $\mu_q$ and $\sigma_q^2$ corresponding to the $N - N_1$ largest $\tau_q$ as the parameters for the rest of the candidates.

## 6. TWO IMPLEMENTATIONS FOR EMK

We use the four global measuring functions given in Section 3 and estimate the parameters for the local measuring functions, the Gaussian kernels, as introduced in Section 5. Using (5) and (6), we calculate $E\{\mathbf{r} - \boldsymbol{\alpha}\}$ and the Jacobian matrix, $\mathbf{J}$, respectively. Then, the Lagrangian multipliers, $\boldsymbol{\lambda}$, are found using the Newton iteration as in (4). We keep adding Gaussian kernels as local measuring functions until the MDL cost in (7) stops decreasing. The parameters $\varepsilon$ and $Q$ are empirically chosen to be $0.1$ and 3, respectively.

Here, we propose two versions of the new method, EMK and EMK-lite (EMK-L). In EMK-L, we choose $N$ to be 1 and use histogram to estimate $\bar{p}$ in (8). For EMK, $N$ and $N_1$ are chosen to be 4 and 2, respectively, and the parameters for these 4 candidates are estimated by PDF-DF and EM as described in Section 5. The pseudocode is given in Algorithm 1.

**Data**: $\mathbf{x} \in \mathbb{R}^T$
**Result**: $p(x)$
$r_0 = 1, r_1 = x, r_2 = x^2$, and $r_3 = x/(1 + x^2)$;
Estimate $\boldsymbol{\lambda}$ using Newton iteration in (4);
Estimate $\bar{p}$;
Calculate $\mathcal{J}(3)$ using (7);
$\boldsymbol{\lambda}_{\text{opt}} = \boldsymbol{\lambda}$, $\mathcal{J}_{\text{opt}} = \infty$, and $M = 3$;
**while** $\mathcal{J}(M) < \mathcal{J}_{opt}$ **do**
  $\boldsymbol{\lambda}_{\text{opt}} = \boldsymbol{\lambda}$, $\mathcal{J}_{\text{opt}} = \mathcal{J}(M)$, and $M = M + 1$;
  Get current estimation, $p$, from (3) using $\boldsymbol{\lambda}_{\text{opt}}$;
  Calculate $d_p$ using (8);
  Estimate $N$ candidates, $\{\mu, \sigma\}$, by $d_p$ and EM;
  Estimate $\boldsymbol{\lambda}$ using (4) and choose $\{\mu_M, \sigma_M\}$ to be the one giving min $\mathcal{L}(\mathbf{x})$ among candidates;
  Calculate $\mathcal{J}(M)$ using (7);
**end**
$M = M - 1$ ;
The estimation, $p$, is given by (3) using $\boldsymbol{\theta} = \{\boldsymbol{\lambda}_{\text{opt}}, \mu_4, \sigma_4, \ldots, \mu_M, \sigma_M\}$ ;

**Algorithm 1:** EMK

By jointly using global and local measuring functions, EMK provides a very simple exponential form of the PDF, which allows for easy derivation in many applications. In addition, EMK enjoys the flexibility property of non-parametric methods, which is shown in the next section.

## 7. EXPERIMENTAL RESULTS

In this section, we study the performance of EMK by using simulated and natural data. We show the flexibility of EMK by comparing its performance to that of other widely used and competitive density estimation methods using simulations. In addition, we use natural data to

demonstrate the effectiveness of EMK by comparing its result with that of the given histogram visually.

Experiment 1: In order to show the flexibility of the new method, we show the performance of EMK using a mixture of GGD data. The PDF of GGD is given by [5]

$$p_{\text{GGD}}(x \mid \beta, \mu, \sigma) = \frac{\beta}{2^{\frac{1}{2\beta}}\Gamma(\frac{1}{2\beta})\sigma} \exp\left(-\frac{(x-\mu)^{2\beta}}{2\sigma^{2\beta}}\right).$$

Thus, the PDF of mixture of GGD is given by

$$p_*(x) = \sum_{i=1}^{N} \pi_i p_{\text{GGD}}(x \mid \beta_i, \mu_i, \sigma_i).$$

We choose the number of mixtures $N \in \{1, 2, 3, 4, 5\}$ where each choice is equally likely and weights $\pi_i \sim \mathcal{U}(0, 1)$ such that $\sum_{i=1}^{N} \pi_i = 1$. The shape parameter is generated by $\beta_i = 2^a$, where $a \sim \mathcal{U}(-2, 2)$, i.e., $p_{\text{GGD}}$ is equally likely to be super or sub-Gaussian. The mean is generated by $\mu_n \sim \mathcal{U}(-6, 6)$. The data is normalized to be zero mean and unit variance after generation. We generate 25 PDFs, and perform 8 independent trials for each PDF. Hence, the results are the average of 200 trials. The performance is measured in terms of the KL divergence [12], $\int_{-\infty}^{\infty} p_* \log p_*/p$, between the true PDF and the estimates. Fig. 1 demonstrates the diversity of possible true densities that can be estimated. In Fig. 2, we compare the performance of our method to several of the most popular PDF estimation methods in terms of the KL divergence and the CPU time. The bin width for histogram (Hist), number of samples for KNN, and the maximum number of kernels for GMM are chosen to be $3.49T^{-1/3}$, 10, and 12, respectively, where $T$ is the sample size. We can see that EMK provides the best performance among the methods we compare with in terms of KL divergence showing the flexibility property of EMK. Additionally, for CPU time, EMK is more efficient than the methods with comparable KL divergence.

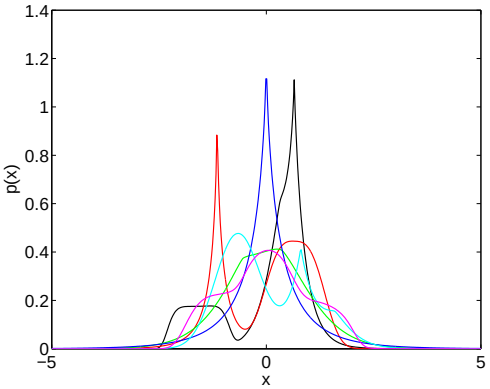


**Fig. 1**. 6 randomly selected true PDFs.

Experiment 2: We show the effectiveness of EMK using two $512 \times 512$ standard test images [15]. As we can see from Fig. 3, EMK provides very desirable performance in terms of matching with the histogram. In addition, EMK provides flexible density matching through simple parametric forms using only 17—automatically chosen—Gaussian kernels in each image.

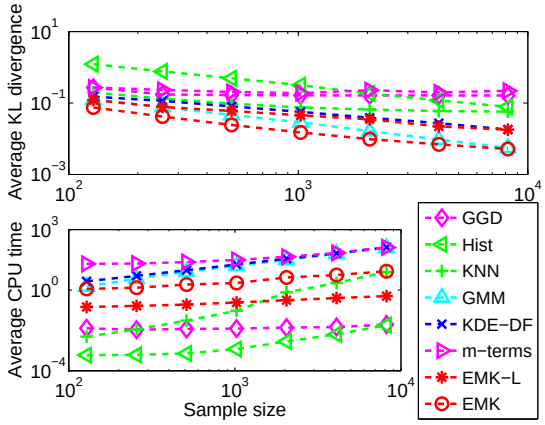


**Fig. 2**. Performance comparison in terms of KL divergence and CPU time for mixture of GGD data.

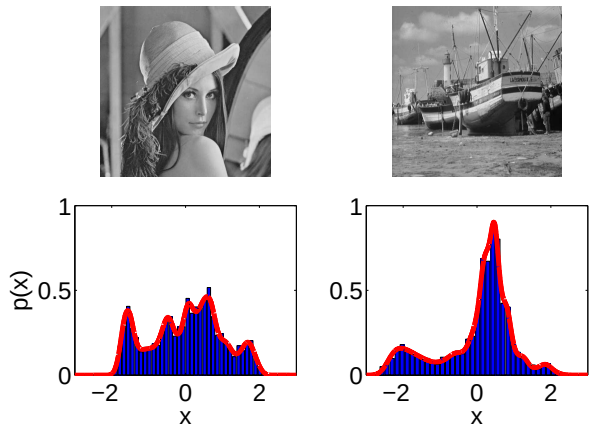


**Fig. 3**. Two test images and their estimated PDF by EMK, red lines, and histogram, blue bars.

## 8. DISCUSSIONS

In this paper, we propose a new PDF estimation technique, EMK, using the principle of maximum entropy with Gaussian kernels. By jointly using global and local measuring functions, EMK enjoys a high level of flexibility while providing a simple exponential form for the overall PDF. We show by experiments that EMK yields the best performance among others in terms of KL divergence. Although we only consider the univariate case in this paper, the new method can be extended to multivariate case in a fairly straightforward manner.

## 9. REFERENCES


[1] T. Adalı, M. Anderson, and G.-S. Fu, "Diversity in independent component and vector analyses: Identifiability, algorithms, and applications in medical imaging," *Signal Processing Magazine, IEEE*, vol. 31, no. 3, pp. 18–33, May 2014.

[2] 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.

[3] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[4] Z. I. Botev, J. F. Grotowski, and D. P. Kroese, "Kernel density estimation via diffusion," *Annals of Statistics*, 2010.

[5] S. Nadarajah, "A generalized normal distribution," *Journal of Applied Statistics*, vol. 32, no. 7, pp. 685–694, 2005.

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

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

[8] R. V. Abramov, "An improved algorithm for the multidimensional moment-constrained maximum entropy problem," *Journal of Computational Physics*, vol. 226, no. 1, pp. 621–644, 2007.

[9] J. Rissanen, "Modeling by shortest data description." *Automatica*, vol. 14, pp. 465–471, 1978.

[10] E. T. Jaynes, "Information theory and statistical mechanics," *Phys. Rev.*, vol. 106, pp. 620–630, May 1957.

[11] ——, "Information theory and statistical mechanics. II," *Physical Review Online Archive*, vol. 108, no. 2, pp. 171–190, Oct. 1957.

[12] T. M. Cover and J. A. Thomas, *Elements of information theory*. New York, NY, USA: Wiley-Interscience, 1991.

[13] M. D. Buhmann, *Radial basis functions: theory and implementations*. New York, NY, USA: Cambridge University Press, 2003.

[14] W. Wang, T. Adalı, S. Y. Kung, and Z. Szabo, "Quantification and Segmentation of Brain Tissues from MR Images: A Probabilistic Neural Network Approach," *IEEE Transactions on Image Processing*, vol. 7, no. 8, pp. 1165–1181, 1998.

[15] A. G. Weber, "The USC-SIPI Image Database," University of Southern California, Signal and Image Processing Institute, Department of Electrical Engineering, Los Angeles, CA 90089-2564 USA, 3740 McClintock Ave, Tech. Rep., Oct. 1997.