entropy

MDPI

Review

Beyond Moments: Extending the Maximum Entropy Principle to Feature Distribution Constraints

Paul M. Baggenstoss

Fraunhofer FKIE, Wachtberg 53343, Germany; p.m.baggenstoss@ieee.org; Tel.: +49-228-9435-150

Received: 27 June 2018; Accepted: 20 August 2018; Published: 30 August 2018

Abstract: The maximum entropy principle introduced by Jaynes proposes that a data distribution should maximize the entropy subject to constraints imposed by the available knowledge. Jaynes provided a solution for the case when constraints were imposed on the expected value of a set of scalar functions of the data. These expected values are typically moments of the distribution. This paper describes how the method of maximum entropy PDF projection can be used to generalize the maximum entropy principle to constraints on the joint distribution of this set of functions.

Keywords: maximum entropy principle; PDF projection; statistical inference

1. Introduction

1.1. Jaynes' Maximum Entropy Principle

The estimation of probability density functions (PDF) is the cornerstone of classical decision theory as applied to real-world problems. The maximum entropy principle of Jaynes [1] proposes that the PDF should have *maximum entropy* subject to constraints imposed by the knowledge one has about the density. Let $\mathbf{x}$ be a set of N random variables $\mathbf{x} = [x_1, x_2 \ldots x_N]$. The entropy of the distribution $p(\mathbf{x})$ is given by

$$H\{p(\mathbf{x})\} = -\int_{\mathbf{x}} p(\mathbf{x}) \log(p(\mathbf{x})) \mathrm{d}\mathbf{x}. \tag{1}$$

Jaynes worked out the case when the knowledge about $p(\mathbf{x})$ consists of the expected value of a set of K measurements. More precisely, he considered the K scalar functions $\phi_1(\mathbf{x}), \phi_2(\mathbf{x}) \ldots \phi_K(\mathbf{x})$ and constrained the expected values:

$$\int_{\mathbf{x}} \phi_k(\mathbf{x})\ p(\mathbf{x})\ \mathrm{d}\mathbf{x} = d_k, \quad 1 \le k \le K. \tag{2}$$

If $\phi_k(\mathbf{x}) = \sum_{i=1}^{N} x_i^k$, then (2) are moment constraints.

The distribution maximizing (1) subject to (2) is:

$$p(\mathbf{x}) = e^{-[\lambda_0 + \lambda_1 \phi_1(\mathbf{x}) + \lambda_2 \phi_2(\mathbf{x}) + \ldots \lambda_K \phi_K(\mathbf{x})]},$$

where λ_0 is the log of the partition function:

$$Z(\lambda_1, \lambda_2 \ldots \lambda_K) = \int_{\mathbf{x}} e^{-[\lambda_1 \phi_1(\mathbf{x}) + \lambda_2 \phi_2(\mathbf{x}) + \ldots \lambda_K \phi_K(\mathbf{x})]}.$$

The constants λ_k are determined by solving

$$d_k = \frac{\partial}{\partial \lambda_k} \log Z, \quad 1 \le k \le K.$$

1.2. Feature Distribution Constraints

Jaynes' results had initial applications in statistical mechanics and thermodynamics [2], and have found more applications in a wide range of disciplines [2–5]. However, we would like to extend the results by replacing constraints (2) with constraints that are more meaningful in real-world inference problems. Instead of knowing just the average values of $\phi_k(\mathbf{x})$, suppose we knew the *joint distribution* of $\mathbf{z} = \Phi(\mathbf{x}) = [\phi_1(\mathbf{x}), \phi_2(\mathbf{x}), \ldots \phi_K(\mathbf{x})]$, denoted by $p_z(\mathbf{z})$. This carries more information than the average values of each measurement $\phi_k(\mathbf{x})$. Because the number of parameters K is small compared with the dimension of $\mathbf{x}$, it is feasible to estimate $p_z(\mathbf{z})$ from a set of training samples using kernel-based PDF estimation methods, for example. This constraint is more general and can be adapted to produce something similar to Jaynes' constraints (2) if the marginal measurement distributions are assumed independent, and Gaussian with mean d_k. This has immediate applications in a wide range of fields, for example in speech analysis and recognition where $\mathbf{z}$ could be MEL frequency cepstrum coefficients (MFCC) [6] extracted from the time-series data $\mathbf{x}$, or in neural networks, where $\mathbf{z}$ could be the output of a network.

Note that the distribution $p_z(\mathbf{z})$ can be obtained from $p(\mathbf{x})$ by marginalization:

$$p_z(\mathbf{z}) = \int_{\mathbf{x} \in \mathcal{M}(\mathbf{z})} p(\mathbf{x}) \, \mathrm{d}\mathbf{x}, \tag{3}$$

where the integral is carried out on the *level set* or manifold given by

$$\mathcal{M}(\mathbf{z}) = \{\mathbf{x} : \phi_1(\mathbf{x}) = z_1, \phi_2(\mathbf{x}) = z_2, \ldots \phi_K(\mathbf{x}) = z_K\}. \tag{4}$$

The constraint problem can then be re-stated as follows:

Problem 1. *Given a known distribution $p_z(\mathbf{z})$, maximize the entropy of $p(\mathbf{x})$ subject to*

$$\int_{\mathbf{x} \in \mathcal{M}(\mathbf{z})} p(\mathbf{x}) \, \mathrm{d}\mathbf{x} = p_z(\mathbf{z}), \quad \forall \mathbf{z}. \tag{5}$$

The solution to this problem is called maximum entropy PDF projection [7–9].

1.3. Significance

The main significance of maximum entropy PDF projection is the de facto creation of a statistical model through the extraction of features. Once a feature extraction $\mathbf{z} = \Phi(\mathbf{x})$ has been identified, and it meets some mild requirements given below, a statistical model has been determined. This has a number of advantages, not the least of which is that the "art" of extracting features, i.e., signal processing, is well established, and many good methods exist to extract meaningful information from data. For example, the extraction MFCC features for processing speech signals has been developed to approximate human hearing [6], and, therefore, with maximum entropy PDF projection, should lead to statistical data models which share some qualities with human perception. Before maximum entropy PDF projection, comparing feature extraction methods had to be done based on secondary factors such as classification results. Maximum entropy PDF projection allows a feature extraction method to be evaluated based its corresponding statistical model.

The use of the maximum entropy principle assures the fairest means of comparing two statistical models derived from competing feature extraction methods. In most real-world applications, we cannot know $p(\mathbf{x})$, and must be satisfied with estimating it from some training data. Suppose that we have a set of K training samples $\mathbf{x}_1, \ \mathbf{x}_2, \ \ldots, \mathbf{x}_K$, and have a number of proposed PDFs computed using (6) for various feature transformations $\mathbf{z}_i = \Phi_i(\mathbf{x})$. Let these projected PDFs be denoted by $p_i(\mathbf{x})$. We would like to determine which projected PDF (i.e., which feature vector) provides a "better" fit to the data. One approach would be to compare the PDFs based on the average log-likelihood $L_i = \frac{1}{K} \sum_{n=1}^{K} \log p_i(\mathbf{x}_n)$, choosing the feature transformation that results in the largest value. However,

likelihood comparison by itself is misleading, so one must also consider the entropy of the distribution, $Q_i = -\int_{\mathbf{x}}\{\log p_i(\mathbf{x})\}\ p_i(\mathbf{x})\mathrm{d}\mathbf{x}$, which is the negative of the theoretical value of L_i. Distributions that spread the probability mass over a wider area have higher entropy since the average value of $\log p(\mathbf{x})$ is lower. The two concepts of Q and L are compared in Figure 1 in which we show three competing distributions: $p_1(\mathbf{x})$, $p_2(\mathbf{x})$, and $p_3(\mathbf{x})$. The vertical lines represent the location of the K training samples. If L_i is the average value of $\log p_i(\mathbf{x})$ at the training sample locations, then clearly $L_1 \ll L_3 \ll L_2$. However, choosing $p_2(\mathbf{x})$ is very risky because it is over-adapted to the training samples. Clearly, $p_2(\mathbf{x})$ has lower entropy since most of the probability mass is at places with higher likelihood. Therefore, it has achieved higher L at the cost of lower Q, a suspicious situation. On the other hand, $Q_1 = Q_3$, but $L_3 > L_1$. Therefore, $p_3(\mathbf{x})$ has achieved higher L than $p_1(\mathbf{x})$ without suffering lower Q, so choosing $p_3(\mathbf{x})$ over $p_1(\mathbf{x})$ is not risky. If we always choose among models that have maximum possible entropy for the given choice of features, we are likely to obtain better features and better generative models.

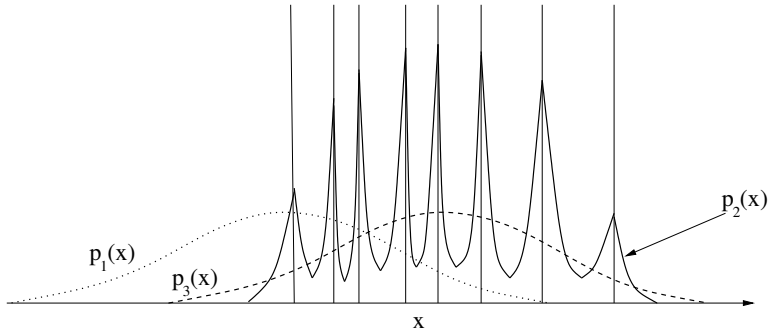

Figure 1. Comparison of entropy Q and average log-likelihood L for three distributions. The vertical lines are the locations of training samples.

2. Main Results

2.1. MaxEnt PDF Projection

The solution to Problem 1 is based on PDF projection [10]. In PDF projection, one is given a feature distribution $p_z(\mathbf{z})$ and constructs a PDF as follows:

$$p_p(\mathbf{x}; p_0, \Phi, p_z) = \frac{p_0(\mathbf{x})}{p_{0,z}(\mathbf{z})}\ p_z(\Phi(\mathbf{x})), \tag{6}$$

where $p_0(\mathbf{x})$ is a reference distribution meeting some mild constraints [10], and $p_{0,z}(\mathbf{z})$ is the corresponding distribution imposed by $p_0(\mathbf{x})$ on the measurements $\mathbf{z}$, i.e., $p_{0,z}(\mathbf{z})$ is Equation (3) applied to $p_0(\mathbf{x})$. It can be shown that

- Equation (6) is a PDF, so $\int_{\mathbf{x}} p_p(\mathbf{x}; p_0, \Phi, p_z)\mathrm{d}\mathbf{x} = 1$.
- $p_p(\mathbf{x}; p_0, \Phi, p_z)$ meets (5), so it is consistent with $p_z(\mathbf{z})$.
- All distributions meeting (5) can be written in the form (6) for some $p_0(\mathbf{x})$.

The last item in the list indicates that, to solve Problem 1, it is only necessary to select the reference distribution $p_0(\mathbf{x})$ for *maximum entropy* (MaxEnt).

To understand the solution to this problem, it is useful to consider the sampling procedure for (6). To sample from distribution (6), one draws a sample $\mathbf{z}^*$ from PDF $p_z(\mathbf{z})$; then, $\mathbf{x}$ is drawn from the set $\mathcal{M}(\mathbf{z}^*)$, defined in (4). Note, however, that to conform to (6), it is necessary to draw sample $\mathbf{x}$ from

$\mathcal{M}(\mathbf{z}^*)$ with probability proportional to the value of $p_0(\mathbf{x})$. The distribution of $\mathbf{x}$ on the manifold $\mathcal{M}(\mathbf{z}^*)$ may be thought of the conditional distribution $p(\mathbf{x}|\mathbf{z}^*)$, and it is proportional to $p_0(\mathbf{x})$. It is in fact

$$p(\mathbf{x}|\mathbf{z}^*) = \frac{p_0(\mathbf{x})}{p_{0,z}(\mathbf{z})}. \tag{7}$$

It can be verified that (7) integrates to 1 on the manifold $\mathcal{M}(\mathbf{z}^*)$. The entropy of (6) can be decomposed as the entropy of $p_z(\mathbf{z})$ plus the expected value of the entropy of the $p(\mathbf{x}|\mathbf{z})$ (see Equation (8) in [8]):

$$H\{p_p(\mathbf{x}; p_0, \Phi, p_z)\} = H\{p_z(\mathbf{z})\} + \int_{\mathbf{z}} H\{p(\mathbf{x}|\mathbf{z})\}\ p_z(\mathbf{z})\ \mathrm{d}\mathbf{z}.$$

Maximizing this quantity seems daunting, but there is one condition under which $H\{p(\mathbf{x}|\mathbf{z})\}$ has the maximum entropy for all $\mathbf{z}$, and that is when $p(\mathbf{x}|\mathbf{z})$ is the *uniform* distribution for all $\mathbf{z}$. This, in turn, is achieved when $p_0(\mathbf{x})$ has a constant value on any manifold $\mathcal{M}(\mathbf{z})$.

This process of selecting $p_0(\mathbf{x})$ for maximum entropy is called maximum entropy PDF projection [8,9]. The maximizing reference distribution is written $p_0^* = \arg\max_{p_0} H\{p_p(\mathbf{x}; p_0, \Phi, p_z)\}$, and the MaxEnt distribution is written

$$p_p^*(\mathbf{x}; \Phi, p_z) = \frac{p_0^*(\mathbf{x})}{p_{0,z}^*(\mathbf{z})}\ p_z(\Phi(\mathbf{x})), \tag{8}$$

which is the unique distribution that solves Problem 1.

In order that it is possible to select $p_0(\mathbf{x})$ for MaxEnt, the feature transformation $\Phi(\mathbf{x})$ must be such that the uniform distribution can be defined on $\mathcal{M}(\mathbf{z})$ for any $\mathbf{z}$. Thus, $\mathcal{M}(\mathbf{z})$ must be bounded and integrable. This condition is easily met if the feature $\mathbf{z}$ contains information about the size of $\mathbf{x}$ so that when $\mathbf{z}$ is fixed to a finite value, the $\mathbf{x}$ has a fixed norm. To say this formally, let there exist a function $f(\mathbf{z})$ such that $f(\Phi(\mathbf{x})) = \|\mathbf{x}\|$ for some valid norm $\|\mathbf{x}\|$ on the range of $\mathbf{x}$.

Once this condition is met, then $p_0^*(\mathbf{x})$ is any distribution that is constant on any level set $\mathcal{M}(\mathbf{z})$. This happens if there exists a function c such that

$$p_0^*(\mathbf{x}) = c(\Phi(\mathbf{x})).$$

Interestingly, any $p_0^*(\mathbf{x})$ meeting these constraints results in the same distribution (6) [8]. This means that, although $p_0^*(\mathbf{x})$ is not unique, $p_p^*(\mathbf{x}; \Phi, p_z)$ is *unique*—it *must* be unique if it is the maximum entropy PDF.

The above conditions can be easily met by inserting an *energy statistic* into the feature set $\Phi(\mathbf{x})$, and defining a reference distribution that depends on $\mathbf{x}$ only through this energy statistic. The energy statistic is a scalar statistic from which it is possible to compute a valid norm on the range of $\mathbf{x}$, denoted by $\mathcal{X}$. In summary, the simplest way to solve for the MaxEnt projected PDF given the range of $\mathbf{x}$, denoted by $\mathcal{X}$, involves these three steps:

1. Identify a norm $\|\mathbf{x}\|$ valid in $\mathcal{X}$ A norm $\|\mathbf{x}\|$ must meet the properties of scalability $\|a\mathbf{x}\| = |a|\|\mathbf{x}\|$, triangle inequality $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$, and $\|\mathbf{0}\| = 0$.
2. Identify a scalar statistic (energy statistic) $t(\mathbf{x})$ from which it is possible to compute $\|\mathbf{x}\|$:

$$\|\mathbf{x}\| = f(t(\mathbf{x})).$$

3. Use a reference hypothesis depending only on $t(\mathbf{x})$.

The above will be demonstrated for three cases of $\mathcal{X}$ in Sections 3.1–3.3.

The data generation process for MaxEnt PDF projection, corresponding to distribution (8) does not depend on $\mathcal{X}$ and is the following:

1. From the known distribution $p_z(\mathbf{z})$, draw a sample denoted by $\mathbf{z}^* = [z_1^*, z_2^* \ldots z_K^*]$.
2. Now identify the set of all samples $\mathbf{x}$ mapping to $\mathbf{z}^*$, denoted by $\mathcal{M}(\mathbf{z}^*)$.

3. Draw a sample $\mathbf{x}$ from this set, uniformly, so that no member of $\mathcal{M}(\mathbf{z}^*)$ is more likely to be chosen than another.

The maximum entropy nature of the solution can be recognized in the uniform sampling on the level set $\mathcal{M}(\mathbf{z}^*)$. The last item above is called uniform manifold sampling (UMS) [9]. The data generation process for three cases of $\mathcal{X}$ are provided in Sections 3.1–3.3.

3. Examples

The implementation of MaxEnt PDF projection depends strongly on the range of the input data $\mathbf{x}$, denoted by $\mathcal{X}$. In this section, examples are provided for three important cases of $\mathcal{X}$.

3.1. Unbounded Data $\mathcal{X} = \mathbb{R}^N$

Let $\mathbf{x}$ range everywhere in $\mathbb{R}^N$. The 2-norm $\|\mathbf{x}\|_2$ is valid in $\mathbb{R}^N$ and can be computed from the total energy

$$t_2(\mathbf{x}) = \sum_{n=1}^{N} x_n^2.$$

The Gaussian reference hypothesis can be written in terms of $t_2(\mathbf{x})$:

$$p_0(\mathbf{x}) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}} e^{-x_i^2/2} = (2\pi)^{-N/2}\ e^{-t_2(\mathbf{x})/2}, \tag{9}$$

so naturally $p_0(\mathbf{x})$ will have a constant value on any manifold $\mathcal{M}(\mathbf{z})$. Naturally, it is not necessary to include $t_2(\mathbf{x})$ explicitly in the feature set—it is only necessary that the 2-norm can be computed from $\mathbf{z}$.

The distribution $p^*_{0,z}(\mathbf{z})$ can be determined in closed form for some feature transformations [11,12]. For others, the moment generating function can be written in closed form, which allows the saddle point approximation to be used to compute $p^*_{0,z}(\mathbf{z})$ [11]. More on this will be presented in Section 4.1.

An important case where a closed-form solution exists is the linear transformation combined with total energy:

$$\mathbf{z} = [\mathbf{A}'\mathbf{x},\ \mathbf{x}'\mathbf{x}].$$

This case is covered in detail in ([8], Section IV.C, p. 2821), and in ([9], Section III.B, p. 2459).

The following simple example demonstrates the main points of this case. Assume input data dimension $N = 3$ and a feature transformation consisting of the sample mean and sample variance:

$$\mathbf{z} = [\hat{\mu}, \hat{v}]\,,$$

where

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i, \quad \hat{v} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \hat{\mu})^2.$$

Note that $t_2(\mathbf{x})$ can be computed from $(\hat{\mu}, \hat{v})$,

$$t_2(\mathbf{x}) = (N-1)\hat{v} + N\hat{\mu}^2,$$

which satisfies the requirement that the 2-norm of $\mathbf{x}$ can be computed from $\mathbf{z}$.

Under the assumption that $\mathbf{x}$ is distributed according to the standard Normal distribution (9), $\hat{\mu}$ will have mean 0 and variance $1/N$,

$$p_0(\mu) = \left(\frac{2\pi}{N}\right)^{-1/2}\ e^{-N\mu^2/2},$$

and $\hat{v}$ will have the chi-square distribution with $N-1$ degrees of freedom and scaling $\frac{1}{N-1}$, which is given by

$$p_0(v) = \frac{k}{2^{k/2}}\,\Gamma^{-1}\left(\frac{k}{2}\right)[kv]^{k/2-1}\,e^{-\frac{vk}{2}},$$

where $k = N-1$. Furthermore, $\hat{\mu}$ and $\hat{v}$ are statistically independent. Therefore, $p^*_{0,z}(\mathbf{z}) = p_0(\mu)\cdot p_0(v)$. For the given feature distribution, we assume components of $\mathbf{z}$ are independent and Gaussian

$$p_z(z_i) = (2\pi v_i)^{-1/2} e^{-(z_i-\mu_i)^2/(2v_i)}$$

with given mean μ_i and variance v_i, where $z_0 = \hat{\mu}$, $z_1 = \hat{v}$. The MaxEnt projected PDF, given by $p^*_p(\mathbf{x};\Phi,p_z) = \frac{p^*_0(\mathbf{x})}{p^*_{0,z}(\mathbf{z})}\;p_z(z)$ is plotted on the left of Figure 2 for slice of x_2, x_3 at $x_1 = 0.0$. The density values shown in the figure, summed over all three axes and properly scaled added to a value 0.9999999998, which validates with numerical integration that $p^*_p(\mathbf{x};\Phi,p_z)$ is a density. Notice that the probability is concentrated on a circular region. This can be understood in terms of the sampling procedure given below.

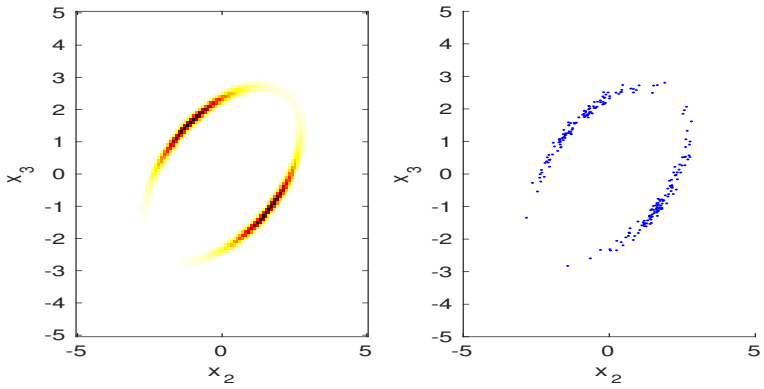

Figure 2. (**Left**) illustration of projected PDF for $\mu_0 = 0.15$, $v_0 = 0.3$, $\mu_1 = 1.85$, $v_1 = 0.025$, on a slice of x_2, x_3 at $x_1 = 0$; (**Right**) samples drawn from the sampling procedure (see text).

To sample from $p^*_p(\mathbf{x};\Phi,p_z)$, we first draw a sample of $\mathbf{z}$ from $p^*_{0,z}(\mathbf{z})$, denoted by $\mathbf{z}^*$, which provides values for the sample mean value μ^* and variance v^*. Then, $\mathbf{x}$ must be drawn uniformly from the manifold $\{\mathbf{x} : \hat{\mu} = \mu^*, \quad \hat{v} = v^*\}$, which are conditions on the sample mean and variance. This is easily accomplished if we note that the sample mean condition is met for any $\mathbf{x}$ of the form

$$\mathbf{x} = [1, 1\ldots 1]'\mu^* + \mathbf{B}\mathbf{u}, \tag{10}$$

where $\mathbf{B}$ is the $N \times (N-1)$ ortho-normal matrix spanning the space orthogonal to the vector $[1, 1\ldots 1]'$. To meet the second (variance) condition, it is necessary that

$$\|\mathbf{u}\|^2 = (N-1)v^*.$$

This condition defines a hypersphere in $(N-1)$ dimensions, which explains the circular region in Figure 2. This hypersphere is sampled uniformly by drawing $N-1$ independent Gaussian random variables, denoted by $\mathbf{u}$, then scaling $\mathbf{u}$ so that $\|\mathbf{u}\|^2 = (N-1)v^*$. Then, $\mathbf{x}$ is constructed using (10). Samples drawn in this manner are shown on the right side of Figure 2. To agree with the left side of the figure, only samples with $|x_1| < 0.01$ are plotted.

Please see the above-cited references for using general linear transformations.

3.2. *Positive Data* $\mathcal{X} = \mathbb{P}^N$

Let $\mathbf{x}$ have positive-valued elements, so $\mathbf{x}$ ranges in the positive quadrant of $\mathbb{R}^N$, denoted by $\mathbb{P}^N$. This holds whenever spectral or intensity data is processed. The appropriate norm in this space is the 1-norm

$$\|\mathbf{x}\| = \frac{1}{N}\sum_{n=1}^{N} x_n.$$

To satisfy conditions for maximum entropy, it must be possible to compute the statistic $t_1(\mathbf{x}) = \sum_{n=1}^{N} x_n$ from the features. The exponential reference hypothesis can be written in terms of $t_1(\mathbf{x})$:

$$p_0(\mathbf{x}) = \prod_{i=1}^{N} e^{-x_i} = e^{-t_1(\mathbf{x})}, \tag{11}$$

so naturally $p_0(\mathbf{x})$ will have a constant value on any manifold $\mathcal{M}(\mathbf{z})$, and is the appropriate reference hypothesis for maximum entropy. The inclusion of $t_1(\mathbf{x})$ explicitly in the feature set is only one way to insure that $\mathcal{M}(\mathbf{z})$ is compact—it is only necessary that the 1-norm can be computed from $\mathbf{z}$.

An important feature extraction is the linear transformation

$$\mathbf{z} = \mathbf{A}'\mathbf{x}.$$

Note that is necessary that statistic $t_1(\mathbf{x})$ can be computed from $\mathbf{z}$, which can be accomplished, for example, to making the first column of $\mathbf{A}$ constant. This case is covered in detail in ([8], Section IV.B, p. 2820), and in ([9], Section IV, p. 2460). Sampling $\mathbf{x}$ is accomplished by drawing a sample $\mathbf{z}^*$ from $p_z(z)$ and then drawing a sample $\mathbf{x}$ uniformly from the set $\{\mathbf{x} : \mathbf{A}'\mathbf{x} = \mathbf{z}^*\}$.

The following simple example demonstrates the main theoretical concepts. We assume a data dimension of $N = 2$ so that the distribution can be visualized as an image. The feature transformation is simply the sum of the samples:

$$z = T(x_1, x_2) = x_1 + x_2.$$

Under the exponential reference hypothesis, the feature distribution is chi-square with $2N$ degrees of freedom and scaling $1/2$:

$$p^*_{0,z}(z) = \frac{2}{\Gamma(k/2)}\, 2^{-k/2}\, (2z)^{(k/2-1)} e^{-z},$$

where $k = 2N$. For the given feature distribution, we assume Gaussian

$$p_z(z) = (2\pi v_z)^{-1/2} e^{-(z-\mu_z)^2/(2v_z)}$$

with a given mean μ_z and variance v_z. The MaxEnt projected PDF, given by $p^*_p(\mathbf{x}; \Phi, p_z) = \frac{p^*_0(\mathbf{x})}{p^*_{0,z}(\mathbf{z})}\, p_z(z)$ is plotted in Figure 3. The density values shown in the figure, when properly scaled, summed to a value 0.9998, which validates with numerical integration that $p^*_p(\mathbf{x}; \Phi, p_z)$ is a density. Note that the distribution is concentrated on the line $x_1 + x_2 = \mu_z = 2$, and is flat on this line, as would be expected for maximum entropy. To sample from this distribution, we first draw a sample z^* from $p_z(z)$ and then draw a sample $\mathbf{x}$ on the line given by $x_1 + x_2 = z^*$. This can be done by sampling x_1 uniformly in $[0, z^*]$, then letting $x_2 = z^* - x_1$. Samples drawn in this way are shown on the right side of Figure 3.

This example generalizes to higher dimension and to arbitrary linear transformations $\mathbf{z} = \mathbf{A}'\mathbf{x}$ for full-rank $N \times M$ matrix $\mathbf{A}$. In this case, $p^*_{0,z}(z)$ is not chi-square, and in fact is not available in closed-form. However, the moment-generating function is available in closed-form so the saddle point approximation may be used (See Section IV.A, p. 2245 in [11]). Samples of $\mathbf{x}$ are drawn by drawing a sample $\mathbf{z}^*$ from $p_z(\mathbf{z})$ and then sampling uniformly in the set $\{\mathbf{x} : \mathbf{A}'\mathbf{x} = \mathbf{z}^*\}$. At high dimensions, this requires a form of Gibbs sampling called hit and run (see Section IV, p. 2460 in [9]).

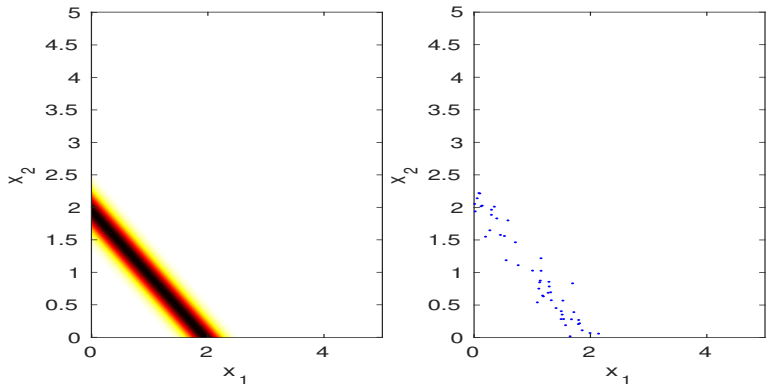

Figure 3. (**Left**) illustration of projected PDF for $\mu_z = 2.0$, $v_z = 0.04$; (**Right**) samples drawn from the sampling procedure (see text).

3.3. *Unit Hypercube, $\mathcal{X} = \mathbb{U}^N$*

Let $\mathbf{x}$ have elements limited to $0 \leq x_i \leq 1$. This case is common when working with neural networks. This is called the unit hypercube, denoted by $\mathbb{U}^N$. The uniform reference hypothesis

$$p_0(\mathbf{x}) = 1. \tag{12}$$

produces maximum entropy. No norm-producing energy statistic is needed. Naturally, $p_0(\mathbf{x})$ will have a constant value on any manifold $\mathcal{M}(\mathbf{z})$.

The following simple example demonstrates the main theoretical concepts. We assume a data dimension of $N = 2$ so that the distribution can be visualized as an image. The feature transformation is simple the sum of the samples:

$$z = T(x_1, x_2) = x_1 + x_2.$$

For this case, the uniform distribution brings maximum entropy, $p_0^*(\mathbf{x}) = 1$. Under the reference hypothesis, the feature distribution is Irwin-Hall, given by

$$p_{0,z}^*(z) = \frac{1}{2(N-1)!} \sum_{k=0}^{N} (-1)^k \binom{N}{k} (z-k)^{N-1} \text{sign}(z-k),$$

where $\text{sign}(0) = 0$. For $N = 2$, this is a triangular distribution

$$p_{0,z}^*(z) = \{z, \;\; 0 \leq z \leq 1; \qquad 2 - z, \;\; 1 \leq z \leq 2\}.$$

For the given feature distribution, we assume Gaussian

$$p_z(z) = (2\pi v_z)^{-1/2} e^{-(z-\mu_z)^2/(2v_z)}$$

with a given mean μ_z and variance v_z. The MaxEnt projected PDF, given by $p_p^*(\mathbf{x}; \Phi, p_z) = \frac{p_0^*(\mathbf{x})}{p_{0,z}^*(\mathbf{z})} \; p_z(z)$ is plotted in Figure 4. The density values shown in the figure, when properly scaled, summed to a value 0.999, which validates with numerical integration that $p_p^*(\mathbf{x}; \Phi, p_z)$ is a density. Note that the distribution is concentrated on the line $x_1 + x_2 = \mu_z$, and is flat on this line, as would be expected for maximum entropy. To sample from this distribution, we first draw a sample z^* from $p_z(z)$ and then draw a sample $\mathbf{x}$ on the line given by $x_1 + x_2 = z^*$. This can be done by finding where the line that intercepts the axes, and sampling uniformly in the interval between the intercepts. Note that this sampling differs from the previous example as a result of the upper bound at 1.

This example generalizes to higher dimension and to arbitrary linear transformations $\mathbf{z} = \mathbf{A}'\mathbf{x}$ for full-rank $N \times M$ matrix $\mathbf{A}$. In this case, $p_{0,z}^*(z)$ is no longer Irwin-Hall and in fact is not available in

closed-form. However, the moment-generating function is available in closed-form so the saddle point approximation may be used (see Appendix in [13]). Samples of $\mathbf{x}$ are drawn by drawing a sample $\mathbf{z}^*$ from $p_z(\mathbf{z})$ and then sampling uniformly in the set $\{\mathbf{x} : \mathbf{A}'\mathbf{x} = \mathbf{z}^*\}$. At high dimensions, this requires a form of Gibbs sampling called hit and run (see p. 2465 in [9]).

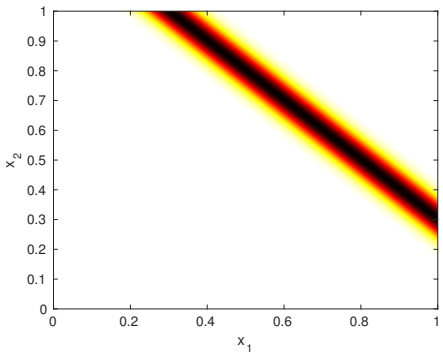

Figure 4. Illustration of projected PDF for $\mu_z = 1.3$, $v_z = 0.002$.

4. Advanced Concepts

4.1. Implementation Issues

Implementing (8) seems like a daunting numerical task, since $p_0^*(\mathbf{x})$ is some canonical distribution, for which a real data sample $\mathbf{x}$ normally lies in the far tails of both $p_0^*(\mathbf{x})$ and $p_{0,z}^*(\mathbf{z})$. However, if the distributions are known exactly, and are represented in the log domain, then the difference

$$\log p_0^*(\mathbf{x}) - \log p_{0,z}^*(\mathbf{z}) \tag{13}$$

typically remains within very reasonable limits. In some cases, terms in $\log p_0^*(\mathbf{x})$ and $\log p_{0,z}^*(\mathbf{z})$ cancel, leaving (13) only weakly dependent on $\mathbf{x}$ (for example, see Section IV.A, p. 2820 in [8]).

Evaluating $\log p_0^*(\mathbf{x})$ is mostly trivial since it is normally a canonical distribution, such as Gaussian, exponential, or uniform. Calculating $\log p_{0,z}^*(\mathbf{z})$, however, remains the primary challenge in maximum entropy PDF projection. However, when evaluating $p_{0,z}^*(\mathbf{z})$ seems daunting, there are several ways to overcome the problem.

1. Saddle Point Approximation. If $p_{0,z}^*(\mathbf{z})$ is not available in closed form, the moment-generating function (MGF) might be tractable. This allows the saddle point approximation (SPA) to be used (see Section III in [11]). Note that the term "approximation" is misleading because the SPA approximates the shape of the MGF on a contour, not the absolute value, so the SPA expression for $\log p_{0,z}^*(\mathbf{z})$ remains very accurate, in the far tails, even when $p_{0,z}^*(\mathbf{z})$ itself cannot be evaluated in machine precision. Examples of this include general linear transformations of exponential and chi-squared random variables (see Section III.C and Section IV in [11]), general linear transformations of uniform random variables (Appendix in [13]), a set of linear-quadratic forms [14], and order statistics [15].
2. Floating reference hypothesis. There are conditions under which the MaxEnt reference hypothesis $p_0^*(\mathbf{x})$ is not unique, so it can depend on a parameter $\boldsymbol{\theta}$, so we write $p_0^*(\mathbf{x};\boldsymbol{\theta})$. An example is when the feature $\mathbf{z}$ contains the sample mean and sample variance (see example in Section 3.1). In this case, a Gaussian reference hypothesis $p_0^*(\mathbf{x};\boldsymbol{\theta})$ can be modified to have any mean and variance $\boldsymbol{\theta} = [\mu_0, \sigma_0^2]$, and can serve as the MaxEnt reference hypothesis with no change at all in the resulting projected PDF. In other words, (13) is independent of $\boldsymbol{\theta}$—this can be verified by cancelling terms. Therefore, there is no reason that $\boldsymbol{\theta}$ cannot be made to track the data—that is, let $\mu_0 = \hat{\mu}(\mathbf{x})$, $\sigma_0^2 = \hat{\sigma}^2(\mathbf{x})$. By doing this, $p_{0,z}^*(\mathbf{z})$ will track $\mathbf{z}$, allowing simple approximations based on central limit theorem to be used to approximate $p_{0,z}^*(\mathbf{z})$.

3. Chain Rule. When $p^*_{0,z}(\mathbf{z})$ cannot be derived for a feature transformation, it may be possible to break the feature transformation into stages, where each stage can be easily analyzed. The next section is devoted to this.

4.2. Chain Rule

The primary numerical difficulty in implementing (8) is the calculation of $p^*_{0,z}(\mathbf{z})$. Solutions for many of the most useful feature transformations are available [9,11–13]. However, in many real-world applications, such as neural networks, the feature transformations cannot be easily written in a compact form $\mathbf{z} = [\phi_1(\mathbf{x}), \phi_2(\mathbf{x}), \ldots \phi_K(\mathbf{x})]$. Instead, they consist of multi-stage transformations, for example, $\mathbf{y} = T_1(\mathbf{x})$, $\mathbf{w} = T_2(\mathbf{y})$, and $\mathbf{z} = T_3(\mathbf{w})$. The individual stages $T_m(\mathbf{x})$ could be the layers of a neural network. In this case, it is best to apply (8) recursively to each stage. This means that the distribution of the first stage features $p(\mathbf{y})$ is written using (6) with $\mathbf{y}$ taking the role of input data, and so forth. This results in the chain-rule form:

$$p(\mathbf{x}) = \left[\frac{p^*_{0,x}(\mathbf{x})}{p^*_{0,x}(\mathbf{y})}\right]\left[\frac{p^*_{0,y}(\mathbf{y})}{p^*_{0,y}(\mathbf{w})}\right]\left[\frac{p^*_{0,w}(\mathbf{w})}{p^*_{0,w}(\mathbf{z})}\right] p(\mathbf{z}), \tag{14}$$

where $p^*_{0,x}(\mathbf{x})$, $p^*_{0,y}(\mathbf{y})$, $p^*_{0,w}(\mathbf{w})$ are canonical reference hypotheses used at each stage, for example (9), (11), and (12), depending on the range of $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{w}$, respectively.

To understand the importance of the chain-rule, consider how we would compute (6) without the chain rule. Let $T(\mathbf{x})$ be the combined transformation

$$T(\mathbf{x}) = T_3(T_2(T_1(\mathbf{x})))$$

and let $p^*_0(\mathbf{x})$ be one of the canonical reference distributions. Consider the difficulty in deriving $p^*_{0,z}(\mathbf{z})$. At each stage, the distribution of the output feature becomes more and more intractable, and trying to estimate $p^*_{0,z}(\mathbf{z})$ is futile because generally a canonical reference distribution is completely unrealistic as PDF for real data. Furthermore, $p^*_{0,z}(\mathbf{z})$ is more often than not evaluated in the far tails of the distribution. With the chain-rule, however, we can assume a suitable canonical reference hypothesis at the start of each stage, and only need to derive the feature distribution imposed on the output of that stage.

As long as the reference hypothesis used at each stage meets the stated requirements given in Section 2.1, then the chain as a whole will indeed produce the desired MaxEnt projected PDF, which is the PDF with maximum entropy among all PDFs that generate the desired output feature distribution $p(\mathbf{z})$ through the combined transformation [8]!

An example of the application of the chain-rule is the computation of MEL frequency cepstral coefficients (MFCC), commonly used in speech processing. Let us consider a frame of data of length N, denoted by $\mathbf{x}$. The processing is broken into the following stages:

1. The first step, denoted by $\mathbf{y} = T_1(\mathbf{x})$ is to convert $\mathbf{x}$ into $N/2 + 1$ magnitude-squared discrete Fourier transform (DFT) bins. Under the standard Gaussian assumption (9), the elements of $\mathbf{y}$ are independent and have chi-squared statistics (see Section VI.D.1, pp. 47–48 in [12]).
2. The second step is to sum energy in a set of K MEL-spaced band functions. This results in a set of K band energies. This can be written using the $(N/2 - 1) \times K$ matrix $\mathbf{A}$ as the linear transformation $\mathbf{w} = \mathbf{A}'\mathbf{y}$. This feature transformation is explained in Section 3.2 above so an exponential reference distribution can be assumed for $\mathbf{y}$. Care must be taken that the K band functions add to a constant—this insures the energy statistic is "contained in the features".
3. The next step is to compute the log of the K band energies, $\mathbf{u} = \log(\mathbf{w})$. This is a 1:1 transformation for which PDF projection simplifies to computing the determinant of the transformation's Jacobian matrix (see Section VI.A, p. 46 in [12]).

4. The last step is the discrete cosine transform (DCT), which can be written as a linear transformation $\mathbf{z} = \mathbf{C}'\mathbf{u}$. If some DCT coefficients are discarded, then the transformation must be analyzed as in Section 3.1 above by including the energy statistic $t(\mathbf{u}) = \mathbf{u}'\mathbf{u}$.

This example illustrates that complex feature transformations can be easily analyzed if broken into simple steps. More on the above example can be found in Sections V and VI in [8].

4.3. Large-N Conditional Distributions and Applications.

When the feature value $\mathbf{z}^*$ is fixed, then sampling $\mathbf{x}$ on the manifold $\mathcal{M}(\mathbf{z}^*)$, called UMS, has some interesting interpretations relative to maximum entropy. Let the conditional distribution be written $p(\mathbf{x}|\mathbf{z}^*)$. Notice that $p(\mathbf{x}|\mathbf{z}^*)$ is not a proper distribution since all the probability mass exists on the manifold $\mathcal{M}(\mathbf{z}^*)$ of zero volume. Writing down $p(\mathbf{x}|\mathbf{z}^*)$ in closed form or determining its mean is intractable. It is useful, however, to know $p(\mathbf{x}|\mathbf{z}^*)$ because, for example, the mean of $p(\mathbf{x}|\mathbf{z}^*)$ is a point estimate of $\mathbf{x}$ based on $\mathbf{z}^*$, a type of MaxEnt feature inversion. However, depending on the range of $\mathbf{x}$, as exemplified by the three cases in Sections 3.1–3.3, $p(\mathbf{x}|\mathbf{z}^*)$ can be approximated by a *surrogate distribution* (See p. 2461 in [9]). The surrogate distribution is a proper distribution that (a) has its probability mass concentrated near $\mathcal{M}(\mathbf{z}^*)$, (b) has constant value on $\mathcal{M}(\mathbf{z}^*)$, and (c) has mean value on the manifold, so $\bar{\mathbf{x}} \in \mathcal{M}(\mathbf{z}^*)$. The surrogate distribution therefore meets the same conditions as $p(\mathbf{x}|\mathbf{z}^*)$ but is a proper distribution. The mean of the surrogate distribution is a very close approximation to the mean of $p(\mathbf{x}|\mathbf{z}^*)$, which can be called th *centroid* of $\mathcal{M}(\mathbf{z}^*)$, but can be computed. In Sections 3.1–3.3, the surrogate distribution is Gaussian, exponential, and truncated exponential, respectively. These are the MaxEnt distributions under applicable constraints. It was shown, for example, when the range of $\mathbf{x}$ is the positive quadrant of $\mathcal{R}$, that the centroid corresponds to the classical Maximum Entropy feature inversion approach for a dimension-reducing linear transformation of intensity data, for example to sharpen images blurred by a point-spread function [9]. The method, however, is more general because it can be adapted to different ranges of $\mathbf{x}$ [9].

5. Applications

5.1. Classification

Assume there are M classes and the M class hypotheses are $H_1, H_2 \ldots H_M$. The general form of the classifier by applying Bayes theorem and (8) is given by

$$\hat{m} = \arg\max_m p(\mathbf{x}|H_m)\ p(H_m), \tag{15}$$

where $p(H_m)$ is the prior class probability, and $p(\mathbf{x}|H_m)$ is a PDF estimate for class hypothesis H_m. For the classification problem, there are many classifier topologies for using (8) to construct $p(\mathbf{x}|H_m)$.

1. Class-specific features. One can specify a different feature transformation per class, $\mathbf{z}^m = \Phi^m(\mathbf{x})$,

$$p(\mathbf{x}|H_m) = \frac{p_0^*(\mathbf{x})}{p_0^*(\mathbf{z}^m)}\ p(\mathbf{z}^m|H_m),$$

but the numerator is common, so the classifier rule becomes

$$\hat{m} = \arg\max_m \frac{p(\mathbf{z}^m|H_m)}{p_0^*(\mathbf{z}^m)}.$$

This amounts to just comparing the likelihood ratio between class hypothesis H_m and the reference distribution, computed using a class-dependent feature [16].

2. It is not necessary to use a common reference hypothesis. A class-dependent reference hypothesis can be selected so that the feature is an approximately sufficient statistic to discriminate the given class from the class-dependent reference hypothesis. Then,

$$p(\mathbf{x}|H_m) = \frac{p_{0,m}(\mathbf{x})}{p_{0,m}(\mathbf{z}^m)}\, p(\mathbf{z}^m|H_m),$$

where $p_{0,m}(\mathbf{x})$ is the class-dependent reference hypothesis. Note that, when using the chain-rule (14), there is not a single reference hypothesis associated with each class, but a series of stage-wise reference hypotheses. Note that here we have relaxed the MaxEnt requirement for the reference hypothesis.

3. Using a different feature to test each class hypothesis is not always a good idea. Some data can be "contaminated" with noise or interference, so they may not be suitable to test a hypothesis with just one feature. In this case, a *class-specific feature mixture* (CSFM) [17–19] may be appropriate. For the CSFM, we define a set of feature transformations $\{\Phi^1(\mathbf{x}), \Phi^2(\mathbf{x}), \ldots \Phi^M(\mathbf{x})\}$. (We assume here that the number of feature transformations equals the number of classes, but this is not necessary.) Then, $p(\mathbf{x}|H_m)$ is constructed as a mixture density using all the features:

$$p(\mathbf{x}|H_m) = \sum_{l=1}^{M} w_{m,l}\, \frac{p^*_{0,l}(\mathbf{x})}{p^*_{0,l}(\mathbf{z}^l)}\, p(\mathbf{z}^l|H_m),$$

where $p^*_{0,l}(\mathbf{x})$ is the MaxEnt reference hypothesis corresponding to each feature transformation $\Phi^l(\mathbf{x})$.

4. To solve the classification problem (15), it is necessary to obtain a segment of data $\mathbf{x}$ that can be classified into one of M classes. The problem is often not that simple, and the location of the classifiable "event" may be unknown within a longer data recording, or the data recording may contain multiple events from multiple classes. Using MaxEnt PDF projection, it is possible to solve the data segmentation problem simultaneously with the classification problem [20,21].

5.2. Other Applications

MaxEnt PDF projection has applications in the analysis of networks and feature transformations. For example in neural networks, it is possible to view a feed-forward neural network as a generative network, a duality relationship between two opposing types of networks [22]. In addition, the restricted Boltzmann machine (RBM) can be used as a PDF estimator with tractable distribution [13]. In feature inversion, MaxEnt PDF projection can be used to find MaxEnt point-estimates of the input data $\mathbf{x}$ based on fixed values of the feature [9].

6. Conclusions

In this short paper, the method of maximum entropy PDF projection was presented as a generalization of Jaynes' maximum entropy principle with moment constraints. The mathematical basis of maximum entropy PDF projection was reviewed and practical considerations and applications were presented.

Funding: This research was funded by Fraunhofer FKIE, Wachtberg, Germany.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Jaynes, E.T. Information Theory and Statistical Mechanics I. *Phys. Rev.* **1957**, *106*, 620–630. [CrossRef]
2. Kesavan, H.K.; Kapur, J.N. The Generalized maximum Entropy Principle. *IEEE Trans. Syst. Man Cybern.* **1989**, *19*, 1042–1052. [CrossRef]

3. Banavar, J.R.; Maritan, A.; Volkov, I. Applications of the principle of maximum entropy: From physics to ecology. *J. Phys. Condens. Matter* **2010**, *22*. [CrossRef] [PubMed]
4. Holmes, D.E. (Ed.) *Entropy, Special Issue on Maximum Entropy and Its Application*; MDPI: Basel, Switzerland, 2018.
5. Martino, A.D.; Martino, D.D. An introduction to the maximum entropy approach and its application to inference problems in biology. *Heliyon* **2018**, *4*, e00596. [CrossRef] [PubMed]
6. Picone, J.W. Signal Modeling Techniques in Speech Recognition. *Proce. IEEE* **1993**, *81*, 1215–1247. [CrossRef]
7. Baggenstoss, P.M. Maximum entropy PDF projection: A review. *AIP Conf. Proc.* **2017**. [CrossRef]
8. Baggenstoss, P.M. Maximum Entropy PDF Design Using Feature Density Constraints: Applications in Signal Processing. *IEEE Trans. Signal Process.* **2015**, *63*, 2815–2825. [CrossRef]
9. Baggenstoss, P.M. Uniform Manifold Sampling (UMS): Sampling the Maximum Entropy PDF. *IEEE Trans. Signal Process.* **2017**, *65*, 2455–2470. [CrossRef]
10. Baggenstoss, P.M. The PDF Projection Theorem and the Class-Specific method. *IEEE Trans. Signal Process.* **2003**, *51*, 672–685. [CrossRef]
11. Kay, S.M.; Nuttall, A.H.; Baggenstoss, P.M. Multidimensional probability density function approximations for detection, classification, and model order selection. *IEEE Trans. Signal Process.* **2001**, *49*, 2240–2252. [CrossRef]
12. Baggenstoss, P.M. The Class-Specific Classifier: Avoiding the Curse of Dimensionality (Tutorial). *IEEE Aerosp. Electron. Syst. Mag. Spec. Tutor. Add.* **2004**, *19*, 37–52. [CrossRef]
13. Baggenstoss, P.M. Evaluating the RBM Without Integration Using PDF Projection. In Proceedings of the EUSIPCO 2017, Kos, Greece, 28 August–2 Septemper 2017.
14. Nuttall, A.H. *Saddlepoint Approximation and First-Order Correction Term to The Joint Probability Density Function of M Quadratic and Linear Forms in K Gaussian Random Variables With Arbitrary Means and Covariances*; NUWC Technical Report 11262; US Naval Undersea Warfare Center: Newport, RI, USA, 2000.
15. Nuttall, A.H. *Joint Probability Density Function of Selected Order Statistics And the Sum of the Remaining Random Variables*; NUWC Technical Report 11345; US Naval Undersea Warfare Center: Newport, RI, USA, 2002.
16. Baggenstoss, P.M. Class-Specific Features in Classification. *IEEE Trans. Signal Process.* **1999**, *47*, 3428–3432. [CrossRef]
17. Baggenstoss, P.M. Optimal Detection and Classification of Diverse Short-Duration Signals. In Proceedings of the International Conference on Cloud Engineering, Boston, MA, USA, 11–14 March 2014; pp. 534–539.
18. Baggenstoss, P.M. Class-specific model mixtures for the classification of time-series. In Proceedings of the 2015 23rd European Signal Processing Conference (EUSIPCO), Nice, France, 31 August–4 September 2014.
19. Baggenstoss, P.M. Class-Specific Model Mixtures for the Classification of Acoustic Time-Series. *IEEE Trans. AES* **2016**, *52*, 1937–1952. [CrossRef]
20. Baggenstoss, P.M. A multi-resolution hidden Markov model using class-specific features. *IEEE Trans. Signal Process.* **2010**, *58*, 5165–5177. [CrossRef]
21. Baggenstoss, P.M. Acoustic Event Classification using Multi-resolution HMM. In Proceedings of the European Signal Processing Conference (EUSIPCO) 2018, Rome, Italy, 3–7 September 2018.
22. Baggenstoss, P.M. On the Duality Between Belief Networks and Feed-Forward Neural Networks. *IEEE Trans. Neural Netw. Learn. Syst.* **2018**, 1–11. [CrossRef] [PubMed]