Advances in Reliability, Safety and Security, Part 9 – Kołowrocki, Dąbrowska (eds.)
© 2024 Polish Safety and Reliability Association, Gdynia, ISBN 978-83-68136-21-0 (printed), ISBN 978-83-68136-08-1 (electronic)

Advances in Reliability, Safety and Security

ESREL 2024
Monograph Book Series

Neural Network Based Probability Density Function Identification

Guillaume Levillain[a], Pierre Beaurepaire[a],Vincent Barra[b]

[a]*Université Clermont Auvergne, Clermont Auvergne INP, CNRS, Institut Pascal, F-63000 Clermont-Ferrand, France*
[b]*Université Clermont-Auvergne, CNRS, Mines de Saint-Étienne, Clermont-Auvergne-INP, LIMOS, 63000 Clermont-Ferrand, France*

Abstract

This paper discusses the identification of probability distributions from samples. We propose a new method based on the principle of maximum entropy and more specifically maximum entropy probability density functions. This method requires prior knowledge or arbitrary selection of some constraint functions, and we identify those with neural networks. Maximum entropy formalism can then be formulated as a special layer in a neural network. This neural network can be trained on samples, using a custom entropy-based loss function. We provide examples of the estimation of non-trivial probability distributions to demonstrate the capability of our method.

Keywords: distribution estimation, neural networks, maximum entropy

1. Introduction

Probabilistic Distributions (PD) are widely used in science and engineering; their identification remains a challenging task. From this, the estimation of Probability Density Functions (PDF) is way to completely characterize a PD. It is however not the only function describing a PD as Cumulated Distribution Functions (CDF) and Characteristic Functions (CF) can also serve this purpose. With the description of a PD, we can curate data and interpret it (e.g., with uncertainty quantification, statistical inference, etc). During the last decades, the acquisition of data evolved with the development, increasing availability and integration of computer. This evolution impacted every step of the lifecycle of data: automated acquisition instead of an operator’s read-out, vast amounts of storage only possible with hard drives or tape recorders and easy exchange of data through the Internet. Overall, this means data is more accessible than ever, and it alters our relation with it. However, in that regard, methods of PD identification did not follow the same evolution in that time. When those methods were developed for small samples, noisy from their limited size, most models seemed to provide a fair fit. Meanwhile, large datasets suffer less from noise, while capturing more subtleties of their generating PDs. Hence, the usual models, often characterized only by a small number of parameters, fail to estimate the complex PDs by their own lack of expressiveness. Considering this, we are aiming for more versatile and expressive estimates, in order to match the detailed patterns offered from large samples. Therefore, we propose to employ neural networks, which were proven useful in a vast variety of applications, as part of our model in the framework of Maximum Entropy Probability Density Functions (ME-PDF). Differential entropy, described in Shannon’s famous article (Shannon, 1948) giving birth to information theory, is a measure of the quantity of information of a PD. It can be interpreted as the average “surprise” provided by the PD. If provided multiple estimations of a PD, differential entropy can discriminate the best. The most parsimonious choice is keeping the distribution with the highest entropy, because maximizing surprise implies minimizing overconfidence. This yields a method to choose between models. An extent of this choice for all models of a PD is the Principle of Maximum Entropy (PME) defined by (Jaynes, 1957). It states the best estimation of a PD is the model that, under constraints from prior data, maximizes entropy. From this principle, we can derive a universal ME-PDF expression:

$$p(x) = exp\left(-\lambda_0 - \sum_{i=0}^{n} \lambda_i \, g_i(x)\right) \tag{1}$$

where $\{\lambda_i\}_n$ can be viewed as parameters of the model, and $\{g_i(x)\}_n$ as functions probing the global behaviour, or constraints, of the PD. In the literature, (Novi Inverardi and Tagliani, 2003) provide an example of ME-PDF implementation. Fractional moments probability density functions use fractional moments as constraints:

$g_i(x) = x^{\alpha_i}, \alpha_i \in \mathbb{Q}^+$.

This choice is justified by analytical results: Lin's theorem (Lin, 1992) states that an infinity of those fractional moments under a given value completely defines a random variable. Then, an optimization over the choice of the best moments, and how many are used, is applied. Finally, employing the BIC criterion avoids over-parametrization of the model from the use of small datasets, which are prone to over-fitting.

We also take interest in Sliced-Normal distributions from (Crespo et al., 2019), which do not explicitly use ME-PDF. First, their method consists in augmenting the dimensionality of a multivariate sample by calculating monomials of its component, projecting the problem in a higher dimensionality latent space. After that, the data projected in the latent space is supposed normal, and a multivariate normal distribution is fitted. An additional specificity is considered, as its tails are arbitrary truncated. In effect, authors only identify the mean vector and the covariance matrix (more exactly the Cholesky decomposition of the inverse of the covariance matrix) in the latent space. From there, their density only has to be re-normalised, as it no longer integrates to one from the tail trimming. The framework of ME-PDF is applicable for the sliced-normal distributions even though (Crespo et al., 2019) did not use it. In this context, the constraint could be identified as a large polynomial combining the monomials. The fitting is based on log-likelihood maximization which can be similar as using an entropy-based objective function, as shown in (Novi Inverardi and Tagliani, 2003).

Several works in the literature discuss the identification of PD using some concepts presented above. (Di Paola and Pinnola, 2012) and (Dai et al., 2018) use complex fractional moments to reconstruct the CF of the PD with the Mellin transform. Meanwhile, (Magdon-Ismail and Atiya, 2002) use neural networks to approximate the CDF from the empirical distribution function. Finally, another work from (Crespo et al., 2018) fits a histogram-shaped PDF by matching moments with maximum likelihood or minimum Kullback-Leibler divergence.

In regard to ME-PDF framework, early works based their choice of constraints from analytical result with fractional moments PDF, whereas more recent works are similar to a ME-PDF approach with a polynomial constraint identification. We inscribe ourselves in a continuity by using neural networks for constraints identification in ME-PDF problems.

2. Method formulation

2.1. Maximum entropy probability density function

Shannon introduced the concept of differential entropy, which is applicable to continuous distributions. Differential entropy is not an absolute measure of quantity of information, as it is not invariant by change of variable. It is however sufficient for comparison in optimization problems as long as the random variable is not changed. The definition of differential entropy is:

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

where p is a continuous PDF supported on $\mathbb{X}$. From that definition, the following optimization problem identifies the ideal density maximizing entropy:

$$\arg\max_{p} H[p]$$

$$\text{subject to} \quad \int_{\mathbb{X}} p(x)dx = 1 \tag{2}$$

$$\mathbb{E}_p[g_i(x)] = \mu_i, \forall i \in [\![1, n]\!] \tag{3}$$

First, the optimization is done not over a scalar or a vector but over a function (in that case p). The first optimization constraint (see (2)) is obvious: in order to be a PDF, the function p has to integrate to one. The other constraints (see (3)) are here to represent large scale properties of the distribution, and reflect the patterns contained in the data. Jaynes derived this method from statistical thermodynamics, where macroscopic quantities measured as μ_i (e.g., temperature, pressure, etc) can reveal microscopic behaviours (e.g., the kinetic energy distribution of gas molecules) and reversely. ME-PDF formalism is a generalization from physics, and in the process constraints are loosing their interpretability. Nonetheless, a certain number of usual PD (e.g., the normal distribution, the exponential distribution, etc) can be formulated as ME-PDF.

The Lagrangian associated with this optimization problem is:

$$J[p] = \int_{\mathbb{X}} p(x)\, log\, p\,(x) dx - \eta_0 \left(\int_{\mathbb{X}} p(x) dx - 1 \right) - \sum_{i=1}^{n} \lambda_i \left(\int_{\mathbb{X}} g_i(x) p(x) dx - \mu_i \right).$$

The extremum of the optimization problem is reached when the functional derivative of the Lagrangian equals zero.

$$\frac{\delta J}{\delta p}[p] = 0 \Leftrightarrow log\, p\,(x) + 1 - \eta_0 - \sum_{i=1}^{n} \lambda_i g_i(x) = 0.$$

It can be proven this extremum is a maximum. From this, after setting $\lambda_0 \doteq 1 - \eta_0$, the optimal expression of p is given by (1). In this equation the $\{\lambda_i\}_n$ are parameters of the density function. However, our method identifies not only the $\{\lambda_i\}_n$, but the $\{g_i\}_n$ as well, instead of setting them arbitrarily. The goal is to estimate the best constraint functions, as the information from the data is partially contained in them.

2.2. Probability density functions defined by neural networks

2.2.1. Introducing neural networks

Neural networks (NN) are a learning model based on multiple layers of neurons, where the neurons from layer n ($\boldsymbol{\ell}_n$) are computed as

$$\boldsymbol{\ell}_n = \phi(\boldsymbol{W}_{n-1} \boldsymbol{\ell}_{n-1} + \boldsymbol{b}_n),$$

where $\boldsymbol{W}_{n-1}$ (respectively $\boldsymbol{b}_n$) are the weights (respectively bias) and are the NN parameters, and ϕ is a non-linear activation function. The layers are applied successively beginning with the input vector and the last being the output, with all the intermediary layers we referred to as hidden layers. The activation functions are very usually chosen from classical results in the literature. NNs are learnt by updating through iteration the parameters (typically with gradient descent methods) to improve the quality of the estimation, measured by a scalar function (usually called loss function) comparing the estimation and the true value selected from given data.

We use here a very classical implementation of an interpolator NN,

$$\boldsymbol{f}: \mathbb{R}^l \mapsto \mathbb{R}^m, (l, m) \in \mathbb{N}^{+2}.$$

The input has the dimension of the data, and the output has a dimension equal to the number of constraints.

2.2.2. Identification of the constraint functions in neural networks

A NN is used for the formulation of the ME-PDF, and it is formulated as discussed below. From the vectorial output of the NN, each of its components can be identified as a constraint function

$$f_i(\boldsymbol{x}) = g_i(\boldsymbol{x}).$$

The set of constraint functions $\{g_i\}_n$ is the output of the NN. Then, another special layer is applied, that we call the Maximum Entropy (ME) layer. The weights are given as

$$\boldsymbol{W}_{1i} = \lambda_i, \forall i \in [\![1, n]\!],$$

and the bias is

$$b_1 = \lambda_0.\ \lambda_0$$

is however treated very differently from a classical bias in NNs, as it ensures the PDF integrates to one, and is analytically computed from all the other parameters (the ME layer and the NN layers). Finally, the activation function of the ME layer is identified as

$\phi(\xi) = exp(-\xi)$.

The discussion above shows that (1) can be formulated in the framework of NNs, using a NN to model the constraint functions, and ME layer to express a ME-PDF.

2.2.3. Formulation of the loss function

Once the problem is formulated in NNs framework, we aim to find an objective function to learn all the parameters. A naive approach would consist of maximizing the entropy of the PDF, which would fail. The PD maximizing the entropy for a given domain is the uniform distribution, as it is the most parsimonious by default, and using the entropy to learn the parameters would ignore all the data and converge to this PD. However, Novi Inverardi and Tagliani (2003) proposed as an objective function the Kullback-Leibler (KL) divergence between the true PDF of the PD generating the data p and the estimated density $\hat{p}$ from a given model. This quantity can be viewed as the average added "surprise" we get sampling from a PD, expecting it to be defined by $\hat{p}$ while it is actually defined by p. The lower the KL divergence, the "closer" (in the sense of reducing surprise) the PDs; a zero value implies that both distributions are identical.

The KL divergence can be defined as:

$$\begin{aligned} D_{KL}(p||\hat{p}) &= \int_{\mathbb{X}} p(x)\, log \frac{p(x)}{\hat{p}(x)} dx \\ &= \int_{\mathbb{X}} p(x)\, log\, p\,(x) dx - \int_{\mathbb{X}} p(x)\, log\, \hat{p}\,(x) dx \\ &= -H[p] - \mathbb{E}_p[log\, \hat{p}\,(X)] \end{aligned} \tag{4}$$

The left-hand term of (4) is independent of the model, as it only involves on the true PD, and can be ignored in the optimization. The right-hand term of (4), i.e., the expected value of $log\, \hat{p}\,(X)$ (while X follows a PD defined by p) can not be calculated as it requires the knowledge of p. However, samples of the distribution $\{x_j\}_M$ are available, and this term can be approximated as:

$$\mathbb{E}_p[log\, \hat{p}\,(X)] \approx \frac{1}{M}\sum_{j=1}^{M} log\, \hat{p}\,(x_j) \tag{5}$$

The ME-PDF, i.e., (1), can be injected into (5), which yields:

$$\mathbb{E}_p[log\, \hat{p}\,(X)] \approx -\frac{1}{M}\sum_{j=1}^{M}\left(\lambda_0 + \sum_{i=1}^{n}\lambda_i g_i(x_j)\right) \tag{6}$$

From this we can conclude that we can minimize $D_{KL}(p||\hat{p})$ by maximizing (6) (i.e., minimizing its opposite).

It is important to note that we use this divergence as a loss function and use gradient descent to update all parameters, excepted λ_0. We have to calculate this last one from the other parameters to ensure $\hat{p}$ integrates to one on $\mathbb{X}$. This is discussed in more details in the example section.

3. Examples

We use two examples to demonstrate the capability of our method, one univariate, one multivariate, and both are non-trivial. We limited ourselves to 10^3 data points for each example, as it is a usual sample size in the literature. It would be possible to use bigger dataset as the training of NNs allows to handle the data by batches, and to not exceed the working memory of the computer, but this remains outside the scope of this paper.

We implement a quantification of the quality of the estimation, as the literature usually only present plots of estimated PDF or of the value of the optimized function. This allows objective comparison with past and future works. This quantity requires needing moderate computational efforts, even in high dimension. In that sense, we decided to use here the Maximum Mean Discrepancy (MMD) squared, which is suitable to compare PDFs, and can be numerically approximated (Gretton et al., 2012). It requires the choice of a kernel function, and we use the Gaussian RBF (with $\sigma = 0.5$). The use of other kernels (other classical kernels, or custom created ones (Shawe-Taylor and Cristianini, 2004)) or the use the Dvoretzky-Kiefer-Wolfowitz inequality (Naaman, 2021) as other metrics is left for future work.

3.1. Univariate example

For the first example, we try to estimate the density of the random variable

$X = U^2 - V^3$,

where U and V are independent random variables uniformly distributed between zero and one. (Crespo et al., 2018) proposed this example. In the implementation of the NN, we use here a model of 3 hidden layers of 8/4/1 neurons respectively, with ReLU activation function. With this NN, 57 parameters characterize the constraint function. The total number of parameters might appear high, as the frequently used PDs (normal, Weibull, gamma, etc) or the methods described in the literature (Novi Inverardi and Tagliani, 2003; Crespo et al., 2019) involves a reduced number of parameters. However, the ratio over the dataset size and dimension (1000 samples of univariate data) is still perfectly acceptable. The NN has only a single neuron as an output, which implies only one constraint function is considered in (1). It was empirically observed that this leads to satisfactory results, and considering additional constraints did not improve the results.

In order to enforce the integration to one, we need to approximate the integral between the minimum and maximum sample values, using the rectangle quadrature rule. For univariate distributions, this approach is totally feasible, but the curse of dimensionality forbids us to extend it to multivariate case.

Figure 1 shows that the results are imperfect, but are acceptable considering the complexity of the example. The reference PDF has a peak and is continuous, but its derivative is discontinuous at that peak. Our method captures relatively well that peak. Models that try to get a better fit usually require many more parameters; polynomials would not be able to capture such a shape.

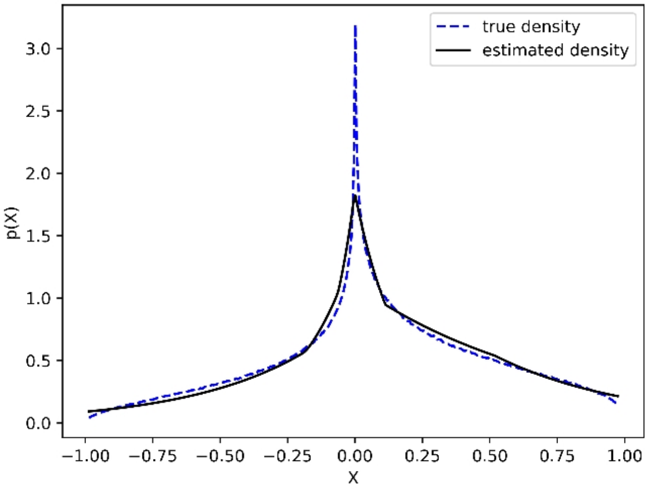

Fig 1. Univariate PDF estimations

When generating multiple fits, there are some variabilities in the results as shown in Figure 2. In order to grasp the variability of the proposed method, Figure 2 represents 110 runs of the algorithm, each one with a different training sample. This variability comes partially from the regeneration of a sample at each run. With a new 1000 sample points for each run, we can expect some sample to be less representative of the distribution, which causes worse PDF identifications. The random initialization of the weights may as well have an influence. This last factor should not matter, as the NN is supposed to converge to the same result, but the diversity of curves for similar values of the loss function seems to point that the loss function is very flat when close to the minimum. However, all the estimated PDFs globally respect the shape of the PD, with only some difficulties on the spike and on the extremities. Here, for all estimations, we have $MMD^2 < 10^{-4}$.

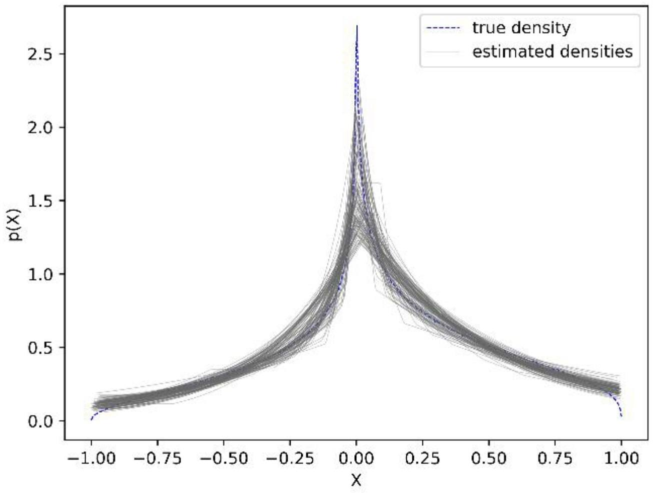

Fig. 2. Bundle of 110 PDF estimations.

Further works intend to study the repeatability, i.e., with a fixed learning sample and different neural networks parameters initialization, and the reproducibility, i.e., with a different sample and neural networks parameters initialization.

3.2. Multivariate example

The second example involves two random variables with a complex joint PDF. (Colbert et al., 2020) proposed a similar example. The NN is here six layers deep with respectively 40/30/20/10/5/1 neurons, activated by the "swish" function. This NN involves then 2241 parameters in total. Compared to the univariate problem, the problem's estimation domain is $\mathbb{R}^2$ instead of $\mathbb{R}$, and we could have expected the need to square the number of parameters to get satisfying results. The number of parameters is significantly lower than that, which seems to indicate that the method is able to scale to higher number of dimensions without an explosion of the number of parameters. The integration here is slightly more complex: we defined the bounding box of the samples, to then use a Sobol' design of experiment scaled on that domain to estimate the integral by Quasi Monte Carlo.

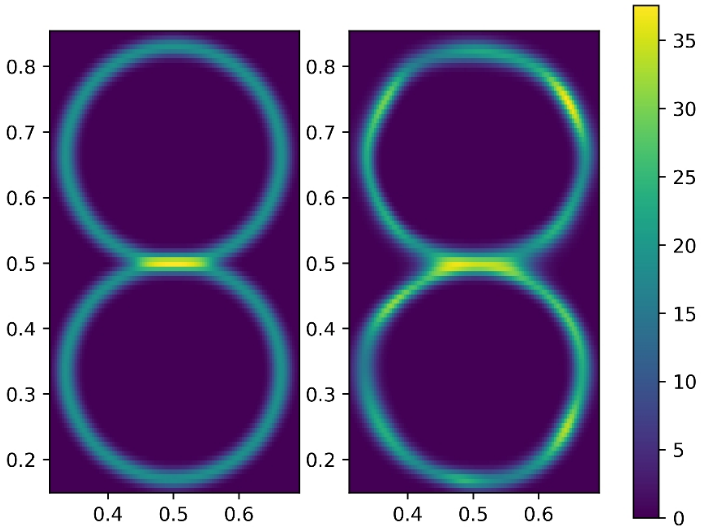

Fig. 3. True multivariate PDF value and one estimated PDF.

In Figure 3, the true PDF is on the left, and the estimated PDF is on the right. Figure 3 presents the difference between the true PDF and the estimated PDF. This result is satisfying as our model captures the global complex shape of this PDF, with however imperfections highlighted by Figure 4. When repeating this method, the estimated density always satisfies $MMD^2 = 10^{-3}$.

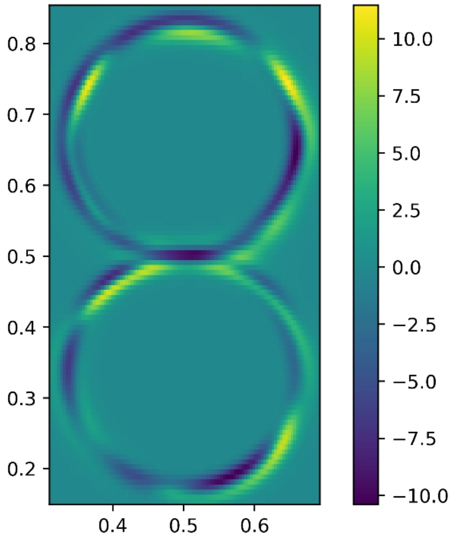

Fig 4. Difference between the true and an estimated PDF.

4. Conclusions and future work

We propose a new method based on ME-PDFs, which combines this statistical framework with NNs. This paper empirically shows that it is possible to use NNs to estimate PDFs. We believe that such PDF estimations would inherently benefit from the strengths of NNs: ability to handle large scale datasets, good extrapolation behaviour, etc.

Multiple parameters of the problem can be further investigated: the sample size, the dimension of the data, the selection of the activation functions, the use of regularization, the architecture of the NN (i.e., the number of layers and neurons per layer), or even the type of NN used. This is left for future work.

However, there remain multiple limitations and questions to address. The evaluation of the integration constant (λ_0 involved in (1)) is more and more difficult as the dimension increases. Some efficient integration methods, like importance sampling or Markov Chain Monte Carlo (Llorente et al., 2023) might be necessary to push the method further. Some questions arise as well on the identification of the domain. We use here an interval closely framing all the sample but, neglecting any physical limitation, we can not know the bounds of the PD. The PD can even be unbounded. The method has to be adapted to account for the unknown nature of the bounds. In a related manner, some methods usually reject some samples by considering them as “outliers”, given some criterion. Similar thoughts could be of interest for us, but the concept of outliers clashes with the principle of maximum entropy. Moreover, ME-PDFs can be defined with multiple constraint functions, but we empirically observed that a unique constraint function is sufficient and the introduction of additional constraints does not improve the fit. Some time could also be dedicated on investigating the separation of the information in the multiple constraints to improve the expressiveness of the network. Furthermore, an extensive amount of work is left to measure the quality of the estimated PDFs, to guarantee the relevance of our method. It will come first with a great care in the metrics used, and their ability to reflect an objective quantitative measure of the density estimation. Finally, the robustness of the method will be investigated, with either some mathematical bounds on the error given by an estimation, or by some numerical benchmarks testing the repeatability and reproducibility of the method.

Acknowledgements

This research was financed by the French government IDEX-ISITE initiative 16-IDEX-0001 (CAP 20-25) under the program DATA. This organization is gratefully acknowledged for the support.

References

Colbert, B. K., Crespo, L. G., Peet, M. M. 2020. A Convex Optimization Approach to Improving Suboptimal Hyperparameters of Sliced Normal Distributions. 2020 American Control Conference (ACC).

Crespo, L. G., Colbert, B. K., Kenny, S. P., Giesy, D. P. 2019. On the quantification of aleatory and epistemic uncertainty using Sliced-Normal distributions. Systems & Control Letters.

Crespo, L. G., Kenny, S. P., Giesy, D. P., Stanford, B. K. 2018. Random variables with moment-matching staircase density functions. Applied Mathematical Modelling.

Dai, H., Ma, Z., Li, L. 2018. An improved complex fractional moment-based approach for the probabilistic characterization of random variables. Probabilistic Engineering Mechanics.

Di Paola, M., Pinnola, F. P. 2012. Riesz fractional integrals and complex fractional moments for the probabilistic characterization of random variables. Probabilistic Engineering Mechanics.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., Smola, A. 2012. A Kernel Two-Sample Test. Journal of Machine Learning Research.

Jaynes, E. T. 1957. Information theory and statistical mechanics. Physical review.

Lin, G. D. 1992. Characterizations of Distributions via Moments. Sankhyā: The Indian Journal of Statistics, Series A (1961-2002).

Llorente, F., Martino, L., Delgado, D., López-Santiago, J. 2023. Marginal Likelihood Computation for Model Selection and Hypothesis Testing: An Extensive Review. SIAM Review.

Magdon-Ismail, M., Atiya, A. 2002. Density estimation and random variate generation using multilayer networks. IEEE Transactions on Neural Networks.

Naaman, M. 2021. On the tight constant in the multivariate Dvoretzky–Kiefer–Wolfowitz inequality. Statistics & Probability Letters.

Novi Inverardi, P. L., Tagliani, A. 2003. Maximum Entropy Density Estimation from Fractional Moments. Communications in Statistics - Theory and Methods.

Shannon, C. E. 1948. A mathematical theory of communication. The Bell System Technical Journal.

Shawe-Taylor, J., Cristianini, N. 2004. Kernel methods for pattern analysis.