# Maximum Entropy PDF Design Using Feature Density Constraints: Applications in Signal Processing

Paul M. Baggenstoss

***Abstract*—This paper revisits an existing method of constructing high-dimensional probability density functions (PDFs) based on the PDF at the output of a dimension-reducing feature transformation. We show how to modify the method so that it can provide the PDF with the highest entropy among all PDFs that generate the given low-dimensional PDF. The method is completely general and applies to arbitrary feature transformations. The chain-rule is described for multi-stage feature calculations typically used in signal processing. Examples are given including MFCC and auto-regressive features. Experimental verification of the results using simulated data is provided including a comparison with competing generative methods.**

***Index Terms*—Maximum entropy, statistical learning, statistical distributions, PDF estimation.**

## I. Introduction

### A. Why PDF Estimation on the Input Data?

In the last decades, the fields of classification, machine learning, and computer vision have witnessed remarkable advances in classification methods based on discriminative methods (neural networks, deep learning, support vector machines), while generative methods have taken a back seat. Generative methods, which are based on estimating the probability density function (PDF), are often applied in the lower-dimensional feature space, and usually compare unfavorably to discriminative methods. The use of generative methods in the high-dimensional input data space is rarely seen, except in cases when the data is well defined, such as for detecting known waveforms in Gaussian noise (replica correlation). With the arrival of the PDF projection method, generative methods can now be extended to the input data, without requiring PDF estimation at high dimensions, even for data that is not well defined (aside from knowing a suitable feature transformation) [1]. In PDF projection, the feature PDF is estimated, then mathematically "projected" to the input data domain. Input data PDFs can even be constructed using multiple features by forming kernel mixtures of the separate projected PDFs.

There are intuitive arguments for going to the input data when constructing classifiers: collections of real data may contain samples from a variety of sources. It is not hard to imagine a situation where two data classes originating from different signal generation mechanisms have similar spectral content, and so are indistinguishable by spectral analysis alone. The only way to distinguisn them would be by separate input-data generative models adapted to their respective generation mechanisms. PDF projection offers this possibility if based on features matching the data generation processeses. Mathematical arguments also exist. Steven Kay [2] presents a constructed example where no single minimal sufficient statistic exists for testing between two hypotheses $H_1$ and $H_2$, whereas PDF projection (then known as class-specific features) results in an optimal test. Despite these arguments, PDF projection remains largely unknown. One reason for this is the lack of a clear understanding of the optimality of the method. After all, there are an infinite number of input data PDFs that are consistent with the given feature PDF. Why should the one provided by PDF projection be better than any other? In this paper, we answer this question.

### B. Maximum Entropy Principle

The maximum entropy (ME) principle is a well-established criterion for design of probability density functions (PDFs) [3]–[5]. The entropy of a distribution $p(\mathbf{x})$ is given by

$$Q = -\mathcal{E}_{\mathbf{x}}\{\log p(\mathbf{x})\} = -\int_{\mathbf{x}} \log\{p(\mathbf{x})\}p(\mathbf{x})\mathrm{d}\mathbf{x}. \quad (1)$$

The PDF $p(\mathbf{x})$ that maximizes (1) is the best representation of the current state of knowledge of random variable $\mathbf{x}$ [6]. Prior knowledge about $p(\mathbf{x})$ is introduced through constraints in the maximization. If there is nothing known about $\mathbf{x}$, other than that it is contained in the interval $[a, b]$, the uniform distribution is the maximum entropy density [3], [4]. This extends to bounded regions in higher dimensions. Moment constraints lead to PDFs in the exponential family. The exponential and Gaussian distributions are examples of maximum entropy PDFs satisfying moment constraints.

### C. Goal and Problem Statement

Our goal is to adapt the method of moments to PDF projection, and that means re-formulating the problem using features

Manuscript received November 26, 2014; revised March 21, 2015; accepted March 23, 2015. Date of publication April 02, 2015; date of current version April 30, 2015. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Cedric Fevotte. This work was partially funded by internal research grant by the Naval Udersea Warcere Center, Newport, RI.

The author is with the Naval Undersea Warfare Center, Newport, RI 02840 USA (e-mail: p.m.baggenstoss@ieee.org; web: http://class-specific.com/csf).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSP.2015.2419189

instead of moments. Consider an arbitrary mapping from $\mathcal{R}^N$ to $\mathcal{R}^D$

$$\mathbf{z} = T(\mathbf{x}), \quad \mathbf{z} \in \mathcal{R}^D, \quad \mathbf{x} \in \mathcal{R}^N, \tag{2}$$

where $D \leq N$. We are given an arbitrary multivariate feature PDF $g(\mathbf{z})$. Consider the family of all PDFs that generate $g(\mathbf{z})$ through $T(\mathbf{x})$, written $G_T(\mathbf{x})$ to distinguish them from PDFs $p(\mathbf{x})$ that might not generate $g(\mathbf{z})$. In this paper we seek to solve the following problem.

*Problem 1:* Let feature $\mathbf{z}$ be computed from $\mathbf{x}$ using feature transformation (2). Find the PDF $G_T(\mathbf{x})$ that maximizes (1) under the constraint that the PDF generated on $\mathbf{z}$, through $T(\mathbf{x})$, is $g(\mathbf{z})$.

We will show that if $\mathbf{z}$ contains an *energy statistic* (ES), Problem 1 can be easily solved. The ES is typically a scalar statistic and is related to a norm and serves the purpose of a moment constraint.

### D. Prior Work

Most existing methods incorporating maximum entropy into machine learning and classification apply the concept to the features only, ignoring the raw data distribution and feature extraction itself [7]–[10].

Closer to our goal is the work of Basu *et al.* [11] which seeks a maximum entropy density defined on the input data, but is based only on uni-variate marginals of the features and considers only linear feature transformations. We seek a more general method to use the ME principle to design PDFs on the high-dimensional data based on the low-dimensional feature density. In our own prior work on PDF projection [1], we construct high-dimensional PDFs based on low-dimensional feature PDFs, but do not investigate maximum entropy. Later, Kay examined PDF projection from the point of view of Kullback-Leibler divergence (KLD) [12], showing that PDF projection achieves the lowest KLD to a target PDF. KLD optimality generally implies maximum entropy [13], so Kay's result is of relevance to Problem 1 and deserves close examination. In Section 2.1, we discover that while Kay's result points to PDF projection as an optimal form for a given reference hypothesis, we still need to maximize the entropy over the chosen reference hypothesis.

### E. Applications

Working with the input data PDF $G_T(\mathbf{x})$ opens a wide variety of possibilities that are not available when working with just feature distributions.

1. **Hypothesis testing using mixed features**. We may construct projected PDFs based on several feature transformations $T_1(\mathbf{x}), T_2(\mathbf{x}), \ldots$ using kernel mixtures:

$$P(\mathbf{x}) = \sum_i \alpha_i G_{T_i}(\mathbf{x}).$$

   It is also possible to determine which feature transformation is "best" based on maximum likelihood.
2. **Monte Carlo methods**. The raw data PDF $G_T(\mathbf{x})$ is a generative model from which we may generate samples of $\mathbf{x}$. This opens new research directions in applying Monte Carlo methods, which rely on generating samples of a model density, to high-dimensional data.

TABLE I
REFERENCE PDFS AND THEIR ENERGY STATISTICS

| Name | $p(\mathbf{x}\|H_0)$ | $t(\mathbf{x})$ |
|---|---|---|
| Exponential | $\prod_{i=1}^n e^{-x_i}$ | $\sum_{i=1}^N x_i$ |
| Gaussian | $\prod_{i=1}^n \frac{e^{-x_i^2/2}}{\sqrt{2\pi}}$ | $\sum_{i=1}^N x_i^2$ |
| Chi-squared(1) | $\prod_{i=1}^n \frac{e^{-x_i/2}}{\sqrt{2\pi x_i}}$ | $\sum_{i=1}^N (\log x_i + x_i)$ |
| Laplacian | $\prod_{i=1}^n \frac{1}{\sqrt{2}} e^{-\sqrt{2}\,\|x_i\|}$ | $\sum_{i=1}^N \|x_i\|$ |
| Uniform | $\left(\frac{1}{b-a}\right)^N$, $a \leq x_i \leq b$ | n/a |

3. **Sensor Fusion**. When data must be compressed, and later inference about the original data must be made, $G_T(\mathbf{x})$ can be used as an optimal PDF estimate. When the same original data is observed through multiple sensors, the information can be fused statistically [14].

## II. MATHEMATICAL PRELIMINARIES

### A. PDF Projection

Consider a statistical reference hypothesis $H_0$ under which the distributions of both $\mathbf{x}$ and $\mathbf{z}$ are known and given by $p(\mathbf{x} \mid H_0)$ and $p(\mathbf{z}|H_0)$, respectively. This is a theoretical reference hypothesis that does not necessarily correspond to any realistic data. Examples of reference hypotheses are given in Table I. Let there be an arbitrary PDF defined on the feature space $g(\mathbf{z})$. The PDF projection theorem (PPT) [1] states that the function

$$G_T(\mathbf{x}; H_0) = \left[\frac{p(\mathbf{x} \mid H_0)}{p(\mathbf{z} \mid H_0)}\right] g(\mathbf{z}), \tag{3}$$

is a PDF (it integrates to 1 over $\mathbf{x}$), and that the PDF $G_T(\mathbf{x}; H_0)$ indeed generates $g(\mathbf{z})$ through $T(\mathbf{x})$.

To understand this form in terms of *statistical sufficiency* [15], we re-organize (3) as

$$\frac{G_T(\mathbf{x}; H_0)}{p(\mathbf{x} \mid H_0)} = \frac{g(\mathbf{z})}{p(\mathbf{z} \mid H_0)},$$

which is the classical requirement for $\mathbf{z}$ to be sufficient in distinguishing $G_T(\mathbf{x}; H_0)$ from $p(\mathbf{x}|H_0)$. It follows that if $T(\mathbf{x})$ is a sufficient statistic for the binary test between $H_0$ and some hypothesis $H_1$, and if $g(\mathbf{z}) \to p(\mathbf{z}|H_1)$, then, $G(\mathbf{x}) \to p(\mathbf{x}|H_1)$. But, in the absence of knowing $H_0$ and $H_1$ for which $\mathbf{z}$ is sufficient, the argument appears circular—$G(\mathbf{x})$ exists based on the sufficiency of $\mathbf{z}$, and $\mathbf{z}$ is sufficient based on the existence of $G(\mathbf{x})$, leaving us unsure if (3) is always a PDF.

The following constructive way to arrive at (3) is due to Kay [16]. Let there exist a complete invertible version of feature transformation (2), given by $[\mathbf{z}, \mathbf{u}] = T_c(\mathbf{x})$, with inverse $\mathbf{x} = T_c^{-1}(\mathbf{z}, \mathbf{u})$, where $\mathbf{u}$ is the "ancillary" statistic. We then generate a sample $\mathbf{x}$ in the following manner: (a) draw a feature value $\mathbf{z}^*$ from PDF $g(\mathbf{z})$, (b) then draw a sample $\mathbf{u}^*$ from the conditional PDF $p(\mathbf{u}|\mathbf{z}^*)$, which is assumed to exist, (c) finally transform $\{\mathbf{z}^*, \mathbf{u}^*\}$ using $T_c^{-1}(\,)$ to get a sample $\mathbf{x}^*$. First, it is obvious that $\{\mathbf{z}^*, \mathbf{u}^*\}$ are samples of the PDF $p(\mathbf{z}, \mathbf{u}) = p(\mathbf{u}|\mathbf{z})\, g(\mathbf{z})$. Next, by the change of variables theorem, $p(\mathbf{x}) = \alpha\, p(\mathbf{u}|\mathbf{z})\, g(\mathbf{z})$, where $\alpha$ is the determinant of the Jacobian of the transformation from $\mathbf{x}$ to $\{\mathbf{z}, \mathbf{u}\}$, which can be written $\alpha = \frac{p(\mathbf{x}|H_0)}{p(\mathbf{z},\mathbf{u}|H_0)}$, for

any hypothesis $H_0$. If we now assume that the $p(\mathbf{u}|\mathbf{z})$ we used previously was in fact $p(\mathbf{u}|\mathbf{z};H_0)$, then, (3) follows.

This constructive proof by Kay also shows the *completeness property* of (3), meaning that *all* PDFs that generate $g(\mathbf{x})$ can be sampled in the manner shown and therefore can be written in the form (3). In what follows, we prefer to avoid ancillary statistics or any approach that presumes $T_c(\,)$ is known since it would complicate the proof of the maximum entropy property. Incidentally, the original proof of (3) was done without ancillary statistics [17]. In the course of proving the maximum entropy property in Section 3.4, we will provide another proof of (3) and of completeness.

The method of (3) is called PDF projection because the feature PDF $g(\mathbf{z})$ is *projected* back to the input data space. It can also be called "embedding" since $g(\mathbf{z})$ is embedded in $G_T(\mathbf{x};H_0)$. We call the first term $J_T(\mathbf{z};H_0) = \left[\frac{p(\mathbf{x}|H_0)}{p(\mathbf{z}|H_0)}\right]$ the "J-function" because it reduces to the determinant of the Jacobian matrix for 1:1 transformations and has other interpretations that we will explain.

We now examine the role of Kay's KLD optimality [12] in solving Problem 1. Kay's theorem 3.1 shows that PDF projection minimizes the KLD between $p(\mathbf{x})$ and a target density $p(\mathbf{x};H_1)$ among all PDFs in the family of PDF written

$$p(\mathbf{x}) = Ce^{f(T(\mathbf{x}))}p(\mathbf{x};H_0). \tag{4}$$

We can consider this the family of all PDFs which embed the feature $\mathbf{z} = T(\mathbf{x})$ using reference hypothesis $p(\mathbf{x};H_0)$. This family, not surprisingly, includes PDF projection (3). Kay's result shows that the optimal form of the term $e^{f(T(\mathbf{x}))}$ is $p(\mathbf{z}|H_1)/p(\mathbf{z}|H_0)$, where $p(\mathbf{z}|H_1)$ plays the role of $g(\mathbf{z})$ in PDF projection. But to solve Problem 1, we still need to maximize the entropy over $H_0$. We will see shortly that it is impossible to maximize the entropy over $H_0$ until we add one last missing piece to the puzzle. Thus, the extension of Kay's KLD optimality is only conceptual.

A very instructive example to illustrate this missing piece is the example of the linear transform examined by Kay ([12] Section IV on page 727). Let $\mathbf{z} = \mathbf{A}'\mathbf{x}$, where $\mathbf{A}$ is $N \times p$ and $p < N$, and where $H_0$ is the standard independent Gaussian assumption. While Kay shows that for this feature transformation, PDF projection provides the PDF with minimum KLD to a suitable target PDF, Problem 1 can not be solved. This is because the feature contains insufficient information to limit the size of vector $\mathbf{x}$, allowing it to go to infinity in directions orthogonal to the column space of $\mathbf{A}$, even if it still generates the desired feature PDF on $\mathbf{z}$. Maximizing the entropy will, in fact, drive the variance in the orthogonal subspace to infinity!

Maximizing entropy is analogous to inflating a balloon within a container—the container is the analog of the constraints imposed on the PDF or of moment constraints—the solution will fill all available space within the container. To maximize entropy, we need a leak-proof container otherwise, entropy can go to infinity and there will be no "maximum". The orthogonal subspace in the example is an "energy leak" in the feature transformation, the analog of a hole in the container. We begin to see, at least intuitively, the role of the energy statistic that we introduce in the next section—it serves the same role as a moment constraint. Indeed, we consider the same linear transform in Sections 4.2 and 4.3, but with an *energy statistic* to plug the leak.

In summary, (3) defines a class of PDFs that generate $g(\mathbf{z})$. We will show that all PDFs that generate $g(\mathbf{z})$ can be written in this form and therefore this class of densities contains the maximum entropy density that we seek.

### B. Energy Statistic

*1) Definition of Energy Statistic (ES):* A scalar or multi-dimensional function of the data, denoted by $t(\mathbf{x})$, is an energy statistic if there exists a function $f_t(t)$ satisfying $f_t(t(\mathbf{x})) = \|\mathbf{x}\|$, where $\|\mathbf{x}\|$ is a norm on defined on $\mathcal{R}^N$. Thus, an energy statistic has sufficient information to compute a norm. The feature $\mathbf{z}$ is said to *contain* $t(\mathbf{x})$ if there exists another function $f_z(\mathbf{z})$ such that $f_z(\mathbf{z}) = t(\mathbf{x})$. Therefore, if $\mathbf{z}$ contains an energy statistic, then for fixed $\mathbf{z}$, the size of $\mathbf{x}$ is constrained.

A reference hypothesis corresponding to energy statistic $t(\mathbf{x})$ is a PDF $p(\mathbf{x}|H_0)$ defined on $\mathbf{x} \in \mathcal{R}^N$ that depends on $\mathbf{x}$ only through $t(\mathbf{x})$,

$$p(\mathbf{x}\,|\,H_0) = C \cdot H(t(\mathbf{x})), \tag{5}$$

for some function $H(t)$, where $C$ does not depend on $\mathbf{x}$. We will see that all reference hypotheses that admit a given ES result in the same projected PDF $G_T(\mathbf{x};H_0)$ in (3). Therefore, an ES defines a family of reference hypotheses that are equivalent with respect to (3). They may, however, differ in computation and tractability. As long as the features contain an energy statistic of some sort, we can always define an exponential family density

$$p(\mathbf{x}|H_0) = C(p)e^{-[t(\mathbf{x})]^p},$$

for $p >= 1$.

In most applications, the ES is a useful statistic and is already included, either explicitly or implicitly. In classifier applications where the ES is not wanted because it is deemed detrimental to classification, it's effect can be "neutralized" by assigning the feature a non-informative prior in $g(\mathbf{z})$. The energy statistic will then have no effect on classification as long as all projected PDFs use features that have the same energy statistic. This property was shown for PDF projection in another context where the classification decision could be made invariant to scaling as long as the sample variance was always included in the features ([1], Section II.B). Below, we provide examples of energy statistics and their corresponding canonical reference hypotheses.

*2) Choosing $H_0$ and ES:* The idea "choosing an energy statistic" apparently contradicts Problem 1, in which it is assumed that $T(\mathbf{x})$ and $g(\mathbf{z})$ are given. But we consider the process iterative—you change or add an ES to the feature, then solve Problem 1 with $T(\mathbf{x})$ and $g(\mathbf{z})$ fixed. A primary concern when choosing an ES is the tractability of $p(\mathbf{z}|H_0)$. Distributions have been derived for several useful feature transformations using Gaussian and exponential $H_0$ [18]. But, more work needs to be done to extend this work to additional cases.

That said, the choice of $H_0$ and ES may seem somewhat arbitrary, but it is not. If there exists an ES, the choice of $H_0$ is theoretically irrelevant, so can be made based on tractability

and computation. If $T(\mathbf{x})$ can be modified, and we are free to choose any ES, we choose $H_0$ based on the source and range of the data—usually this means making $p(\mathbf{x} \mid H_0)$ as close as possible to the ideal reference or noise-only condition. This will have the effect of optimizing the approximate sufficiency of the features for the problem at hand. Acoustic or seismic recordings generally call for the Gaussian reference hypothesis. Un-averaged spectral or intensity data call for the exponential reference hypothesis. Data limited to a fixed interval call for the uniform reference hypothesis (no energy statistic is needed). When the choice of $H_0$ is ambiguous, multiple choices can be quantitatively tested by maximum likelihood. The best choice maximizes the total likelihood over $n$ data samples: $L = \sum_{i=1}^{n} \log G_T(\mathbf{x}_i; H_0)$. When the feature PDF $g(\mathbf{z})$ needs to be estimated from training data, it is best to partition the data into training and testing parts.

## III. Mathematical Results

### A. Notation and Terminology

For simplicity, we use the same notation for a random variable and an instance of the random variable. A density is defined by the argument, so that $p(\mathbf{x})$ and $p(\mathbf{z})$ are different functions, understood to be the densities of the random variables $\mathbf{x}$ and $\mathbf{z}$, respectively. To avoid notation conflicts we use $Q$ for entropy instead of $H$.

### B. Entropy Chain Rule

We now return to the problem of maximizing (1) under feature constraints. Consider two jointly-distributed random variables $\mathbf{x}$ and $\mathbf{z}$. The conditional entropy relationship is (see [5], Theorem 2.2.1 on p. 16),

$$Q_{x,z} = Q_z + \int_z Q_{x\,|\,z}(\mathbf{z})\, g(\mathbf{z})\, \mathrm{d}\mathbf{z}, \tag{6}$$

where $Q_{x,z}$ is entropy of the joint density $p(\mathbf{x}, \mathbf{z})$,

$$Q_{x,z} = \int_{\mathbf{x}} \int_{\mathbf{z}} \log\{p(\mathbf{x}, \mathbf{z})\}\, p(\mathbf{x}, \mathbf{z})\, \mathrm{d}\mathbf{z}\, \mathrm{d}\mathbf{x},$$

$Q_z$ is the marginal entropy of $\mathbf{z}$,

$$Q_z = \int_{\mathbf{z}} \log\{p(\mathbf{z})\}\, p(\mathbf{z})\, \mathrm{d}\mathbf{z},$$

and $Q_{x|z}(\mathbf{z})$ is the entropy of $p(\mathbf{x}|\mathbf{z})$,

$$Q_{x|z}(\mathbf{z}) = \int_{\mathbf{x}} \log\{p(\mathbf{x}|\mathbf{z})\}\, p(\mathbf{x}|\mathbf{z})\, \mathrm{d}\mathbf{x}. \tag{7}$$

This is called the chain-rule because it can be applied recursively. In our application, however, $\mathbf{z}$ is completely dependent on $\mathbf{x}$ through (2). This leads to the conclusion that $Q_{x,z} = Q_x$ (see [5], (2.167) on p. 43). Therefore (6) becomes,

$$Q_x = Q_z + \int_z Q_{x\,|\,z}(\mathbf{z})\, g(\mathbf{z})\, \mathrm{d}\mathbf{z}. \tag{8}$$

Applying (8) to $G_T(\mathbf{x})$ decomposes the entropy in terms of $Q_z$, which is given, and the conditional entropy averaged over $\mathbf{z}$. We need only maximize the second conditional entropy term. Note that the conditional PDF $p(\mathbf{x}|\mathbf{z})$ does not exist in the usual sense since all the probability mass is concentrated on a surface, or *manifold*. To evaluate (7), we will need to use manifold integration.

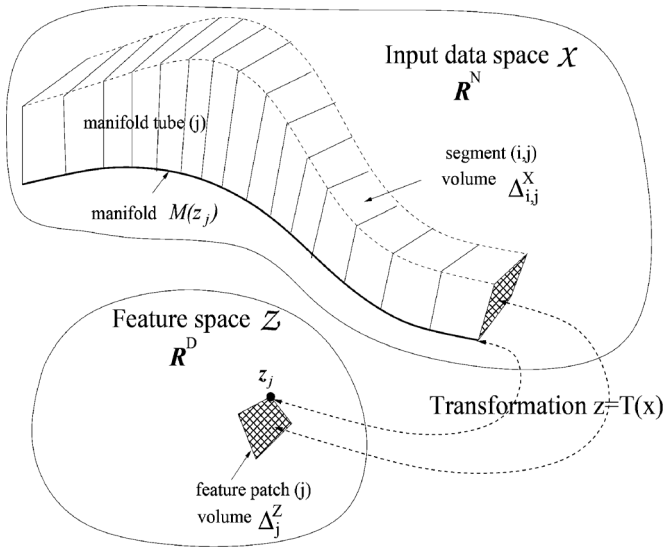


Fig. 1. Geometry of the manifold and manifold tube.

### C. Manifold Integration

We now re-derive (8) in the special case that $\mathbf{z}$ is a function of $\mathbf{x}$. In their book, Cover and Thomas re-affirm the chain-rule in this case for discrete random variables (see [5], (2.168) on p. 43), but do not provide for the continuous case. Relationships involving discrete random variables typically extend to continuous random variables by approximating integrals as discrete Riemann sums. Then, the result is valid for distributions with a finite number of discontinuities (Riemann integrable [19]).

The *manifold* $\mathcal{M}(\mathbf{z})$ is defined as the set of all points $\mathbf{x}$ that map to a given feature value $\mathbf{z}$ through transformation $T(\mathbf{x})$:

$$\mathcal{M}(\mathbf{z}) = \{\mathbf{x} : T(\mathbf{x}) = \mathbf{z}\}. \tag{9}$$

Refer to Fig. 1, which illustrates a manifold and a related construction that we call the manifold "tube". In the figure, a point $\mathbf{z}_j$ in the $D$-dimensional feature space $\mathcal{Z} \in \mathcal{R}^D$ maps to the manifold $\mathcal{M}(\mathbf{z}_j)$ in the $N$-dimensional input space $\mathcal{X} \in \mathcal{R}^N$, represented as a curve in $\mathcal{R}^N$. Now consider a volume patch $j$ in $\mathcal{Z}$ with volume $\Delta^z_j$, one of a countably infinite number of non-overlapping patches that span $\mathcal{Z}$. We assume that $\mathbf{z}_j$ is either inside or on the edge of patch $j$—in the illustration, it is on an edge. The dark curve on the lower edge of the manifold tube represents the manifold $\mathcal{M}(\mathbf{z}_j)$ and the tube itself is the envelope of all points that map to the feature patch $j$. The manifold tube is divided into non-overlapping segments of volume $\Delta^x_{i,j}$. The volume of the tube $j$ is therefore $\Delta^x_j = \sum_i \Delta^x_{i,j}$.

As we have defined the problem, the feature space $\mathcal{Z}$ is spanned by a summation over patch $j$, and the input data space $\mathcal{X}$ is spanned by the double summation over manifold tube $j$ and segment $i$. It follows that the integral of any function $h(\mathbf{x})$ in $\mathcal{X}$ is the limiting form of the double summation

$$\sum_j \sum_i h(\mathbf{x}_{i,j})\, \Delta^x_{i,j} \to \int_{\mathbf{x}} h(\mathbf{x}) \mathrm{d}\mathbf{x} \tag{10}$$

as the volume units $\Delta^x_{i,j}$ tend to zero. We can also write (10) as the double integral

$$\int_{\mathbf{z}} \int_{\mathbf{x}\in\mathcal{M}(\mathbf{z})} h(\mathbf{x})\,\mathrm{d}\mathbf{x} = \int_{\mathbf{x}} h(\mathbf{x})\mathrm{d}\mathbf{x}, \tag{11}$$

where the manifold integral corresponds to the summation over the segments $i$ on the manifold tube. The assumption here is that $h(\mathbf{x})$ is Riemann integrable, and so has a finite number of discontinuities [19]. We will use summations and integrals interchangeably depending on which is most intuitive. Integrating a density, we have

$$\sum_j \sum_i G_T(\mathbf{x}_{i,j})\,\Delta^x_{i,j} \to \int_{\mathbf{z}} \int_{\mathbf{x}\in\mathcal{M}(\mathbf{z})} G_T(\mathbf{x})\,\mathrm{d}\mathbf{x} = 1.$$

We can recover the marginal density $g(\mathbf{z})$ from $G_T(\mathbf{x})$ by integrating the manifold tube for a given patch $j$. If $G_T(\mathbf{x})$ generates $g(\mathbf{z})$, then it follows that the total probability mass in the manifold tube divided by the size of the differential volume in $\mathcal{Z}$ approaches $g(\mathbf{z})$, or

$$\sum_i G_T(\mathbf{x}_{i,j})\,\Delta^x_{i,j} \to g(\mathbf{z}_j)\,\Delta^z_j, \tag{12}$$

or

$$\int_{\mathbf{x}\in\mathcal{M}(\mathbf{z})} G_T(\mathbf{x})\,\mathrm{d}\mathbf{x} = g(\mathbf{z})\,\mathrm{d}\mathbf{z}. \tag{13}$$

### D. The Manifold Distribution

Using the idea of the manifold distribution, we now prove two useful theorems. Since all PDFs that generate $g(\mathbf{z})$ must obey (12), (13), the integral on the manifold is constrained by $g(\mathbf{z})$. The only freedom lies in the choice of the distribution on the manifold. We define the manifold distribution $\mu_z(\mathbf{x})$ using the decomposition

$$G_T(\mathbf{x}) = \mu_z(\mathbf{x})\,g(\mathbf{z}), \tag{14}$$

Integrating a finite-valued manifold distribution, over the manifold, which has zero width, must result in zero. Therefore, we can only integrate it within the framework of the manifold tube, so that it's integral equals the thickness of the tube:

$$\sum_i \mu_z(\mathbf{x}_{i,j})\,\Delta^x_{i,j} = \Delta^z_j, \tag{15}$$

or

$$\int_{\mathbf{x}\in\mathcal{M}(\mathbf{z})} \mu_z(\mathbf{x})\,\mathrm{d}\mathbf{x} = \mathrm{d}\mathbf{z}, \tag{16}$$

which can be verified by substituting (14) into (13). Also, comparing (14) with (3), we see that the J-function and the manifold density are the same: $\mu_z(\mathbf{x};H_0) = \left[\frac{p(\mathbf{x}|H_0)}{p(\mathbf{z}|H_0)}\right]$.

We can now prove the following completeness theorem, which was shown by S. Kay using ancillary statistics (See Section 2.1).

*Theorem 1: Completeness of PDF Projection Method for Specified Marginal Density:* (S. Kay) Any PDF that generates a given feature density $g(\mathbf{z})$ through a given dimension-reducing transformation $\mathbf{z} = T(\mathbf{x})$ may be written as a PDF projection (3) for some reference hypothesis $H_0$.

*Proof:* First, any manifold density $\mu_z(\mathbf{x})$ can be written as a J-function using an arbitrary feature density $h(\mathbf{z})$. To show this, form the function $H(\mathbf{x}) = \mu_z(\mathbf{x})\,h(\mathbf{z})$. It is easy to show using the double integral just presented that $H(\mathbf{x})$ integrates to 1, so is a density. Therefore, $\mu_z(\mathbf{x}) = \left[\frac{H(\mathbf{x})}{h(\mathbf{z})}\right]$ is a J-function where hypothesis $p(\mathbf{x};H_0)$ is identified with $H(\mathbf{x})$ and $p(\mathbf{z};H_0)$ is identified with $h(\mathbf{z})$. Since all PDFs that generate $g(\mathbf{z})$ can be put in form (14), and any manifold density $\mu_z(\mathbf{x})$ can be written as a J-function, it follows that all PDFs that generate $g(\mathbf{z})$ can be created using PDF projection (3) for some $H_0$.

The following theorem re-states the PDF projection theorem in light of the current results.

*Theorem 2: PDF Projection Theorem:* Consider any PDF pair $H(\mathbf{x}), h(\mathbf{z})$ where $h(\mathbf{z})$ is the marginal of $H(\mathbf{x})$ generated through transformation $\mathbf{z} = T(\mathbf{x})$. Then $H(\mathbf{x})$ becomes a PDF that generates $g(\mathbf{z})$ when multiplied by $g(\mathbf{z})/h(\mathbf{z})$.

*Proof:* Let

$$I(\mathbf{x}) = H(\mathbf{x})\frac{g(\mathbf{z})}{h(\mathbf{z})}.$$

Using (12),

$$\begin{aligned}\sum_i I(\mathbf{x}_{i,j})\,\Delta^x_{i,j} &= \frac{g(\mathbf{z}_j)}{h(\mathbf{z}_j)} \sum_i H(\mathbf{x}_{ij})\,\Delta^x_{i,j} \\ &= \frac{g(\mathbf{z}_j)}{h(\mathbf{z}_j)}\,h(\mathbf{z}_j)\,\Delta^z_j \\ &= g(\mathbf{z}_j)\,\Delta^z_j,\end{aligned} \tag{17}$$

or

$$\int_{\mathbf{x}\in\mathcal{M}(\mathbf{z})} I(\mathbf{x})\,\mathrm{d}\mathbf{x} = g(\mathbf{z})\,\mathrm{d}\mathbf{z}, \tag{18}$$

proving that $I(\mathbf{x})$ generates $g(\mathbf{z})$. Also, integrating (18) over $\mathbf{z}$ gives 1, which shows that $I(\mathbf{x})$ is a density. This is an alternate proof of the PPT (3), where $H(\mathbf{x})$ is identified with $p(\mathbf{x}|H_0)$, and $h(\mathbf{z})$ is identified with $p(\mathbf{z}|H_0)$.

### E. Entropy Maximization

We return to the problem of maximizing entropy and re-derive the entropy chain rule in the process. Let $\theta$ index all the densities $G_T(\mathbf{x};\theta)$ in the class of densities that generate $G(\mathbf{z})$ through $\mathbf{z} = T(\mathbf{x})$. Our goal is to maximize

$$Q_x(\theta) = -\int_{\mathbf{x}} \log\{G_T(\mathbf{x};\theta)\}\,G_T(\mathbf{x};\theta)\,\mathrm{d}\mathbf{x}$$

over $\theta$. We first re-write this equation as a sum of two terms. We can write

$$Q_x(\theta) = -\int_{\mathbf{x}} \log\left\{g(\mathbf{z})\,\frac{G_T(\mathbf{x};\theta)}{g(\mathbf{z})}\right\}\,\frac{G_T(\mathbf{x};\theta)}{g(\mathbf{z})}\,g(\mathbf{z})\,\mathrm{d}\mathbf{x},$$

which forms the two terms $Q_x(\theta) = Q_z(\theta) + Q_\mu(\theta)$, where

$$Q_z(\theta) = -\int_{\mathbf{x}} \log\{g(\mathbf{z})\}\;G_T(\mathbf{x};\theta)\,\mathrm{d}\mathbf{x}, \tag{19}$$

and

$$Q_\mu(\theta) = -\int_{\mathbf{x}} \log\left\{\frac{G_T(\mathbf{x};\theta)}{g(\mathbf{z})}\right\}\,\left\{\frac{G_T(\mathbf{x};\theta)}{g(\mathbf{z})}\right\}\,g(\mathbf{z})\,\mathrm{d}\mathbf{x}. \tag{20}$$

We now reduce (19) using (11). Since $\log\{g(\mathbf{z})\}$ is independent of $\mathbf{x}$ for all $\mathbf{x} : T(\mathbf{x}) = \mathbf{z}$, we bring the manifold integral to the inside, obtaining

$$Q_z(\theta) = -\int_{\mathbf{z}} \log\{g(\mathbf{z})\} \int_{\mathbf{x}\in\mathcal{M}(\mathbf{z})} G_T(\mathbf{x};\theta)\,\mathrm{d}\mathbf{x},$$

which from (13), gives the term independent of $\theta$

$$Q_z = -\int_{\mathbf{z}} \log\{g(\mathbf{z})\}\; g(\mathbf{z})\,\mathrm{d}\mathbf{z}, \tag{21}$$

which is just $Q_z$ in (8). Using (14), we substitute $\frac{G_T(\mathbf{x};\theta)}{g(\mathbf{z})} = \mu_z(\mathbf{z};\theta)$ into (20), and write in summation form

$$Q_\mu(\theta) \simeq -\sum_{i,j} \log\{\mu_z(\mathbf{x}_{i,j};\theta)\}\; \mu_z(\mathbf{x}_{i,j};\theta)\, g(\mathbf{z}_j)\, \Delta^x_{i,j}. \tag{22}$$

We can re-write this as $Q_\mu(\theta) \simeq \sum_j Q_\mu(j,\theta)\, g(\mathbf{z}_j)\, \Delta^z_j$, where

$$Q_\mu(j,\theta) = -\sum_i \log\{\mu_z(\mathbf{x}_{i,j};\theta)\}\, \mu_z(\mathbf{x}_{i,j};\theta)\, \frac{\Delta^x_{i,j}}{\Delta^z_j}, \tag{23}$$

which in integral form becomes

$$Q_\mu(\mathbf{z};\theta)\,\mathrm{d}\mathbf{z} = -\int_{\mathbf{x}\in\mathcal{M}(\mathbf{z})} \log\{\mu_z(\mathbf{x};\theta)\}\; \mu_z(\mathbf{x};\theta)\,\mathrm{d}\mathbf{x}, \tag{24}$$

which is the entropy of the manifold density $\mu_z(\mathbf{x};\theta)$ at a fixed value of $\mathbf{z}$. Therefore, $Q_\mu(\theta)$ is just the expected value of the manifold entropy, with expectation over $g(\mathbf{z})$. The complete formula for the entropy of $G_T(\mathbf{x};\theta)$ is therefore

$$Q_x(\theta) = Q_z + \int_{\mathbf{z}} Q_\mu(\mathbf{z};\theta)\, g(\mathbf{z})\,\mathrm{d}\mathbf{z}. \tag{25}$$

This is the special form of (8) that we seek.

### F. Applying the Energy Statistic

Equation (25) has the interpretation as the expected value of the manifold entropy, with expectation over $g(\mathbf{z})$. The value of $\theta$ that maximizes the manifold entropy $Q_\mu(\mathbf{z};\theta)$ at a fixed $\mathbf{z}$ will in general depend on $\mathbf{z}$, so the overall optimal value of $\theta$ will depend on $g(\mathbf{z})$. This prevents us from obtaining a general expression and points to numerical maximization, a futile effort except in very low-dimensional problems. The only hope of finding a general solution is if one choice of $\theta$ maximizes the manifold entropy $Q_\mu(\mathbf{z};\theta)$ for *all* $\mathbf{z}$. We now show that this happens if $\mathbf{z}$ contains an energy statistic, concluding the proof of our main theorem.

*Theorem 3: Maximum Entropy Property of PDF Projection Using Energy Statistic:* Let feature $\mathbf{z}$ be obtained by dimension-reducing transformation of the input data $\mathbf{x}$ according to (2), where $\mathbf{z}$ contains an energy statistic $t(\mathbf{x})$ with corresponding reference hypothesis $p(\mathbf{x}|H_0)$ as defined in Section 2.2. Let $g(\mathbf{z})$ be an arbitrary PDF defined on the feature space. Then the PPT density (3) has the highest entropy among all PDFs that generate feature PDF $g(\mathbf{z})$ through transformation (2).

*Proof:* If $\mathbf{z}$ contains an ES for $H_0$ then, as a result of (3) and (5), the manifold density in (24) is constant on the manifold. In other words,

$$\mu_z(\mathbf{x};\theta) = f(\mathbf{z}) \quad \forall\, \mathbf{x} \text{ s.t. } T(\mathbf{x}) = \mathbf{z},$$

meaning that the manifold distribution is the *uniform density*. Incidentally, any reference hypothesis admitting the same ES will result in a J-function that is constant on the manifold and produce the same manifold density, thus the same projected PDF. Furthermore, the norm-forming property of the ES forces the manifold to be bounded for a fixed $\mathbf{z}$. And, over bounded regions, the uniform density has the maximum entropy [3], [4]. Thus, $Q_\mu(\mathbf{z};\theta)$ is maximized for all $\mathbf{z}$ when $\mathbf{z}$ contains an ES, and therefore $Q_\mu(\theta)$ is maximized.

Intuitively, for any PDF that generates $g(\mathbf{z})$, the total probability mass in the manifold tube is already specified by $g(\mathbf{z})$. All we can do is specify the manifold distribution. The entropy argument says that with lack of any other information, and assuming the manifold is bounded, we should make the density uniform in the tube, or constant on any manifold. All we need is to find one reference PDF that is constant on the manifold because any density that is constant on the manifold can be converted into a PDF that generates $g(\mathbf{z})$ through the method of PDF projection. Fortunately, we've discovered that if the feature contains an ES, both of these requirements are met simultaneously: the boundedness of the manifold is guaranteed, and the there is a reference hypothesis that is constant on the manifold. We have only shown sufficiency—whether the ES is necessary has not been established.

## IV. Building Blocks

Below, we provide simple feature transformations and associated reference hypothesis and ES. We will use these building blocks in cascade to construct more complex feature transformations using the chain rule.

### A. Magnitude Squared FFT Bins

This building block considers the FFT followed by the magnitude-squared of the bins. Let $N$ be even and

$$z_k = \left|\sum_{i=0}^{N-1} x_i\, e^{-j2\pi(k-1)(i-1)/N}\right|^2, \quad 1 \le k \le N/2+1.$$

This building block is covered in detail in [20], page 47, Section D.1. The densities $p(\mathbf{x}|H_0)$ and $p(\mathbf{z}|H_0)$ are shown on page 48, first column (numerator and denominator PDFs). The log J-function is just $\log J = \log p(\mathbf{x}|H_0) - \log p(\mathbf{z}|H_0)$. These densities are separately provided in the reference, but log J-function can be simplified for even $N$ to

$$\log J = \frac{\log z_1 + \log z_{N/2+1}}{2} + \frac{N}{2}\log N - \frac{N-2}{2}\log(2\pi),$$

which interestingly is data independent except with respect to the zero and Nyquist frequency bins. The reference hypothesis is Gaussian (Table I), and the ES is contained implicitly (Parseval's theorem).

### B. Linear Transform (Exponential $H_0$)

This building block considers the general linear transformations of spectral or intensity data, which is widely used in signal and image processing and spectral analysis. Despite the simplicity of the feature transformation, the positive-valued input data makes the problem more challenging. The ES is the sum of the input samples. The most widely-used application of this

feature is in spectral and intensity image classification and analysis.

Let $\mathbf{z} = \mathbf{A}'\mathbf{x}$, where $\mathbf{A}$ is a non-singular $N \times D$ matrix. We use the exponential reference hypothesis and ES from Table I. In order that $\mathbf{z}$ contain the ES, we need that $\mathbf{1} = \mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\mathbf{1}$, where $\mathbf{1} = [1, 1, 1, \ldots 1]'$. The feature PDF $p(\mathbf{z}|H_0)$ is not available in closed form, but the moment generating function is available. Therefore, the saddle-point approximation can be used. Details may be found in Kay *et al.* [18], Sections III.A–III.C, on pages 2243–2246. This approximation is accurate, even in the tails and can be used in place of an exact formula.

### C. Linear Transform (Gaussian $H_0$)

This building-block considers the linear reduction of Gaussian-like data. This type of input data includes measured acoustic or seismic data that is not constrained to be positive, or spectral or intensity data that is averaged or non-linearly transformed (i.e., the logarithm), or both. The ES is formed from the sum of the squares of the input samples. Applications are very wide including principal component analysis and linear filtering.

Let $\mathbf{z}_A = \mathbf{A}'\mathbf{x}$, where $\mathbf{A}$ is any non-singular $N \times D$ matrix. The feature $\mathbf{z}$ is the union of $\mathbf{z}_A$ with $t(\mathbf{x}) = \sum_{i=1}^{N} x_i^2$, and $\mathbf{z} = [t(\mathbf{x}), \mathbf{z}_A]$. Despite the apparent similarity to the previous problem, the problem is quite different as a result of using a different reference hypothesis and ES. Using the Gaussian reference hypothesis (Table I), $t(\mathbf{x})$ is the ES.

The PDF $p(\mathbf{z}|H_0)$ can be easily written down. We must first orthogonalize the features. Let $\rho = \mathbf{x}'\left\{\mathbf{I} - \mathbf{A}\left(\mathbf{A}'\mathbf{A}\right)^{-1}\mathbf{A}'\right\}\mathbf{x}$, which can also be written $\rho = t(\mathbf{x}) - \mathbf{x}'\mathbf{A}\left(\mathbf{A}'\mathbf{A}\right)^{-1}\mathbf{A}'\mathbf{x}$. This is the energy in $\mathbf{x}$ orthogonal to the columns of $\mathbf{A}$. Accordingly, $\rho$ is statistically independent of $\mathbf{z}_A$ under $H_0$. Thus, $p(\rho, \mathbf{z}_A|H_0) = p(\rho|H_0)\, p(\mathbf{z}_A|H_0)$. Also, $p(\rho|H_0)$ is chi-squared with $N - D$ degrees of freedom,

$$p(\rho \mid H_0) = \frac{\rho^{(\kappa/2-1)}\, e^{-\rho/2}}{2^{\kappa/2}\, \Gamma(\kappa/2)},$$

where $\kappa = N - D$ and $p(\mathbf{z}_A|H_0)$ is the Gaussian distribution with mean $\mathbf{0}$ and co-variance $\mathbf{A}'\mathbf{A}$:

$$p(\mathbf{z}_A|H_0) = (2\pi)^{-\frac{D}{2}}\, |\mathbf{A}'\mathbf{A}|^{-1/2} e^{-\frac{1}{2}\mathbf{z}_A'\left(\mathbf{A}'\mathbf{A}\right)^{-1}\mathbf{z}_A}$$

And, since $(\rho, \mathbf{z}_A)$ can be obtained from $\mathbf{z}$ using a linear transformation with Jacobian of determinant 1, we can write $p(\mathbf{z}|H_0) = p(\rho|H_0)\, p(\mathbf{z}_A|H_0)$.

## V. Feature Chain Rule

We now show how to cascade the above building blocks into more complex feature transformations using the chain-rule. Consider the two-stage feature extraction process $\mathbf{w} = T(\mathbf{x})$, followed by $\mathbf{z} = U(\mathbf{w})$. Suppose we are given the feature density $g(\mathbf{z})$. Then, the chain-rule form of (3) is

$$G_{TU}(\mathbf{x}) = \frac{p(\mathbf{x}|H_{0x})}{p(\mathbf{w}|H_{0x})}\, \frac{p(\mathbf{w}|H_{0w})}{p(\mathbf{z}|H_{0w})} g(\mathbf{z}), \tag{26}$$

where $H_{0x}$ and $H_{0w}$ are reference hypotheses for $\mathbf{x}$ and $\mathbf{w}$, respectively. The chain-rule extends to an arbitrary number of stages. Clearly, if $H_{0x}$ and $H_{0w}$ are the same, this simplifies to (3). In general, however, it may be extremely difficult to solve for $p(\mathbf{z}|H_0)$ since under a fixed $H_0$, the feature distributions become increasingly complicated after each stage. This is greatly simplified by changing reference hypotheses at each stage.

Because of the entropy chain-rule, chaining the feature extraction transformations carries forward the maximum entropy property. In other words, $G_{TU}(\mathbf{x})$ in (26) is the maximum entropy density over all densities that generate $g(\mathbf{z})$ through the combined transformation. This is evident from the expression (25) where it does not matter that $g(\mathbf{z})$ is itself a projected PDF.

### A. Chain Rule Example—MFCC

A good example of the utility of the chain-rule is the calculation of MEL frequency cepstral coefficients (MFCC) [21]. Let $\mathbf{x}$ be a length-$N$ time-series segment. We break the MFCC extraction process into four stages:

1. FFT/magnitude-squared. The feature $\mathbf{y}$ is the length $N/2+1$ magnitude-squared spectrum. We apply the FFT building block in Section 4.1.
2. MEL band analysis. Let $\mathbf{y}$ be the magnitude-squared FFT bins of length $n = N/2 + 1$. Let
$$\mathbf{w} = \mathbf{A}'\mathbf{y} \tag{27}$$
be the MEL band energies, where $\mathbf{A}$ is the $n \times K$ matrix of MEL spectral band functions. As a building block for this stage, we use the exponential example in Section 4.2. Since the ES is the sum of the input samples, we need the columns of $\mathbf{A}$ add to a constant—true for the MEL band functions that include the zero and Nyquist bands.
3. Log. Let $\mathbf{u} = \log(\mathbf{w})$ be the element-wise log function. The J-function of this invertible transformation is the determinant of the Jacobian, $J(\mathbf{w}) = \frac{1}{\prod_{i=1}^{K} w_i}$.
4. DCT. Let $\mathbf{z} = [\mathbf{z}_B,\, t(\mathbf{u})]'$, where $\mathbf{z}_B = \mathbf{B}'\mathbf{u}$ computes the lower $D$ DCT bins, and $t(\mathbf{u}) = \sum_{i=1}^{K} u_i^2$. There are two possibilities.
   a) If $D < K$, we use the Gaussian building block in Section 4.3.
   b) If $D = K$, the DCT step is 1:1, invertible, and unitary, so $J(\mathbf{u}) = 1$. There is no energy statistic so $\mathbf{z} = \mathbf{z}_B$.

### B. Chain Rule Example—AR

Auto-regressive (AR) analysis is a widely-used spectral estimation and time-series analysis method. We use the frequency-domain method starting with computing the magnitude-squared bins of the FFT of the input data, then the auto-correlation function (ACF) by inverse FFT.

1. FFT/magnitude squared. The FFT stage has been covered in the MFCC chain in Section 5.1.
2. Auto-correlation (ACF). Calculation of the ACF from the magnitude-squared FFT bins is accomplished by linear transform. This is another application of the building-block in Section 4.2. However, there is a slight anomaly caused by the FFT edge bins that needs attention. Assuming for

the moment that $\mathbf{y}$ is extended to length $N$ by replicating the redundant FFT bins, we use the inverse FFT

$$z_k = \frac{1}{N^2}\sum_{i=1}^{N} y_i \cos\{2\pi(i-1)(k-1)/N\},$$

for $1 \le k \le P+1$. This computes the order-$P$ circular ACF using the frequency-domain method. The ES is the zero lag output $z_1$. In matrix form, we write $\mathbf{z} = \mathbf{C}'\mathbf{y}$ where $\mathbf{C}$ is the $n \times (P+1)$ ACF matrix, where $n = \left(\frac{N}{2}+1\right)$, and $\mathbf{y}$ has be collapsed down to it's original size. Note that because the redundant bins are used twice, the elements in $\mathbf{C}$ in all rows except the first and last, are multiplied by 2. This implies that the ES is equal to $z_1 = \mathbf{c}_0'\mathbf{y}$, where $\mathbf{c}_0 = [1\,2\,2\,\cdots\,2\,1]$. The reference exponential hypothesis (Table I) needs to be slightly modified to be a function of $z_1$,

$$p(\mathbf{y}|H_{0y}) = e^{-y_1}\, e^{-y_n} \prod_{i=2}^{n-1} e^{-2y_i}. \tag{28}$$

The J-function for this stage is $J(\mathbf{y}; H_{0y}) = p(\mathbf{y}|H_{0y})/p(\mathbf{z}|H_{0y})$.

3. Reflection coefficients. The Levinson algorithm and the log-Bilinear transformation, are invertible transformations and are detailed in [20], page 49, Sections VI.D.3 and VI.D.4.

## VI. Simulations

We test both of the models proposed in Sections 5.1 and 5.2 using simulated data, then use the models in a classification experiment which is compared with the optimal Neyman-Pearson classifier.

### A. Experimental Approach

Let $p(\mathbf{x}|H_a)$ stand for a theoretical PDF from which we can generate an unlimited amount of data. We use training data to estimate the feature PDF $\hat{g}(\mathbf{z})$, then form the projected PDF estimate $\log \hat{G}(\mathbf{x}) = \log J(\mathbf{x}; H_0) + \log \hat{g}(\mathbf{z})$, which we compare with $\log p(\mathbf{x}|H_a)$. If the feature extraction is based on a chain of building-blocks, then $\log J(\mathbf{x}; H_0)$ is the accumulation of the building-block log J-functions.

When another theoretical class $\log p(\mathbf{x}|H_b)$ is introduced, we can determine the classification performance of the optimal Neyman-Pearson classifier: $\arg\max_k\{p(\mathbf{x}|H_k)\}$, which can be compared with the performance of the classifier that uses the projected PDFs in place of $p(\mathbf{x}|H_k)$.

### B. Circular Power Spectral Models

When working with length-$N$ time-series, it is convenient to define the spectral and PDF models in the frequency-domain using the DFT. Without using a windowing function, these spectral models assume circular continuity and will only be approximations to spectral models based on stationary processes that are infinitely long in theory. But, they will be exact PDF models from which data can be easily generated. Methods exist to extend our results to windowed data, but are outside the scope of this paper [22].

We define a circular spectral model by the circular power spectrum defined by,

$$\mathcal{E}\{|X_k|^2\} = N\rho_k, \quad 1 \le k \le N, \tag{29}$$

where $X_k$ are the DFT coefficients of the length-$N$ input data $\mathbf{x}$. The PDF of $\mathbf{x}$ for a circular power spectral process is written in the frequency domain

$$\log p(\mathbf{x}; \boldsymbol{\rho}) = -\frac{1}{2}\sum_{k=1}^{N}\left\{\log 2\pi\rho_k + \frac{|X_k|^2}{N\rho_i}\right\}. \tag{30}$$

Although written using the DFT coefficients $X_k$, it is a PDF defined on $\mathbf{x}$. Given a circular power spectrum $\boldsymbol{\rho}$, we can generate data by generating complex FFT output bins with the specified power spectrum, then inverting the FFT to obtain a time-series. Using the definition (29), we may generate random data at the DFT output as

$$X_k = \sqrt{N\rho_k}\, u_k, \tag{31}$$

where $u_k$ is a Gaussian random variable with mean zero and variance 1. When $X_k$ is the zero or Nyquist bin, $u_k$ is real, otherwise, it is a complex Gaussian random variable. We consider two spectral models, the auto-regressive (AR) and MFCC models.

*1) AR Circular Power Spectral Model:* An order-$P$ auto-regressive (AR) process is defined by the innovation variance $\sigma^2$ and the AR coefficients $a_1, a_2 \ldots a_P$. The circular power spectrum of a circular AR process is given by

$$\rho_k = \frac{\sigma^2}{|A_k|^2}, \tag{32}$$

where $\{A_k\}$, is the DFT of the AR parameters zero-padded to length $N$ : $\mathrm{DFT}([1, -a_1, -a_2 \ldots - a_P, 0, 0 \ldots])$.

*2) MFCC Circular Power Spectral Model:* We now construct an "MFCC-like" data model. The structure of MFCC is defined in the frequency-domain by the MEL band functions. Using matrix $\mathbf{A}$ is defined in (27), we define the circular power spectrum of an MFCC-like process as

$$\boldsymbol{\rho}_w = \mathbf{A}\mathbf{b}, \tag{33}$$

where $\mathbf{b}$ is a $K \times 1$ vector of positive values.

### C. Simulation Results

We used a data size of $N = 256$ in our simulations and an AR feature model order of $P = 2$ (matching the data generation) and $K = D = 8$ for MFCC, also matching the data generation. The MFCC spectral model we used is specified by (33), where $\mathbf{b}$ is given in Table II and the columns of $\mathbf{A}$ are the MEL band functions (Hanning-shaped MEL band functions calculated for a sample rate of 16000 Hz). To make the simulations more "interesting", we made the AR and MFCC model spectra as similar as possible by approximating the MFCC power spectrum with an AR model. Thus, we created the AR model by inverting the MEL cepstrum to obtain the ACF, then used the Levinson algorithm to obtain the following AR(2) model $a_1 = -0.04101841$, $a_2 = -0.55743570$. These coefficients were used to create the circular AR models (32). The AR and MFCC-like spectra are

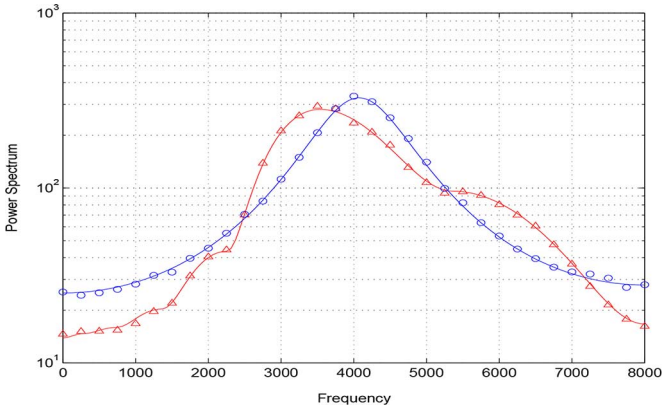


Fig. 2. Power spectra, theoretical (smooth curves) and estimated (symbols). Circles: AR, Triangles: MFCC. Both spectra produce the same AR features.

TABLE II
COEFFICIENTS THAT DEFINE THE MFCC-LIKE MODEL USED IN THE SIMULATIONS

| Coef | Value | Coef | Value |
|---|---|---|---|
| $b_1$ | 6.940937 | $b_5$ | 21.734748 |
| $b_2$ | 7.371341 | $b_6$ | 140.720039 |
| $b_3$ | 7.981074 | $b_7$ | 47.916180 |
| $b_4$ | 10.167393 | $b_8$ | 8.350037 |

plotted in Fig. 2, which also shows the sample mean power spectra from 1000 samples of each model.

We now test the PDF approximation accuracy of the PDF projection. Using 1000 samples of data, we computed features and estimated the feature PDFs $\hat{g}(\mathbf{z})$ using Gaussian mixture approximation with three mixture components. Then, for 1000 samples of independent testing data, we plotted the PDF estimation error $L_e = \log \hat{G}(\mathbf{x}) - \log p(\mathbf{x}|H_a)$. We used feature extraction chain described in Section 5.2 for the AR assumption and that described in Section 5.1 for the MFCC assumption (with $D = K$). Results are shown in Fig. 3 which demonstrates a very important result. All data points displayed on the left side of the figure were generated using the circular AR model. On the X-axis is the theoretical PDF value of each sample, given by (30) and (32). On the Y-axis is the PDF approximation error $L_e$, We only displayed 300 of the 1000 samples for clarity. In the ideal situation, most of the points would lie near zero on the Y-axis. Important to note is that we attempted the approximation using both the AR-based features (dots) and the MFCC-based features (circles). It is clear from the left graph in the figure that PDF projection using the AR-based features chain, described in Section 5.2 produces a much better PDF estimate than PDF projection using the MFCC-based chain.

Similarly, it can be seen on the right side of the figure that the MFCC-based features (Section 5.1) produce a better PDF estimate than the AR features.

Although better, the MFCC-based PDF estimate has still more error than the AR model on the left graph. We can improve the MFCC features if we recall from spectral estimation fundamentals that the AR features are solutions to the Yule-Walker equations, which are a form of maximum-likelihood (ML) estimation [23]. To obtain an improved MFCC-like

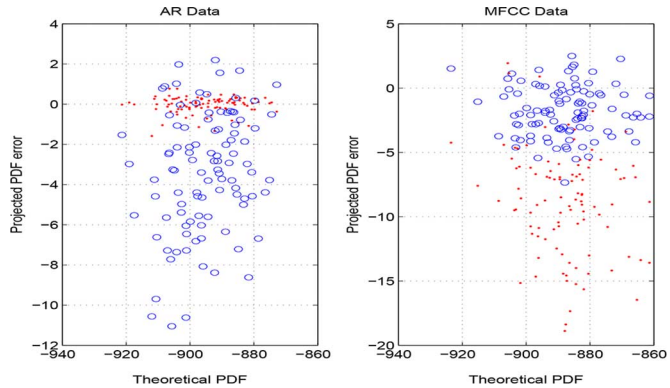


Fig. 3. Comparison of projected PDF with theoretical for AR data (left) and MFCC data (right). Circles: MFCC features, Dots: AR features.

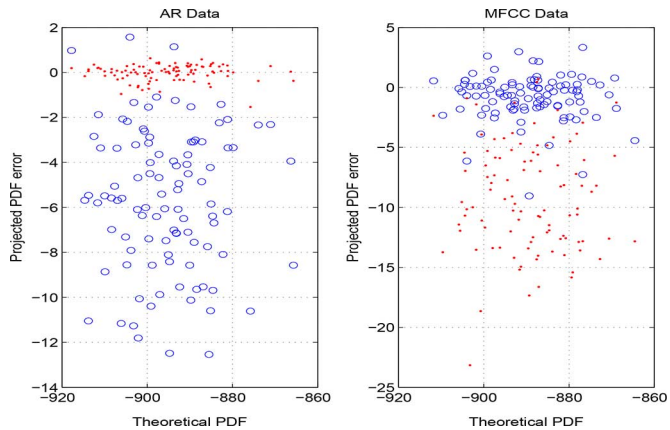


Fig. 4. Re-generation of Fig. 3 using MFCC-ML features.

feature, we need an ML approach. The parametric model for the MFCC-like data is defined by (30) and (33). Our approach is to maximize (30) over $\mathbf{b}$ for each data sample $\mathbf{x}$. Then, $\mathbf{w} = \mathbf{A}'\mathbf{A}\mathbf{b}$, and the feature $\mathbf{z}$ is determined in the usual way from $\mathbf{w}$ using a 1:1 transformation. This maximizing value, denoted by $\mathbf{z}_{\mathrm{ML}}(\mathbf{x})$, is an iteratively-determined feature. The PDF projection method for iteratively-determined ML features is given in [1], page 675, Section II.C, (12), and page 677, Section III.C. The associated J-function is given by

$$J(\mathbf{x}) = \frac{p(\mathbf{x};\hat{\boldsymbol{\theta}})}{(2\pi)^{-p/2}|\mathbf{I}(\hat{\boldsymbol{\theta}})|^{1/2}}, \tag{34}$$

where $\mathbf{I}(\boldsymbol{\theta})$ is Fisher's information matrix. Using this new MFCC feature, called MFCC-ML, we re-generated Fig. 3. The result is shown in Fig. 4. Now we see that the MFCC feature attains even better PDF approximation accuracy. This demonstrates the importance of choosing the correct features when applying the PDF projection method and the power of PDF projection to find weak features before they are introduced into a classifier.

To demonstrate the effectiveness of the method in classification, we conducted an experiment based on data from both theoretical models. We first evaluated the optimal performance using the theoretical Neyman-Pearson classifier constructed using the theoretical PDFs using (30) and the two circular PDF models (32) and (33). Fig. 5(left), shows the theoretical AR model log

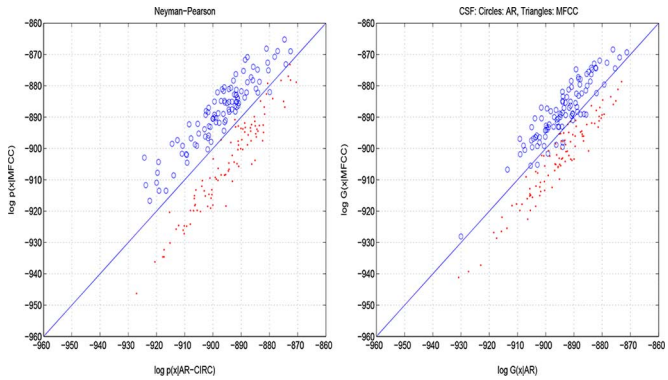


Fig. 5. One hundred generated data samples from each model. The log-likelihood of each sample is displayed for each model assumption (AR on X axis and MFCC on Y axis). Circles: MFCC data, Dots: AR data. Left: using theoretical PDFs. Right: using PDF projection.

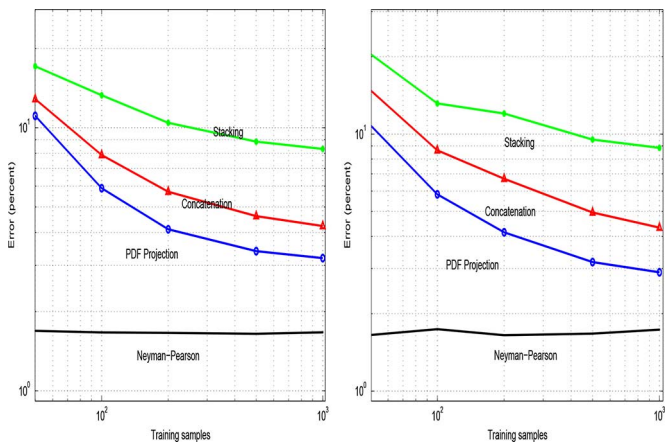


Fig. 6. Classification performance as a function of training set size for 80 000 testing samples. Using MFCC feature (left), using MFCC-ML feature (right).

likelihood on the X-axis and the theoretical MFCC log likelihood on the Y-axis for 100 samples each of MFCC data (circles) and AR data (dots). A few errors can be seen. The Optimal classification error probability was determined to be 1.68% using 80 000 test samples. Fig. 5(right) shows the experiment repeated using the projected PDFs using AR and MFCC features. It is difficult to see a difference between the projected and theoretical values. To obtain a more quantitative result, we need to measure error probability in trials.

Next, we re-ran the experiment using a variety of training sample sizes, measuring classification performance. We compared the method with (a) Neyman-Pearson (optimal) classifier, (b) additive combination of the AR and MFCC feature log-likelihood functions, sometimes called "stacking", and (c) feature concatenation in which the union of the AR and MFCC features was formed. The same feature PDF estimation approach was used as for the feature density $\hat{g}(\mathbf{z})$ in PDF projection. We ran the experiments using both MFCC and MFCC-ML features. The results are shown in Fig. 6 which shows the classification error probability in percent. For the left graph we used MFCC features, and for the right graph MFCC-ML features. After the optimal Neyman-Pearson classifier, PDF projection was best over-all, with MFCC-ML slightly better than MFCC. For MFCC features, feature concatenation showed about 35% more errors than PDF projection. For MFCC-ML, that ratio went up to about 50%. Likelihood stacking did much worse than feature concatenation, indicating that feature concatenation took advantage of the statistical dependence between the ML and AR features.

## VII. Summary and Conclusions

### A. Summary of Paper

We have proved mathematically that under certain conditions, the PDF projection method produces the maximum entropy PDF among all PDFs that generate the given feature PDF. This requires that the feature contain an "energy" statistic for the chosen reference hypothesis. We have also described how the chain-rule may be used to construct maximum entropy PDFs based on chains of feature transformations and have provided examples of two widely-used feature transformations.

We've also presented simulation results to verify many of our claims. It is not possible to verify the maximum entropy principle empirically. Thus, the theoretical derivations in Section 3 must stand on their own. But, we have verified in carefully controlled simulations the PDF approximation accuracy of the feature extraction chains presented in Sections 5.2 and 5.1 against the theoretical PDF. Furthermore, we've shown that when the features can be individually matched to the data generation process of each class, a generative classifier constructed on the raw data space using class-dependent features and PDF projection can perform better than traditional generative classifiers constructed on the feature space.

### B. Interpretation of the J-Function

The J-function is the only physical difference between PDF projection and the competing methods we tested, so it clearly contains useful information. The J-function is a measure of the ability of the features to describe the input data. Mathematically, the J-function is the same as the manifold density, which is a uniform density on the manifold when an ES is included. The manifold is a range of input data values that map to a given feature value. So, if the features are very descriptive, and accurately describe the peculiarities of the given data sample, the range of possible input data values shrinks, increasing the value of the uniform density. Another interpretation, based on asymptotic maximum likelihood (ML) theory, starts by assuming that there exists some parametric model $p(\mathbf{x};\boldsymbol{\theta})$ such that the features are maximum likelihood estimates of the parameters, $\mathbf{z} = \hat{\boldsymbol{\theta}}$. The J-function for ML, given in (34), is dominated by the numerator, which is the likelihood function of the data evaluated at $\hat{\boldsymbol{\theta}}$. Thus, the J-function has the interpretation as a quantitative measure of how well the parametric model can describe the raw data. The better the features, the better this notional parametric model. Interestingly, because the J-function can be computed without actually implementing the ML estimator, this information is available without needing to know the parametric form nor needing to maximize it! Naturally, there are situations where this information is detrimental to classification—specifically if the data contains nuisance information or interference. There are work-arounds that significantly improve classification performance, for example the class-specific feature mixture ([24],

Section II.B). Since all features are available for all class assumptions, it effectively allows the data to "choose" the feature.

### C. Looking Forward

PDF projection is a new idea that has the potential to revitalize the concept of the generative classifier, which has been nudged aside by leading-edge discriminative methods in recent years. The proof of the maximum entropy property should make the method attractive to researchers in coming years. The fact that PDF projection has been demonstrated to perform better as a classifier under controlled conditions should also invoke interest. But, challenges exist to establish the PDF projection method in real data applications. These include

1. The classification scheme we presented makes the one-class, one-feature assumption, a bad assumption in real-data problems. We recommend the more realistic view that each class is a *mixture* of models by modeling the data PDF as a kernel mixture consisting of several PDF-projection models [24].
2. Data window functions are needed prior to FFT processing for classification of real-world acoustic data. This adds additional complexity to the PDF projection, but can be solved [22].
3. There are applications which require accurate data generation according to a known PDF model (see Section 1.5). Thus, methods of generating data from $G(\mathbf{x}; H_0)$ need to be developed.
4. Additional work is needed to tie our method to existing maximum entropy methods such as the Burg spectral estimation method [25], which relates AR features to the maximum entropy rate process. We believe there is a close connection that should be explored. Our method appears more general since Burg's method is based on constraining the ACF, whereas ours is based on constraining an arbitrary feature.

## REFERENCES


[1] P. M. Baggenstoss, "The PDF projection theorem and the class-specific method," *IEEE Trans. Signal Process.*, vol. 51, no. 3, pp. 672–685, Mar. 2003.

[2] S. Kay, "Sufficiency, classification, and the class-specific feature theorem," *IEEE Trans. Inf. Theory*, vol. 46, pp. 1654–1658, Jul. 2000.

[3] J. N. Kapur, *Maximum Entropy Models in Science and Engineering*. New York, NY, USA: Wiley, 1993.

[4] A. Y. Khincin, *Mathematical Foundations of Information Theory*. Mineaola, NY, USA: Dover, 1957.

[5] T. Cover and J. Thomas, *Elements of Information Theory*. New York, NY, USA: Wiley, 1991.

[6] E. T. Jaynes, "On the rationale of maximum-entropy methods," *Proc. IEEE*, vol. 70, no. 9, pp. 939–952, 1982.

[7] J. Cai and F. Song, "Maximum entropy modeling with feature selection for text categorization," presented at the 4th Asia Inf. Retrieval Symp. (AIRS), Marbin, China, Jan. 2008.

[8] S. K. Saha, S. Sarkar, and P. Mitra, "Feature selection techniques for maximum entropy based biomedical named entity recognition," *J. Biomed. Inf. (Biomed. Natural Lang. Process.)*, vol. 42, pp. 905–911, Oct. 2009.

[9] A. McCallum, D. Freitag, and F. Pereira, "Maximum entropy Markov models for information extraction and segmentation," in *Proc. 17th Int. Conf. Mach. Learn.*, 2000, pp. 591–598.

[10] J. R. Quinlan, *Programs for Machine Learning*. San Mateo, CA, USA: Morgan Kaufman, 1993.

[11] S. Basu, C. A. Micchelli, and P. Olsen, "Maximum entropy and maximum likelihood criteria for feature selection from multivariate data," in *Proc. 2000 IEEE Int. Symp. Circuits Syst. (ISCAS)*, Geneva, May 2000, vol. 3, pp. 267–270.

[12] S. Kay, "Asymptotically optimal approximation of multidimensional pdf's by lower dimensional pdf's," *IEEE Trans. Signal Process.*, vol. 55, no. 2, p. 627, Feb. 2007.

[13] S. Eguchi, O. Komori, and A. Ohara, "Duality of maximum entropy and minimum divergence," *Entropy*, vol. 16, pp. 3552–3572, 2014.

[14] T. Luginbuhl, NUWC unpublished memorandum, 2012.

[15] D. R. Cox and D. V. Hinkley, *Theoretical Statistics*. London, U.K.: Chapman & Hall, 1974.

[16] S. M. Kay, private communication, Apr. 2013.

[17] P. M. Baggenstoss, "A modified Baum-Welch algorithm for hidden Markov models with multiple observation spaces," *IEEE Trans. Speech Audio*, pp. 411–416, May 2001.

[18] S. M. Kay, A. H. Nuttall, and P. M. Baggenstoss, "Multidimensional probability density function approximation for detection, classification and model order selection," *IEEE Trans. Signal Process.*, vol. 49, no. 10, pp. 2240–2252, Oct. 2001.

[19] R. Bartle, *The Elements of Real Analysis*. New York, NY, USA: Wiley, 1964.

[20] P. M. Baggenstoss, "The class-specific classifier: Avoiding the curse of dimensionality (tutorial)," *IEEE Aerosp. Electron. Syst. Mag., Special Tutorial Addendum*, vol. 19, no. 1, pp. 37–52, Jan. 2004.

[21] P. Mermelstein, "Distance measures for speech recognition, psychological and instrumental," *Pattern Recogn. Artif. Intell.*, pp. 374–388, 1976.

[22] P. M. Baggenstoss, "On the equivalence of Hanning-weighted and overlapped analysis windows using different window sizes," *IEEE Signal Process. Lett.*, vol. 19, pp. 27–30, Jan. 2012.

[23] S. Kay, *Modern Spectral Estimation: Theory and Applications*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1988.

[24] P. M. Baggenstoss, "Optimal detection and classification of diverse short-duration signals," in *Proc. Int. Conf. Cloud Eng.*, Boston, MA, USA, 2014, pp. 534–539.

[25] J. P. Burg, "The relationship between maximum entropy and maximum likelihood spectra," *Geophysics*, vol. 37, no. 2, pp. 375–376, 1971.


**Paul M. Baggenstoss** received his Ph.D. in electrical engineering (statistical signal processing) at the University of Rhode Island (URI) in 1990. From 1979 to 1996, he was with Raytheon Co., Portsmouth, RI. He joined the Naval Undersea Warfare Center (NUWC) Newport, RI, in 1996 where he has applied statistical signal processing and classification theory to problems in underwater acoustics. He is the author of several patents and numerous conference and journal papers in the field of signal processing and classification and has taught as an adjunct professsor of Electrical Engineering at the University of Connecticut, Storrs. During 2000, he was a visiting scientist at the University of Erlangen, Erlangen, Germany. During 2010, he was a exchange scientist at Fraunhofer FKIE, Bonn, Germany. He is the recipient 2002 URI Excellence Award in Science and Technology, the 2004 NAVSEA Scientist of the year award, and the 2004 NUWC Excellence in Science award.