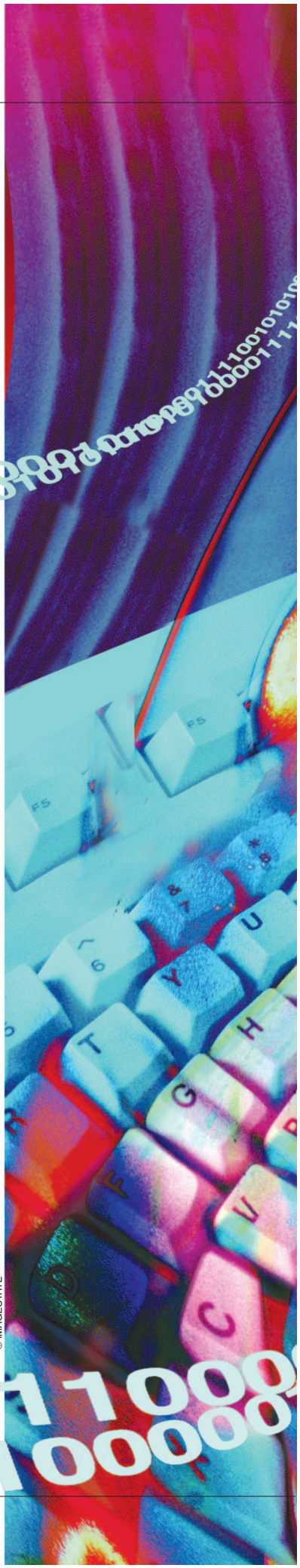
Deniz Erdogmus and Jose C. Principe

From Linear Adaptive Filtering to Nonlinear Information Processing

The design and analysis of information processing systems



Traditional approaches to statistical and adaptive signal processing have exploited the second order statistical properties of signals. This was motivated by the low complexity of the resulting algorithms and the existence of analytical solutions typically in the form of eigendecompositions. Recent advances in computing capabilities and the interest in new challenging signal processing problems that cannot be successfully solved using traditional techniques have sparked an interest in information-theoretic signal processing techniques. Adaptive nonlinear filters that process signals based on their information content have become a major focus of interest. The design and analysis of such nonlinear information processing systems is demonstrated in this article. Theoretical background on necessary information theoretic concepts are provided, nonparametric sample estimators for these quantities are derived and discussed, the use of these estimators for various statistical signal processing problems have been illustrated. These include data density modeling, system identification, blind source separation, dimensionality reduction, image registration, and data clustering.

INTRODUCTION

Wiener and Kolmogorov's framework of seeking *optimal projections* in spaces defined by stochastic processes initiated modern optimal filtering and changed forever our thinking about signal processing [37], [69]. The roots of adaptive model building go even further back to the 19th century, when mathematicians and scientists started describing real data by linear relationships and correlations between independent variables. The combination of the Gaussian assumption and second-order statistical criteria withstood the test of time and led to mathematically convenient and analytically tractable optimal solutions, which could be easily studied through conventional calculus, linear algebra, and functional analysis. The most familiar examples are the mean-square-error (MSE) in least-squares linear regression and output variance in principal components analysis (PCA).

1053-5888/06/$20.00©2006IEEE

The potential of optimal filtering became fully realized with the advent of digital computers, when the Wiener solution could be solved analytically for finite impulse response (FIR) filters using least-square type algorithms. Adaptive methodologies that search for the optimal solution very efficiently such as Widrow's least-mean-square (LMS) [50] could be implemented in digital signal processors to perform *optimally* (in the MSE sense) and *in real time* various challenging signal processing tasks. A curiosity at first, stochastic adaptive algorithms (i.e., processing the incoming data samples on a one-by-one basis) have become pervasive in signal processing and machine learning because they can be applied to problems where analytical solutions are unknown, as in the case of nonlinear filters. A noteworthy example is the backpropagation algorithm from neural networks [25].

In adaptive systems research (which is broadly used here to encompass traditional adaptive filtering as well as neural networks and various branches of machine learning), the user must specify a parametric mapper (a projector or a filter), which can be linear or nonlinear, an adaptation algorithm for the parameters (weights). and a criterion for optimality. The emphasis on second-order statistics as the choice of optimality criterion is still prevalent today. This is understandable because of three main reasons: 1) the success of linear systems combined with second-order statistics, due to the inevitable Gaussianization effect of convolution, 2) the well-established framework, and 3) the abundance of efficient adaptive algorithms. However, DSP engineers are addressing new problems of ever increasing difficulty and seeking more and more often solutions that involve nonlinear systems. Moreover, the criterion of optimality should be closely scrutinized; after all, it defines which statistical properties are being transferred from the measurements (input and/or desired responses) into the parameters of the model. With MSE, the optimal solutions obey just a second-order statistical constraint, and much broader and meaningful properties and frameworks are often required. For

instance, in blind separation of sources and blind deconvolution of linear channels, the insufficiency of second-order statistics in stationary environments have led to new approaches incorporating higher-order statistics into adaptation. Specifically, the field of independent component analysis (ICA) has benefited greatly from the use of information theoretic performance measures [33].

At the same time that Wiener was developing optimal signal processing, Shannon was laying the foundations of information theory to optimally design messages and systems to control stochastic fluctuations (noise) in the transmission of data [53]. Information theory deals with the quantification of statistical uncertainty in random processes and statistical dependencies between multiple random processes. The two main statistical descriptors proposed by information theory to design messages and systems are *entropy* and *divergence*: entropy, a measure of uncertainty of the random vector $\mathbf{X}$ with joint probability distribution function (pdf) $p(\mathbf{x})$, is a generalization of variance to processes with non-Gaussian distributions, and is defined by Shannon as [53], [10]

$$H_S(\mathbf{x}) = -\int p(\mathbf{x}) \log p(\mathbf{x}) d\mathbf{x} = -E[\log p(\mathbf{x})]. \tag{1}$$

As for divergence, a measure of statistical similarity, one can think of it as a generalization of algebraic distance measures (such as the Euclidean norm) to probability distribution spaces. In general, the Kullback-Leibler divergence (KLD) between two distributions $p(\mathbf{x})$ and $q(\mathbf{x})$ is defined as [10]

$$D_{\mathrm{KL}}(p; q) = \int p(\mathbf{x}) \log \frac{p(\mathbf{x})}{q(\mathbf{x})} d\mathbf{x}. \tag{2}$$

(The argument of a pdf is also used to denote the random variable which this pdf describes. Expectations are with respect to the random variable that is obvious from the context if not specified. Otherwise, a subscript shows the pdf with respect to which the expectations are computed.)

This is an asymmetric measure of distance (hence the term divergence) of the probability distribution p to q. This measure becomes zero if and only if p and q are identical distributions (except possibly at a zero-measure set of isolated points) and is positive otherwise.

In communications theory, it is more common to discuss mutual information, which is a measure of statistical dependency and a generalization of correlation to arbitrary nonlinear functional relationships between multiple processes with arbitrary probability distributions. Mutual information is a special case of KLD, when one measures the distance between the joint probability distribution and the product of the marginal distributions

$$I_S(\mathbf{x}; \mathbf{y}) = \int p(\mathbf{x}, \mathbf{y}) \log \frac{p(\mathbf{x}, \mathbf{y})}{p(\mathbf{x}) p(\mathbf{y})} d\mathbf{x} d\mathbf{y}. \tag{3}$$

Unfortunately, these descriptors were never fully incorporated in the mainstream of optimal signal processing. The goal of this article is exactly to outline the framework and the algorithms needed to move from quadratic to information costs, which lead to adaptive *information filtering*.

There are important differences between the application of information theory to communication systems and the reality of adaptive signal processing and machine learning. First, adaptive systems must handle continuous-valued random processes rather than discrete-valued processes. Noting this fact, we will focus our discussion in continuous random variables, described by their pdf. Second, adaptation algorithms require smooth cost functions; otherwise, the local search algorithms become difficult to apply. Third, the Gaussianity assumption so widespread in communications is normally a poor descriptor for the data statistics in machine learning and modern signal processing applications, especially when nonlinear topologies are considered. This means that the analytic approach taken in information theory must be modified with continuous and differentiable nonparametric estimators of entropy and divergence. To meet requirements two and three, we are convinced that Parzen window estimation [13] is a productive research direction. As we will see, Parzen estimation has the added advantage of linking information theory, adaptation, and kernel methods.

In this article, we will provide highlights of a methodology to implement adaptive information filtering, which we named *information theoretic learning* (ITL). ITL synergistically integrates the general framework of information theory in the design of new cost functions for adaptive systems, and it is posed to play an expanding role in adaptive signal processing. ITL does not only affect our understanding of optimal signal processing but also influences the way we approach machine learning, data compression, and adaptation as we will demonstrate in the sequel.

ADAPTIVE INFORMATION FILTERING

The conventional adaptive filtering framework (least squares view) describes the optimal filtering problem as one of obtaining the minimal error in the MSE sense between the desired response z and the system output y:

$$J(\mathbf{w}) = E\left[(z - y(\mathbf{w}))^2\right]. \tag{4}$$

Effectively this corresponds to estimating the orthogonal projection (in an Euclidean fashion) of the desired response z in the space spanned by the states of the system or, for FIR filters, the space spanned by the input signal.

Alternatively, the problem of estimating the system parameters $\mathbf{w}$ can be framed as model-based inference, since it relates measured data, uncertainty, and the functional description of the system and its parameters. The desired response z can be thought of as being created by an unknown transformation of the input vector $\mathbf{x}$. Therefore, the joint pdf $p(\mathbf{x}, z)$ fully characterizes this relationship including any noise that exists in the signal measurements. The job of the mapper is to construct an output y, which is a parametric function of the input, with parameter vector $\mathbf{w}$, that will approximate the unknown mapping. We can therefore think of the system output as an estimator $\tilde{p}_{\mathbf{w}}(\mathbf{x}, z)$ of $p(\mathbf{x}, z)$; here the subscript $\mathbf{w}$ denotes the dependency on the model parameters/weights. In probability

spaces, the role of optimization is therefore to minimize the KLD between these two distributions:

$$\min_{\mathbf{w}} J(\mathbf{w}) = \iint p(\mathbf{x}, z) \log \frac{p(\mathbf{x}, z)}{\tilde{p}_{\mathbf{w}}(\mathbf{x}, z)} d\mathbf{x} dz. \tag{5}$$

It can be shown that if we write $z = f(\mathbf{x}) + e$, where e is the error and $f(\mathbf{x})$ is the mapping function, minimizing the KLD is equivalent to minimizing the entropy of the error [16]

$$\min_{\mathbf{w}} H_S(e) = -\int p_{\mathbf{w}}(e) \log p_{\mathbf{w}}(e) de. \tag{6}$$

In this formulation, $f(\mathbf{x})$ is assumed to be a member of some family of models such as a linear filter, a neural network, or a parametric model derived from first physical principles. Its parameters are summarized by the weight vector $\mathbf{w}$, and the model is simply fit to the input-output data $(\mathbf{x}, z)$ obtained from the unknown system for system identification purposes. The dependence of the error on $\mathbf{w}$ comes from this underlying assumption. The output signal z is assumed to be generated by an unknown system $g(\mathbf{x})$, which may or may not belong to the parametric family that $f(.)$ comes from. Note that we have not imposed any constrains of linearity on the mapper nor any special conditions on the function f or on the pdfs of the input and desired responses. This is the power of the adaptive information filtering formulation. The issue of how to estimate all these quantities in practical cases remains. Before addressing estimation, we would like to expand on the generality of this result versus the conventional adaptive filtering formulation.

OPTIMIZATION IN THE ABSENCE OF DESIRED RESPONSE

Indeed, we can optimize the mapper even if no desired response is available, by constraining in some way the statistics of its output. An important case is ICA where one assumes a multi-input, multi-output (MIMO) mapper and the goal is to create statistically independent outputs [33]. For a nonlinear MIMO system $\mathbf{z} = \mathbf{f}(\mathbf{x}; \mathbf{w})$, the nonlinear ICA problem seeks to determine the parameters $\mathbf{w}$ of $\mathbf{f}(.)$ such that the mutual information between the components of $\mathbf{z}$ are minimized (preferably to zero)

$$\min_{\mathbf{w}} I(\mathbf{z}) = \int p(\mathbf{z}) \log \frac{p(\mathbf{z})}{\prod_d p(z_d)} d\mathbf{z}. \tag{7}$$

OPTIMIZING FOR EXTREMES OF SYSTEM OUTPUT

An alternative is to simply maximize (or minimize) the entropy of the output of the system (subject to some constraint on the weight vector norm or the nonlinear topology), which leads to an information theoretic factor analysis to discover interesting structures in the high dimensional input data

$$\min_{\mathbf{w}} | \max H(\mathbf{z}) = -\int p(\mathbf{z}) \log p(\mathbf{z}) d\mathbf{z}$$
$$\text{subject to} \quad E[h_i(\mathbf{z})] = \alpha_i \; i = 1, \ldots, m. \tag{8}$$

This formulation is useful in blind equalization, nonlinear principal component analysis, ICA, and novelty filtering [26], [27], [33].

MAXIMUM INFORMATION TRANSFER

An alternative optimization problem is to maximize the transfer of information between the input and the output of the system. This is called the principle of maximal information transfer and is related to the channel capacity theorem and the information bottleneck framework [56]. One could maximize the mutual information between the original input data and the transformed output data to preserve information maximally while reducing noise. For transformations with bounded outputs, this could be achieved by maximizing the joint output entropy as in the InfoMax principle [3]

$$\max_{\mathbf{w}} I(\mathbf{z}, \mathbf{x}) \equiv \max_{\mathbf{w}} H(\mathbf{z}) \text{ if range}(\mathbf{f}) \text{ is bounded.} \tag{9}$$

This formulation has also been suggested as a self-organization principle in distributed systems.

FEATURE EXTRACTION

Suppose that the desired response is a discrete-valued indicator label whose actual value is practically inconsequential (e.g., integer class labels as used in classification). In this case, one important question is how to project the high-dimensional input to a possibly nonlinear subspace, such that the discriminability with respect to the labels is preserved. The problem can be solved by maximizing the mutual information between the projection $\mathbf{z}$ and the class labels c

$$\max_{\mathbf{w}} I(\mathbf{z}, c) = H(\mathbf{z}) - \sum\nolimits_c p_c H(\mathbf{z}|c). \tag{10}$$

This principle generalizes the concepts behind PCA and linear discriminant analysis (LDA) for finding effective reduced-dimensionality nonlinear feature projections.

CLUSTERING

Finally, assume that the goal of the mapper is to divide the input data into a preselected number of structurally and/or statistically distinct groups (clusters). Here, the weights become the assigned cluster membership values and the criterion is to assign samples to a cluster such that the clusters are defined as compactly and distinctly as possible, measured by cluster entropy and divergence. In the case of two clusters, one could use a symmetric KLD measure, for example

$$\max_{\mathbf{w}} D_{\mathrm{KL}}(p_1(\mathbf{x}); p_2(\mathbf{x})) + D_{\mathrm{KL}}(p_2(\mathbf{x}); p_1(\mathbf{x})). \tag{11}$$

We therefore conclude that entropy and divergence help bridge the traditional gap between supervised and unsupervised learning as well as provide a convenient framework that allows the treatment of discrete, continuous, and mixed-type random variables with the same type of tools. Adaptive information filtering creates effectively a unifying principle for system adaptation.

Finally, the adaptive information filtering framework benefits from a differential geometry treatment to better understand the issues and solutions to the extension of adaptive filtering

theory to nonlinear systems and information theoretic cost functions. In fact, probability distributions are differential forms, and both the measurements and the parametric system define manifolds in probability space. The Fisher information is the natural metric in manifolds of probability distributions. In every manifold, one can define the tangent space, and the corresponding cotangent space, that defines the projection back to the real line so we have sufficient topological structure to solve optimization problems. Therefore, the problem of nonlinear adaptive filtering (parameter estimation) is one of finding the closest distance (as measured by the information measure used) in the system manifold to the measurement manifold, a principle that is a direct generalization of the orthogonality principle of linear adaptive filtering.

INFORMATION THEORETIC LEARNING

ITL is nothing but a set of algorithms to implement adaptive information filtering. As typical in many optimal signal processing and machine learning problems, the probability density function of the data is unknown, and the Gaussianity assumption is a stretch. The fundamental issue in ITL, therefore, is how to estimate entropy and divergence directly from samples. As stated earlier, the estimators have to be smooth and defined for continuous random variables; smooth because they will be used as cost functions to be searched by local algorithms, and continuous because the problems of interest in optimal signal processing and machine learning often are formulated for continuous random variables (e.g., function approximation). Instead of utilizing Shannon's definition of information, we pursue Renyi's entropy definition for the reasons that will be apparent below.

RENYI'S DEFINITIONS

Following the postulate-based derivation of information theory, Renyi relaxed the additivity property of information (by considering exponential additivity rather than linear additivity). These new set of postulates lead to the following generalized definitions of entropy and mutual information providing the flexibility of a parametric family, while maintaining Shannon's definitions as the special case $\alpha = 1$. (The definitions of Renyi are discontinuous at $\alpha = 1$, but using L'Hopital's Rule one can show that their limit as $\alpha \to 1$ is equal to Shannon's definitions for the corresponding quantity.) The order-α entropy of $\mathbf{X}$ is defined as [48]

$$H_\alpha(\mathbf{x}) = \frac{1}{1-\alpha}\log\int p^\alpha(\mathbf{x})d\mathbf{x} = \frac{1}{1-\alpha}\log E[p^{\alpha-1}(\mathbf{x})]. \quad (12)$$

Conditional entropy $H_\alpha(\mathbf{y}|\mathbf{x})$ is defined similarly using the conditional distribution $p(\mathbf{y}|\mathbf{x})$ and averaged over $\mathbf{x}$. Notice that Renyi's entropy is related to the L_α-norm of the data distribution $p(.)$.

Renyi's definition of mutual information is based on the α-divergence between two distributions $p(\mathbf{x})$ and $q(\mathbf{x})$, which converges to KLD as $\alpha \to 1$ [48]

$$D_\alpha(p;q) = \frac{1}{\alpha-1}\log\int p(\mathbf{x})\left(\frac{p(\mathbf{x})}{q(\mathbf{x})}\right)^{\alpha-1}d\mathbf{x}. \quad (13)$$

Consequently, the order-α mutual information is

$$I_\alpha(\mathbf{x};\mathbf{y}) = \frac{1}{\alpha-1}\log\iint p(\mathbf{x},\mathbf{y})\left(\frac{p(\mathbf{x},\mathbf{y})}{p(\mathbf{x})q(\mathbf{y})}\right)^{\alpha-1}d\mathbf{x}d\mathbf{y}. \quad (14)$$

Generalization of mutual information to more than two arguments is obtained in analogy with the Shannon case.

ALTERNATIVE DISTANCE MEASURES

Alternative measures of divergence between two pdfs are available in the literature. A well-known example is the Csiszar divergence. For an arbitrary convex function $h(.)$ such that $h(1) = 0$, we define [11]

$$D_h(p;q) = \int p(\mathbf{x})h\left(\frac{q(\mathbf{x})}{p(\mathbf{x})}\right)d\mathbf{x}. \quad (15)$$

For the specific choice of $h(.) = -\log(.)$, (15) reduces to KLD. The flexibility in selecting $h(.)$ allows the use of certain a priori information about the problem to be incorporated into the solution (such as suppressing the effects of tails to improve robustness to outliers).

On the other hand, these classical information divergence measures are asymmetric. Thus they are not distances, but it is straightforward to define a distance from divergence by adding two opposite-direction divergences to obtain symmetry. Probability spaces have a Riemannian structure, so when the pdfs are close to each other, one can naturally adopt formal distance measures from Euclidean space such as the L_β-norm

$$D_\beta(p;q) = \left(\int |p(\mathbf{x}) - q(\mathbf{x})|^\beta\, d\mathbf{x}\right)^{1/\beta}. \quad (16)$$

Of these measures, Euclidean distance ($\beta = 2$) is especially interesting due to its quadratic properties. We can also define an inner product distance using the Cauchy-Schwartz inequality, to obtain

$$D_{CS}(p;q) = -\frac{1}{2}\log\frac{\left(\int p(\mathbf{x})q(\mathbf{x})d\mathbf{x}\right)^2}{\int p^2(\mathbf{x})d\mathbf{x}\int q^2(\mathbf{x})d\mathbf{x}}. \quad (17)$$

Notice that the argument of the log(.) is always between [0,1], since it is the cosine of the angle between the two distributions, it can be considered as a Riemannian distance in the Hilbert space. From these alternative distance measures, we can obtain alternative mutual information measures as desired.

NONPARAMETRIC SAMPLE ESTIMATES

ITL requires the evaluation and optimization of performance indices based on information theoretic concepts, such as entropy and mutual information. Since in typical ITL applications the data distributions are not known, analytical evaluation of these performance indices is not possible. These cost functions must be evaluated using sample estimators. Sample estimators for information theoretic quantities typically rely on the *plug-in density estimation* principle. That is, using the

available samples, one needs to obtain an estimate of the underlying probability distributions, which is in turn substituted into the cost function.

For illustration purposes and aiming for simplicity, we focus on estimating the entropy of a random variable X from its scalar samples $\{x_1, \ldots, x_N\}$. There are three possible techniques one can assume towards estimating the pdf of a random variable from its independent and identically distributed (iid) samples: parametric, semiparametric, and nonparametric. There are many parametric and semiparametric approaches for entropy estimation in the literature and a good review of entropy estimation methods can be found in [1]. Here we focus on the nonparametric estimators.

The most straightforward nonparametric approach in entropy estimation is to consider a histogram approximation for the underlying distribution. Fixed-bin histograms lack the flexibility of sliding histograms, where the windows are placed on every sample. A generalization of sliding histograms is obtained by relaxing the rectangular window to assume smoother functional forms in the form of continuous and differentiable (and preferably symmetric and unimodal) pdfs. This generalization is referred to as kernel density estimation (KDE). Another generalization of histograms is obtained by letting the bin-size vary in accordance with local data distribution. In the case of rectangular windows, this corresponds to nearest neighbor density estimation [13], and for KDE this means variable kernel size [12], [13]. The corresponding entropy estimates are presented below.

ENTROPY ESTIMATION BASED ON SAMPLE SPACING

Suppose that the ordered samples $\{x_1 < x_2 < \cdots < x_N\}$ drawn from $q(x)$ are provided. We assume that the distribution support is $[x_0, x_{N+1}]$ and that the distribution is piecewise constant [67], leading to the following approximation:

$$p(x) = \begin{cases} 1/((x_1 - x_0)(N+1)) & x_0 \le x < x_1 \\ 1/((x_2 - x_1)(N+1)) & x_1 \le x < x_2 \\ \vdots & \vdots \\ 1/((x_{N+1} - x_N)(N+1)) & x_N \le x < x_{N+1}. \end{cases} \quad (18)$$

Denoting the empirical cdf by $P(x)$, for ordered statistics, it is known that

$$E[P(x_{i+m}) - P(x_i)] = \frac{m}{N+1}, \quad i = 1, \ldots, N-m \quad (19)$$

where the expectation is evaluated with respect to the joint data distribution $q(x_1) \ldots q(x_N)$, assuming iid samples. Substituting (18) in Renyi's entropy and using the identity in (19), we obtain the m-spacing estimator for Renyi's entropy as

$$H_\alpha(x) \approx \frac{1}{1-\alpha} \log \left[\frac{1}{N-m} \sum_{i=1}^{N-m} \left(\frac{(N+1)}{m} (x_{i+m} - x_i) \right)^{1-\alpha} \right]. \quad (20)$$

The spacing interval m is chosen to be a slower-than-linear increasing function of N to guarantee asymptotic consistency and efficiency. Typically, $m = N^{1/2}$ is preferred in practice due to its simplicity.

Using L'Hopital's rule, we obtain the sample spacing estimator for Shannon's entropy, as expected [67]. A difficulty with the sample spacing approach is its generalization to higher dimensionalities. Computational issues as well as nonsmoothness of the resulting estimator hamper their usefulness in learning and adaptation.

Perhaps the most popular extension of sample-spacing estimators to multidimensional random vectors is the one based on the minimum spanning tree recently popularized by Hero [29]. This estimator relies on the fact that the integral in Renyi's definition of entropy is related to the sum of the lengths of the edges in the minimum spanning tree (this is the tree—a graph that connects all data points without any loops—that has minimum total length), with useful asymptotic convergence guarantees. One drawback of this approach is that it only applies to entropy orders of $0 < \alpha < 1$. Another drawback is that finding the minimum spanning tree itself is a computationally cumbersome task that is also prone to local minima due to the heuristic selection of a neighborhood search radius by the user.

Another generalization of sample spacing estimates to multidimensional entropy estimation has relied on the L_1-norm as the distance measure between the samples instead of the usual Euclidean norm [39]. This technique, in principle, can be generalized to arbitrary norm definitions. The drawback of this method is its nondifferentiability, which renders it next to useless for traditional iterative gradient-based adaptation. This approach essentially corresponds to extending (18) to the multidimensional case as data-dependent, variable-volume hyperrectangles. One could easily make this latter approach differentiable through the use of smooth kernels rather than rectangular volumes. Such modification will also form the connection between the sample-spacing methods and kernel based methods described next.

PARZEN WINDOWING BASED ENTROPY ESTIMATION

Kernel density estimation, also referred to as Parzen windowing, is a well-understood and useful nonparametric technique that can be employed for entropy estimation in the plug-in estimation framework [1]. For a given set of iid samples $\{x_1, \ldots, x_N\}$ drawn from $q(x)$, the Parzen window estimate for the distribution, assuming a fixed-size kernel function $K_\sigma(\xi)$ for simplicity, is given by

$$p(x) = \frac{1}{N} \sum_{i=1}^{N} K_\sigma(x - x_i). \quad (21)$$

The kernel function and its size can be optimized in accordance with the maximum likelihood (ML) principle [14], [51], or other rules-of-thumb could be employed to obtain approximate optimal parameter selections [12], [55].

For a given kernel function, the Parzen window estimator exhibits the following properties:

1) For fixed σ, $E[p(x)] = \lim_{N\to\infty} p(x) = q(x)^* K_\sigma(x)$.

2) For fixed σ, $\lim_{N\to\infty} Var[p(x)] = 0$.

3) If $\lim_{N\to\infty}\sigma(N) = 0$ and $\lim_{N\to\infty} N\sigma(N) = \infty$, then $\lim_{N\to\infty} p(x) = q(x)$ in probability.

These conditions guarantee that for analytic probability distribution functions, the Parzen window estimate is asymptotically unbiased and consistent (using a suitable annealing rate for the kernel size).

We will first treat the nonparametric estimation of Renyi's quadratic entropy ($\alpha = 2$) with Parzen windows to stress a surprising simplifying result. Substituting (21) in Renyi's entropy definition (12), we obtain the following nonparametric kernel entropy estimator, letting $H_2(X) = -\log V_2(x)$ [45]:

$$V_2(X) = \int p^2(x)dx = \int \left(\frac{1}{N}\sum_{i=1}^{N} K_\sigma(x - x_i)\right)^2 dx = \frac{1}{N^2}\sum_{j=1}^{N}\sum_{i=1}^{N} K_{\sigma\sqrt{2}}(x_j - x_i) \quad (22)$$

where $V_2(X)$ is called the *quadratic information potential* and K is Gaussian. Note that the information potential for the continuous random variable can be exactly estimated by the double sum over the samples due to the well now property of the integral of product of Gaussians is still a Gaussian, but with larger variance. The only approximation in the estimator of (22) stems from the finite number of samples and the kernel size. If we had followed the general plug-in strategy, the information potential would be written as the expected value of $p(X)$, leading to an estimate identical to (22), but with kernel size σ instead of $\sigma\sqrt{2}$. This means that the sample mean approximation error in the plug-in estimator is exactly compensated by utilizing a wider kernel.

The kernel estimator introduces another interesting interpretation of the Renyi entropy of a random variable. Specifically, suppose that a Mercer kernel (a finite-energy function that is symmetric in its two arguments similar to a symmetric matrix) with the following eigendecomposition is utilized [25]:

$$K(x - x') = \sum_{k=1}^{\infty} \lambda_k \varphi_k(x)\varphi_k(x') = \varphi^T(x)\Lambda\varphi(x'). \quad (23)$$

In the literature, almost all work deals with positive semidefinite kernels (i.e., Mercer kernels with nonnegative real eigenvalues λ_k). The eigendecomposition in (23) illustrates that the kernel calculation on the original data is equivalent to an inner product calculation in the feature space defined by the eigenfunction vector φ. This is in fact, the underlying principle of the well-known support vector machine formalism. Substituting (23) in (21), we observe that

$$p(x) = \frac{1}{N}\sum_{i=1}^{N} \varphi^T(x)\Lambda\varphi(x_i) = \varphi^T(x)\Lambda\mu_\varphi(x). \quad (24)$$

That is, Parzen density estimation is equivalent to nonlinearly transforming the data to a *feature space* through the eigenfunctions of the kernel and taking the weighted inner product of the evaluation point with the mean of the training samples. It is also interesting to note that all of the transformed samples lie on a hyperellipsoid (with axes determined by the eigenvalues) in the feature space, since the norm of the transformed values is constant: $\varphi^T(x)\Lambda\varphi(x) = K(0)$, regardless of x.

Now consider Renyi's quadratic entropy as estimated by (22) and the spectral decomposition of the kernel function. (To avoid too many parentheses, the arguments of log functions are not enclosed. Products following a log operation have precedence.)

$$\begin{aligned} H_2(x) &= -\log \frac{1}{N^2}\sum_{j=1}^{N}\sum_{i=1}^{N} K_\sigma(x_j - x_i) \\ &= -\log \frac{1}{N^2}\sum_{j=1}^{N}\sum_{i=1}^{N} \varphi^T(x_j)\Lambda\varphi(x_i) \\ &= -\log \mu^T_{\varphi(x)}\Lambda\mu_{\varphi(x).} \end{aligned} \quad (25)$$

Therefore, Renyi's quadratic entropy can be estimated by a moment of the distribution (just like the mean and variance estimators), but in a transformed (kernel) space. With the spectral decomposition substitution for the kernel, Renyi's quadratic entropy estimator becomes simply the log-norm-squared of the mean vector of the feature space samples $\varphi(x_i)$.

Returning to the general definition of Renyi's entropy and approximating the expected value by the sample mean, we obtain [17]

$$H_\alpha(x) = \frac{1}{1-\alpha}\log\frac{1}{N^\alpha}\sum_{j=1}^{N}\left(\sum_{i=1}^{N} K_\sigma(x_j - x_i)\right)^{\alpha-1}. \quad (26)$$

As expected, using L'Hopital's rule, the kernel estimator for Shannon's entropy could be obtained from (26) as well as by employing the plug-in strategy directly on Shannon's entropy definition

$$H_S(x) = -\frac{1}{N}\sum_{j=1}^{N}\log\frac{1}{N}\sum_{i=1}^{N} K_\sigma(x_j - x_i). \quad (27)$$

The kernel estimation technique illustrated above can be utilized for nonparametrically estimating other information theoretic quantities such as density divergence and conditional entropy. In particular, it can be used to estimate the algebraic norm between pdfs (16) and the Cauchy Schwartz distance (CSD) (17), as well as Csiszar, KL, and α divergences. Extension of the ideas to multidimensional distributions is trivial and only requires utilizing a multidimensional kernel function in Parzen windowing.

The kernel size (also referred to as bandwidth) is a parameter that is introduced by the nonparametric estimation technique, and there exist effective methods of selecting an appropriate value: 1) leave-one-out type maximum likelihood solution [14], [51] and 2) smoothness-constraint based least-squares fit solution [55].

INFORMATION THEORETIC ALGORITHMS

In supervised training of a network with layers of weights, such as multilayer perceptrons (MLPs), the output error is backpropagated through the layers to determine the gradient update rule for each weight under the minimum mean squared error (MSE) criterion. The conventional LMS algorithm is a special case of backpropagation for a single layer network of linear processing elements. Information theoretic learning, on the other hand, utilizes entropic optimality criteria, which must be non- or semi-parametrically estimated from the available samples, using one of the methods mentioned earlier.

ROLES OF INFORMATION FORCES IN LEARNING

Kernel approaches lend a useful characteristic to the associated ITL algorithms that facilitate the generalization of the backpropagation principle to information forces through an interesting analogy with the interactions of physical particles. In particular, the selected kernel function applied to the sample behaves similar to the potential fields generated by particles in Newtonian physics with an interaction law given by the shape of the kernel, while their spatial gradients represent the information forces. This property is best illustrated via the kernel estimator for Renyi's quadratic entropy given in (22). Consider the quadratic information potential [45], [17]

$$\begin{aligned} V_2(x) &= \frac{1}{N^2}\sum_{J=1}^{N}\sum_{i=1}^{N}K_\sigma(x_j - x_i) \\ &= \frac{1}{N}\sum_{j=i}^{N}\left(\frac{1}{N}\sum_{i=1}^{N}K_\sigma(x_j - x_i)\right) \\ &= \frac{1}{N}\sum_{j=1}^{N} V_2(x_j). \end{aligned} \tag{28}$$

Note that the information potential of the samples in the training, now called *information particles* $\{x_1, \ldots, x_N\}$ is the average of the information potentials experienced by individual particles, denoted by $V_2(x_j)$. Similarly, the information potential experienced by particle x_j is the average of contributions from the other particles in the sample set. In particular, the contribution of xi to the potential of x_j is determined by the kernel function utilized in Parzen windowing as $V_2(x_j \mid x_i) = K_\sigma(x_j - x_i)$. Consequently, each particle exerts an information force to each other evaluated by the gradient of the information potential $F_2(x_j \mid x_i) = \nabla K_\sigma(x_j - x_i)$, where ∇ denotes the gradient operation with respect to the kernel argument.

In analogy to adaptive signal processing algorithms where system parameters are adapted by the gradient of the cost, in ITL system parameters are adapted by information forces created among the pairwise interactions in the sample set. Indeed, in supervised learning if the error samples e_j were generated by an MLP with weights w, the gradient of the information potential with respect to these weights would be [17]

$$\frac{\partial V_2(e)}{\partial \mathbf{w}} = \frac{1}{N^2}\sum_j\sum_i F_2(e_j \mid e_i)\left(\frac{\partial x_i}{\partial \mathbf{w}} - \frac{\partial x_j}{\partial \mathbf{w}}\right). \tag{29}$$

For the adaptation of an MLP, the principle of error backpropagation can be extended to information force backpropagation by simply substituting the injected error of the MSE cost with the information force $F(e_i \mid e_j)$ of ITL. More generally, for Renyi's order-α entropy, it can be shown that the information force becomes $F_\alpha(x_j) = (\alpha - 1)p^{\alpha-2}(x_j)F_2(x_j)$, where $p(x_j)$ is the Parzen density estimate of the sample x_j. Thus, the order of the entropy will emphasize the contributions of samples in dense regions of the data (by increasing α) or sparse regions of the data (by decreasing α). This is consistent with the L_α-norm interpretation of Renyi's entropy. In the recent literature on learning algorithms, the L_1-error-norm and ε-insensitive error functions have been popular due to the weaknesses of the traditional MSE criterion, such as susceptibility to outliers. The utilization of information theoretic criteria estimated using Parzen windowing based nonparametric density estimation naturally results in criteria that have a computational complexity of $O(N^2)$, where N is the number of samples in the training set used for optimizing the models/filters, which compares unfavorable with the $O(N)$ complexity of MSE based batch learning.

STOCHASTIC INFORMATION GRADIENT

Stochastic gradients of cost functions are commonly used in obtained online adaptation algorithms that optimize the weights using samples one-by-one (typically as they arrive as a time-series). The main idea behind stochastic gradients is that the expectation operator in the statistical cost function can be dropped to obtain a gradient update rule that on average follows the batch gradient update direction.

For illustration, we focus our discussion on the stochastic gradient of the order-α information potential. Approximating the expectation by the most recent sample x_k and utilizing a small set of previously available samples for Parzen windowing, the instantaneous cost is [19]

$$\begin{aligned} V_\alpha(x) &= E\left[p^{\alpha-1}(x)\right] \approx p^{\alpha-1}(x_k) \\ &= \left(\frac{1}{L}\sum_{i=1}^{L}K_\sigma(x_k - x_{k-i})\right)^{\alpha-1}. \end{aligned} \tag{30}$$

This results in the following stochastic gradient that is called the stochastic information gradient (SIG):

$$\mathrm{SIG} = (\alpha - 1)p^{\alpha-2}(x_k)\left(\frac{1}{L}\sum_{i=1}^{L}\nabla K_\sigma(x_k - x_{k-i})\right). \tag{31}$$

Other approximations of the original information potential estimator are possible [31]. The stochastic gradient reduces the computational complexity of each gradient update significantly from $O(N^2)$ to $O(L)$. By construction, the expected value of SIG is equal to the batch gradient.

In contrast to the conventional stochastic algorithms like LMS, which depend on the instantaneous sample only, SIG relies on temporal differences of samples due to the pairwise

nature of the kernel estimator for entropy. This results in an interesting observation. For example in supervised linear filter adaptation using minimum quadratic entropy, assuming a Gaussian kernel, the SIG reduces to $\Delta e_k \Delta \mathbf{u}_k$ where $\mathbf{u}$ is the input vector and e is the output error for the adaptive filter. While the LMS update ($e_k \mathbf{u}_k$) tries to decorrelate the error and the input signals, the SIG tries to decorrelate the temporal differences of these signals [19]. This observation leads to linear adaptive filter learning algorithms based on temporal difference statistics that are able to reduce (and even completely eliminate) the bias introduced to the filter solution due to the power of noise in these signals. The resulting algorithms, studied under the error whitening and instrumental variables principles, are not the subject of this review [46], [47].

IMPROVED FAST GAUSS TRANSFORM

Information theoretic learning algorithms based on kernel density estimation require the calculation of sums of Gaussian function evaluations between pairs of data vectors. Therefore, the $O(N^2)$ complexity of the information potential calculation is circumvented (in low-dimensional data analysis situations) by the fast Gauss transform (FGT) [24], [61], which decreases the computational complexity of the batch learning algorithm to $O(Np^n)$, where n is the data dimensionality and p is the order of truncation for the Gram-Charlier polynomial series expansion used in this approximation. The original FGT suffers from the curse of dimensionality [15] because it was derived using the fast multipole methodology [23], which was designed for particle interactions in astrophysics (3-D interactions). Exploiting the differentiability of the Gaussian function leads to a more efficient approximation [71], which we briefly describe below. For simplicity, consider the spherical kernel case with the following exponential (where we introduce an arbitrary expansion center $\mathrm{x}*$ and the variables $\Delta\mathrm{y} = \mathrm{y} - \mathrm{x}^*$ and $\Delta\mathrm{x} = \mathrm{x} - \mathrm{x}^*$):

$$e^{-\|\mathrm{y}-\mathrm{x}\|^2/h^2} = e^{-\|\mathrm{y}\|^2/h^2} e^{-\|\mathrm{x}\|^2/h^2} e^{2\Delta\mathrm{y}\Delta\mathrm{x}/h^2}. \quad (32)$$

The first two exponentials can be evaluated at the data points individually leading to a complexity of $2N$. The last exponential term can be expanded into multivariate Taylor series as

$$e^{2\mathrm{xy}} = \sum_{\mathrm{a} \geq 0} \frac{2^{|\mathrm{a}|}}{\mathrm{a}!} \mathrm{x}^{\mathrm{a}} \mathrm{y}^{\mathrm{a}} \quad (33)$$

where the following multivariate notations are employed: $|\boldsymbol{\alpha}| = \alpha_1 + \cdots + \alpha_n$, $\boldsymbol{\alpha}! = \alpha_1! \ldots \alpha_n!$, and $\mathrm{x}^{\mathrm{a}} = x_1^{\alpha_1} \ldots x_n^{\alpha_n}$. If the Taylor expansion in (33) is restricted to order $p-1$, then the number of terms is *Combination* $(p+n-1, n)$, which is substantially smaller than the original FGT with complexity p^n.

ADAPTIVE MIXTURE MODELING BY MINIMIZING DENSITY DIVERGENCE

Information-theoretic techniques have found widespread application in adaptive optimization of mixture density models as well as self-organizing maps (SOMs) and other topographic maps. Most work in this field relies on Shannon's definitions of entropy and mutual information as well as the closely related KLD. This is not without reason; as shown in the appendix, the traditional parametric ML and MAP density estimates asymptotically converge to solutions that minimize the KLD with respect to the true underlying data distribution within the parametric family of densities selected. In the case of a convex parametric family, this corresponds to determining the orthogonal projection of the underlying density onto the convex set in accordance with the Pythagorean theorem [10] of relative entropy and the information geometry of parametric statistical models.

MIXTURE MODEL-ORDER SELECTION

An important property of information theoretic divergence measures, such as the KLD, is that they are invariant under monotonic transformations of the corresponding random variables (including scale changes). This means that if the signals of interest are passed through invertible nonlinear operations (e.g., this is the case in homomorphic signal processing), intuitively and mathematically information is unchanged. The information theoretic divergence measures adhere to this fact and maintain the value regardless of the nonlinearities employed, perhaps for preprocessing. In contrast, traditional distance measures do not possess this invariance property. Thus, information-theoretic measures are more suitable for assessing information content compared to traditional distance measures, such as the L_2-norm of the error between two distributions in the function space. Li and Barron point out this fact to motivate their analysis on determining bounds for the approximation errors of mixture density models utilizing the KLD as their error metric [40].

THEOREM

Consider a family of mixture distributions with possibly infinitely many components

$$p_M(x) = \sum_{k=1}^{M} \alpha_k K(x; \theta). \quad (34)$$

For an arbitrary distribution $f(x)$, there exists an *optimal* M-component mixture density approximation of the form (34) that has an error bounded by

$$D_{\mathrm{KL}}(f; p_{M*}) \leq D_{\mathrm{KL}}(f; p_{\infty *}) + c_f^2 \gamma / M. \quad (35)$$

where $\gamma = 4(\log(3e^{1/2}) + \sup_{\theta_1, \theta_2, x}\{\log K(x; \theta_1)/K(x; \theta_2)\}$ and $c_f^2 = \int f(x) p_{\infty *}^{-2} \sum_{k=1}^{\infty} \alpha_k K^2(x; \theta) dx$. □

This is a significant result showing that, as the number of mixtures increase, the optimal estimate has a KLD that decreases linearly with the number of components M regardless of the choice of the kernel K, as long as this choice leads to finite γ (e.g., Gaussian kernels). Furthermore, one can show that a penalized entropy criterion (equivalent to maximum likelihood) to select the order M that minimizes

$$-\frac{1}{N}\sum_{i-1}^{N}\log p_M(x_i)+\frac{2(Md\,\log(NABe)+2\log(M+1))}{N} \quad (36)$$

where d is the number of parameters in θ, A is the side length of a cube in the θ-space that bounds the allowed values, and B is the smallest value such that

$$\sup_x |\log K(x;\theta)-\log K(x;\theta')|\le B\|\theta-\theta'\|_1 \quad (37)$$

will satisfy the following upper bound on the expected KLD-based approximation error [40]:

$$E[D_{\mathrm{KL}}(f;p_{M*})]-D_{\mathrm{KL}}(f;p_{\infty*})\le\frac{\gamma^2c_f^2}{M}+\frac{2\gamma Md\,\log(NABe)}{N}+\frac{4\log(M+1)}{N}. \quad (38)$$

The significance of this result can be summarized as follows. The average KLD between the optimal model fit with M components and the true density cannot be much greater (determined by the upper bound) than the KLD between the optimal model fit with infinite components and the true density. These results are similar to Akaike's information criterion (AIC) and the minimum description length (MDL) principle and provide a statistically meaningful penalty function for model fitting to select the *optimal* model order. In fact, there are recent results by Stoica that connect model order selection to the minimization of KLD [60].

Another information-theoretic model order penalization technique for mixture models is the normalized entropy criterion (NEC) [7]. For the mixture density model given in (34), the log-likelihood function, which is a finite sample approximation of the KLD optimality criterion as discussed in the appendix, can be broken into two parts as follows:

$$L_*(M)=\sum_{i=1}^{N}\log\sum_{k=1}^{M}\alpha_{k*}K(x_i;\theta_{k*})=C_*(M)+H_*(M) \quad (39)$$

where the subscript "$_*$" denotes these are the optimal values according to ML and

$$\begin{aligned}t_{ik}&=\frac{\alpha_{k*}K(x_i;\theta_{k*})}{\sum_{k=1}^{M}\alpha_{k*}K(x_i;\theta_{k*})}\\ C_*(M)&=\sum_{k=1}^{M}\sum_{i=1}^{N}t_{ik}\log(\alpha_{k*}K(x_i;\theta_{k*}))\\ H_*(M)&=-\sum_{k=1}^{M}\sum_{i=1}^{N}t_{ik}\log t_{ik}.\end{aligned} \quad (40)$$

The NEC is then defined as shown in (41) for $M>1$. Then, the number of components in the mixture can be selected as the minimizer of NEC (M) if NEC $(M_*)\le 1$ and $M=1$ otherwise [4]

$$\mathrm{NEC}(M)=\frac{H_*(M)}{C_*(M)-C_*(1)},\quad M>1. \quad (41)$$

In contrast to the penalized entropy method, the NEC method tries to make sure that a model fit achieves both large data likelihood and a balanced usage of the components (as measured by the normalized entropy) in the mixture model.

LEARNING TOPOGRAPHIC MAPS

From an ITL perspective, nonparametric modeling of data distributions while preserving the topography (i.e., neighborhood relations) of the sample set has drawn more attention. Specifically, it is possible to formulate the learning procedure of a self-organizing topographic map from an entropy maximization perspective. Consider a mixture model

$$y=\sum_{i=1}^{M}K\left(\mathrm{x};\mathrm{w}_i,\sigma_i^2\right) \quad (42)$$

in a d-dimensional data space, where the value of the mixture model is considered to be an equally weighted combination of M processing element (PE) outputs with activation functions defined by the kernel function $K(.)$. Each kernel function is represented by its center location w_i in the data space and for simplicity, we assume circularly symmetric kernels with width parameter σ_i (from a neural network perspective, this corresponds to the imposition that the PEs' receptive fields are circularly symmetric).

The parameters of the kernels (i.e., the component locations and widths) can be optimized by maximizing the joint entropy of the PE outputs, which according to (8), is given by [66]

$$H_S(\mathrm{y})=\sum_{i=1}^{M}H_S(y_i)-I_S(\mathrm{y}). \quad (43)$$

Assuming that the kernel monotonically decays as a function of the Euclidean distance to the center (such as Gaussian), the output entropy of an individual PE can be estimated utilizing the probability distribution of the Euclidean distance of a sample to the center of the receptive field of the neuron [66]. Denoting the pdf of the radius by $p_r(r)$, we have

$$p_{y_i}(y_i)=p_r(r)/|\partial y_i/\partial r|, \quad (44)$$

which leads to

$$H_S(y_i)=H_S(r)+\int_0^{\infty}p_r(r)|\partial y_i/\partial r|dr. \quad (45)$$

The minimization of (43) is then achieved by utilizing $|\partial y_i/\partial r|$ as a stochastic approximation to the marginal entropy part of the criterion. Introducing neighborhood functions (e.g., radial basis functions with certain bandwidth as in SOMs) in the weight updates to preserve topology results in the following rule for the centers and the kernel width:

$$\begin{aligned}\Delta\mathrm{w}_i&=\eta_w\Lambda(i,i^*,\sigma_\Lambda)(\mathrm{x}-\mathrm{w}_i)/\sigma_i^2\\ \Delta\sigma_i&=\eta_\sigma\Lambda(i,i^*,\sigma_\Lambda)\left(||\mathrm{x}-\mathrm{w}_i||^2-d\sigma_i^2\right)/\left(d\sigma_i^3\right)\end{aligned} \quad (46)$$

where Λ is a neighborhood function and i^* is the index of the center nearest to the current sample $\mathbf{x}$. The neighborhood functions help preserve the topology by making sure that nearby components respond to nearby data points, thus components neighboring each other also represent samples that are near each other in the data space. This observation shows that the SOM [36] could be interpreted as approximating the ITL rule in (43) assuming a fixed kernel size. The kernel size allows the objective assessment of the likelihood of a particular sample being represented by a specific component (PE). Provided that the kernel function has infinite support, every sample will be represented by contributions from every center at various levels (as in Gaussian mixtures). However, the degree of confidence that one has in the quantization of a sample vector to a center will be measured by the ratio of the contribution from the nearest PE to the contributions of the other PEs in the network.

WE PROVIDE HIGHLIGHTS OF A METHODOLOGY TO IMPLEMENT ADAPTIVE INFORMATION FILTERING, WHICH WE NAMED INFORMATION THEORETIC LEARNING.

Another possible criterion that naturally arises from information theory is density divergence [40]. Although any valid density divergence measure that we have defined earlier could be employed for the purpose of probability density modeling, we will focus on the KLD for illustrative purposes. Consider iid data being drawn from a distribution $q(\mathbf{x})$. Suppose that a topographic map of the form (42) is utilized. We will rename the output of the network as $p(\mathbf{x})$ to emphasize that it is, in fact, an approximation of the pdf $q(\mathbf{x})$. Now, we can optimize the centers and the width of the kernels to minimize the following KLD criterion:

$$\begin{aligned} D_{\mathrm{KL}}(p;q) &= \int p(\mathbf{x}) \log \frac{p(\mathbf{x})}{q(\mathbf{x})} d\mathbf{x} \\ &= \int p(\mathbf{x}) \log p(\mathbf{x}) d\mathbf{x} - \int p(\mathbf{x}) \log q(\mathbf{x}) d\mathbf{x} \\ &\cong \frac{1}{M} \sum_{i=1}^{M} \log \frac{1}{M} \sum_{j=1}^{M} K(\mathbf{w}_i - \mathbf{w}_j; \Lambda_{\mathbf{w}}) \\ &\quad - \frac{1}{M} \sum_{i=1}^{M} \log \frac{1}{N} \sum_{j=1}^{N} K(\mathbf{w}_i - \mathbf{x}_j; \Lambda_{\mathbf{x}}). \end{aligned} \tag{47}$$

Here the mixture components are centered at $\mathbf{w}_i$ and have bandwidths defined by the covariance matrices (diagonal for simplicity) $\Lambda_{\mathbf{w}}$ in the multidimensional data space. Similarly, the data is represented by a Parzen window density estimate with kernels centered at each data point and with kernel sizes determined by covariance matrices $\Lambda_{\mathbf{x}}$. The data representation and mixture diversity issues are automatically and naturally addressed by this criterion, since the first term essentially maximizes the entropy of the component centers (i.e., diversity), while the second term guarantees accurate data representation by manipulating the centers to the actual data samples drawn from the underlying distribution. An important advantage of this framework has been shown to be the guaranteed global optimization via kernel annealing [17]. The kernel size initially starts from a very large value to result in a smooth criterion (similar to convolution smoothing in global optimization literature), where the resolution of data representation is extremely coarse. The kernel size is allowed to decrease gradually (similar to the temperature in stochastic annealing), while increasing the resolution and data representation capabilities.

The principles presented above under topographic map optimization according to ITL principles can be easily combined with the information theoretic mixture model order selection procedures such as the NEC. This marriage of information theoretic order selection and ITL will result in globally optimal and robust mixture models, assuming that the kernel function is appropriately selected.

SYSTEM IDENTIFICATION USING THE MINIMUM ERROR ENTROPY CRITERION

System identification refers to the problem of optimally determining the parameters of a preselected model to represent the functional relationship between an input variable and an output variable. The case where the output variable takes discrete labels (whose actual values are inconsequential in the domain of pattern recognition) is not considered relevant in this discussion. Specifically, when the output variable takes continuous values, the traditional optimality criterion to optimize the parameters (or the weights) of the model is least squares, or equivalently the MSE [25]. The adaptive filtering and neural network theories deal with the learning of linear and nonlinear (and perhaps dynamical) relationships between these input and output variables. In supervised learning, the error is defined as the difference between the desired output corresponding to a particular input sample and the output of the adaptive system. While MSE has been traditionally the most extensively utilized criterion, various extensions to other even moments of the output error (such as the L_1-norm error and least-mean-fourth-power) as well as the ε-insensitive error function that stems from the support vector theory have also been investigated [50], [57].

The use of minimum error entropy (MEE) criterion as an alternative to MSE in supervised parametric model training (e.g., neural networks) has been studied by the authors [16], [17]. While the main advantage of MEE over MSE has been shown to be better generalization from the same training data [16], an unexpected gain was due to the specific nonparametric entropy estimator utilized in the process: namely kernel-density-plug-in estimate. It was observed that the kernel size acts as a smoothing parameter that could be annealed from an initial large value to a small *optimal* value to avoid the local minima of the optimization process, in a manner similar to stochastic annealing. An equivalence to the deterministic global

optimization scheme called convolution smoothing was conjectured based on these observations [17], and this global optimization through kernel-annealing procedure have been so far successfully used in many other applications involving optimality criteria estimated through kernels.

Given a training set with data pairs $(\mathbf{x}_k, d_k)$ and a neural network or other parametric model topology $g(\mathbf{x}; \mathbf{w})$, for the current weights $\mathbf{w}$, the training error is $e_k = d_k - g(\mathbf{x}_k; \mathbf{w})$. Using these samples of error in criterion (26), the weights are optimized using gradient descent to optimize the error entropy. For example, if error entropy is to be minimized, the gradient with respect to the weights is

$$\frac{\partial H_\alpha(e)}{\partial \mathbf{w}} = \sum_{j=1}^{N}\sum_{i=1}^{N} \frac{F_\alpha(e_j|e_i)}{(1-\alpha)N^2 V_\alpha(e)} \left(\frac{\partial e_j}{\partial \mathbf{w}} - \frac{\partial e_i}{\partial \mathbf{w}}\right). \quad (48)$$

Thus, while the information force and potential of order α [45], [18] determine the contribution of each sample to the gradient via their pairwise interactions with all other samples, the topology determines the specific direction for weight updates.

A simple but effective demonstration of the superiority of MEE over MSE can be achieved by training a time-delay neural network (TDNN) to perform single-step prediction of the Mackey-Glass chaotic time series [16]. Two identical TDNN topologies are trained with MSE and MEE criteria using the same 200-sample training data set and are tested on the same 10,000-sample set independently generated from the same MG30 chaotic attractor. When the performances of these two TDNNs are compared on the test data, the superior generalization capability of the MEE-trained network is obvious from the output errors. Furthermore, the TDNN trained with MEE learns a better approximation of the probability density function of the MG30 attractor as seen in Figure 1.

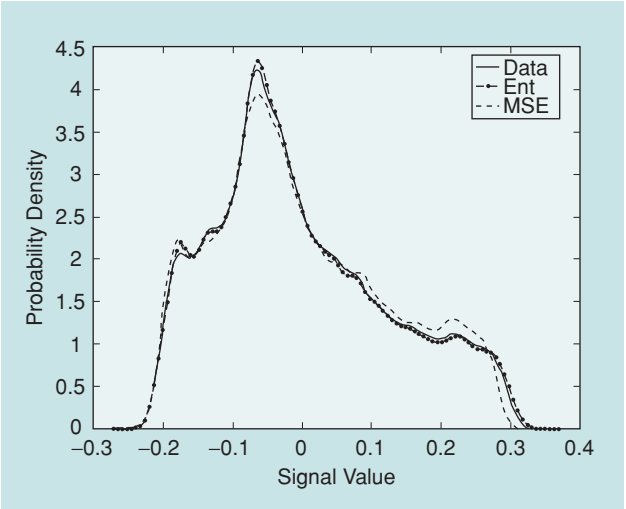
This density-matching property is not surprising; in fact, it is expected since minimizing error entropy is shown to be theoretically equivalent to minimizing the divergence between the conditional distributions of the desired output given the input and the network output given the input [16]. Furthermore, the relationship between minimum KLD and maximum likelihood principles discussed earlier illustrate that minimum entropy is also related to maximum likelihood. A recent formal result demonstrates that: 1) MEE training using the kernel density estimate is equivalent to the so-called nonparametric maximum likelihood in statistics and 2) MEE possesses certain desirable asymptotic qualities, such as outlier rejection [70]. Wolsztynski et al. demonstrate that utilizing MEE (with Shannon's definition) in the nonlinear parametric regression setting, where the noise distributions are *unknown*, is asymptotically equivalent to the employing the ML principle, thus possesses the related unbiasedness and efficiency properties in terms of parameter estimation. Although their analysis is restricted to the case where the true underlying model is part of the nonlinear parametric topology being optimized, the theoretical conclusions can be extended to situations where this is not the case simply by incorporating the residual error due to the topology's incapacity to the unknown error distribution.

Experiments conducted to illustrate the outlier rejection capabilities of MEE versus that of MSE involved generating synthetic data using an exponential function corrupted by white Laplacian noise: $d = ae^{-bx} + n$. The parameters (a, b) are estimated using MEE and MSE with 100 training samples plus 40 outlier samples with distribution $G(10, 4)$.The residual error distributions (made symmetric by creating a data set consisting of e and $-e$ samples) for MEE and MSE optimal solutions are shown in Figure 2. Clearly, MEE does a very good job in ignoring the outliers, while the MSE estimate is corrupted significantly.

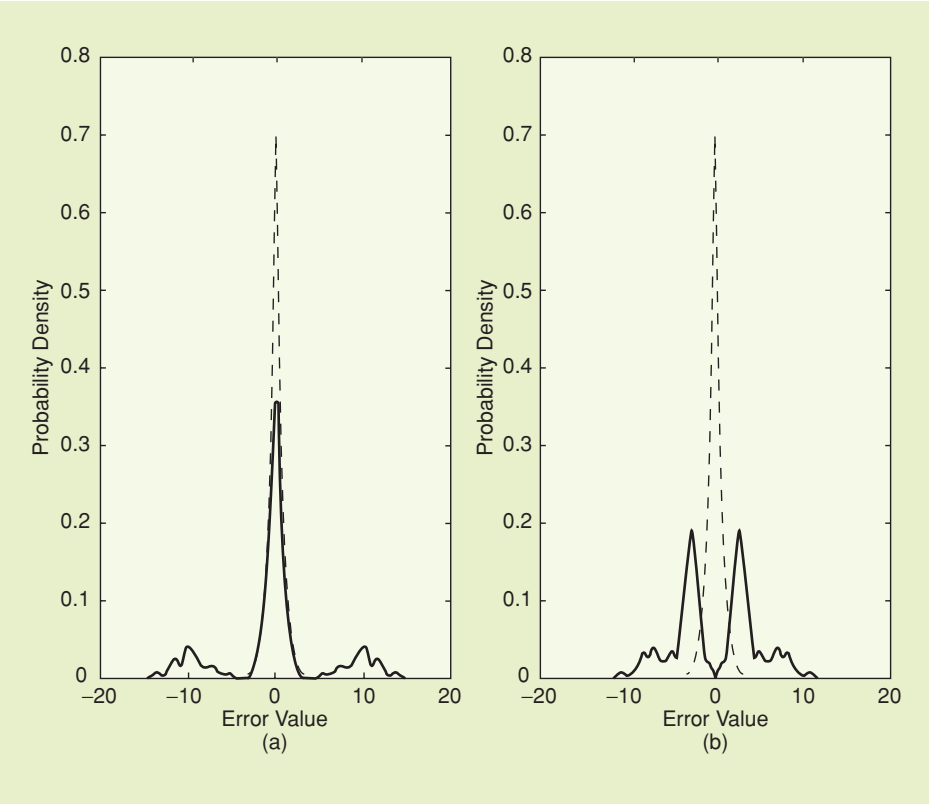
It is not unreasonable to expect similar outlier rejection capabilities from other orders of Renyi's entropy in supervised training. In fact, due to the additional flexibility offered by the entropy order parameter α, it might be even possible to obtain better outlier rejection properties for other orders of entropy. This is clearly seen from the order-α information force expression in (32). Certainly, one can manipulate this parameter to emphasize sparse/dense regions of the data distribution as desired and to improve performance over Shannon's entropy definition.

INDEPENDENT COMPONENTS ANALYSIS VIA MINIMUM MUTUAL INFORMATION

Independent components analysis (ICA) is a generalization of the PCA concept that strengthens the uncorrelated-components condition to mutual independence. It has flourished in the context of blind source separation [33], however, recently it has increasingly found many other applications areas in data analysis, visualization, and dimensionality reduction [22], [64]. The problem of ICA is itself mathematically interesting and has branched out from the original and simplest linear instantaneous square mixture setting to nonlinear,



[FIG1] Probability densities of MG30 test series (solid) and its predictions by MEE-TDNN (dash-dot) and MSE-TDNN (dotted). Reprinted with permission from [52].



[FIG2] Residual error probability distributions over regression error values: (solid) for (a) MEE and (b) MSE optimized models with 100 samples corrupted by Laplacian noise (dashed) and 40 outliers with distribution $N(10,4)$. Reprinted from [70], with permission from Elsevier.

convolutive, and nonsquare mixtures. For illustration purposes, we will restrict our discussion to the simplest case of square linear mixtures.

Various statistical criteria have been proposed and investigated to solve the ICA problem, including cumulants, negentropy, and joint output entropy [33]; the mutual information between the so-called *separated outputs* has been recognized as the "canonical contrast function" for ICA [6]. Not surprisingly, many of these higher-order statistical, criteria-based algorithms have been eventually shown to correspond to minimizing various approximations of output mutual information [33]. Most notably, the widely employed kurtosis-based separation algorithms can be shown to be a course approximation to mutual information when the entropy approximation is obtained by assuming that the data distribution is given by a reference Gaussian distribution multiplied by a fourth-order polynomial (a truncated polynomial series expansion). In essence, most ICA algorithms can be understood as some form of mutual information minimization.

Suppose that a measurement vector of mixed signals are available and the underlying generative model for these mixtures is $\mathbf{z} = \mathbf{H}\mathbf{s}$, where the mixture $\mathbf{z}$ and the source $\mathbf{s}$ are real-valued n-dimensional vectors and $\mathbf{H}$ is the unknown mixing matrix. (Without loss of generality, we assume that $E[\mathbf{z}] = \mathbf{0}$.) The goal of ICA is to determine the separation matrix that will transform $\mathbf{z}$ into $\mathbf{y}$, such that the entries of $\mathbf{y}$ are independent. This can be achieved in two stages: whitening and rotation. Whitening simply refers to transforming the mixtures such that the result has a covariance matrix of identity. This can be simply achieved by determining the eigendecomposition of the mixture covariance matrix $\Sigma_{\mathbf{z}} = \mathbf{Q}\Lambda\mathbf{Q}^T$. The whitening matrix is then $\mathbf{W} = \Lambda^{-1/2}\mathbf{Q}^T$ and $\mathbf{x} = \mathbf{W}\mathbf{z}$ has identity covariance. The whitened mixture $\mathbf{x}$ is then made independent by determining the rotation matrix $\mathbf{R}$ that minimizes the output mutual information. Since the mutual information is the sum of marginal output entropies minus their joint entropy and since the joint entropy is invariant to rotations, the cost function reduces to minimizing the sum of marginal entropies. While the whitening stage is necessary for algorithms based on maximizing non-Gaussianity based on the minimization of the sum of marginal output entropies (or their approximations), it is suggested as a useful initialization step for all algorithms, since many researchers have observed that this procedure increases the speed of convergence.

The rotation matrix is typically parameterized using Givens angles, where each angle determines the rotation in a principal plane of the n-dimensional vector space. Specifically, the overall rotation matrix is given by

$$R(\theta) = \prod_{i=1}^{n-1} \prod_{j=i+1}^{n} R_{ij}(\theta_{ij}) \tag{49}$$

where R_{ij} is the rotation matrix in the ij-plane. While Shannon's mutual information is typically used, which leads to the sum of Shannon marginal entropies, the criterion can be more generally written as the sum of marginal Renyi entropies (to include Shannon as a special case) [30]

$$J(\theta) = \sum_{o=1}^{n} H_\alpha(y^o). \tag{50}$$

The angles can be optimized by minimizing (50) using gradient descent and updating all angles at every gradient iteration. Alternatively, the angles can be updated in a rotating manner in accordance with the Jacobi iteration scheme. The batch expression for the gradient when using entropy estimator (26) will have computational complexity $O(N^2)$, which makes the algorithm prohibitively slow. Therefore, the SIG is typically preferred even in offline training. Multiple epochs of

SIG updates always lead to the optimal solution one would obtain using batch gradient, while the stochastic nature of SIG also usually helps avoid local minima in the cost function [19], [31].

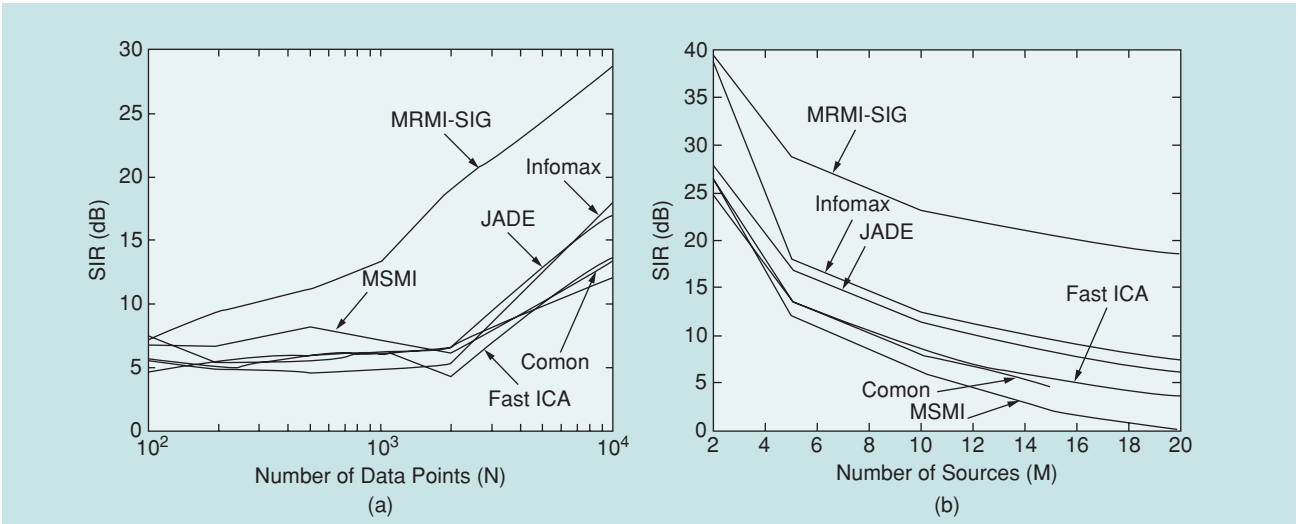
An extensive comparison of various algorithms in separating audio recordings was performed, and a summary of results obtained from the most popular and best-performing algorithms have been provided below, specifically, criterion (50) for Renyi's quadratic entropy minimized using SIG (MRMI-SIG), same criterion for Shannon entropy (MSMI), JADE [5], FastICA [33], Comon's MI method [9], and InfoMax [3]. The experiment consists of separating artificially mixed speech and music randomly selected from a pool of 50 source signals (24 speech, 26 music). In each Monte Carlo run, the mixing matrix is also randomly selected. In a simulation with $M = 5$ sources, where the number of samples N is varied MRMI-SIG consistently outperforms the other algorithms in terms of the signal-to-interference ratio (SIR) of the solution, as shown in Figure 3(a). Repeating a similar experiment with fixed (10,000) samples and varying the number of sources also yields the same comparative result, shown in Figure 3(b).

ENTROPY IS A GENERALIZATION OF VARIANCE TO PROCESSES WITH NON-GAUSSIAN DISTRIBUTIONS.

MAXIMALLY DISCRIMINATIVE PROJECTIONS VIA MUTUAL INFORMATION

Dimensionality reduction is an important procedure in pattern recognition due to at least two important potential benefits: filtering out irrelevant components while maintaining discriminative information in the original high-dimensional features and increasing implementation practicality by reducing computational complexity and memory requirements. In the pattern recognition context, input dimensionality reduction is performed in two ways: wrapper and filter. The wrapper approach attempts to determine the best projection specific to a classifier topology and therefore requires repeated training and testing of the classifier until an acceptable result is achieved. While this technique determines a good projection for the specific classifier, it also eliminates the flexibility of employing another classifier topology. On the other hand, the filter approach tries to determine the best projection based on the optimization of a suitable criterion independent from the specific classifier topology to be utilized. Although the filter approach does not optimize the feature projection based on the classification error of a certain topology, with a suitable selection of the optimality criterion, its results can be expected to be good and consistent over a wide range of classifier topologies.

The literature is rich in feature selection (mostly in the wrapper context), linear and nonlinear feature projection, and feature extraction algorithms based on a variety of optimality criteria. While the traditional approach to dimensionality reduction involves PCA, LDA, and their variations [13], the consensus among many researchers is that the mutual information between the projected features and the class labels is the natural measure of discriminability. This is mainly motivated by the information theoretic bounds on the classification performance: specifically Hellman and Raviv's bound [28] and extended Fano's bounds using Renyi's mutual information [18]. However, due to practical difficulties in estimating this quantity, various heuristic and principled approximations have been employed. Especially, the difficulties in estimating mutual information in high dimensions motivate heuristic single-dimensional approaches in feature selection. Recently, Torkkola proposed utilizing the quadratic mutual information measures based on the divergence measures (16) and (17) and the kernel density estimates of these quantities



[FIG3] **(a) SIR versus number of samples for five audio sources and (b) SIR versus number of sources for 10,000 samples. Reprinted with permission from [32].**

[64]. He has shown that discriminative information in the high-dimensional features can be effectively preserved through linear and nonlinear projections optimized through ITL. Alternatively, other divergence measures and other discriminability measures stemming from information theoretic measures can be utilized. The effective reason for Shannon mutual information being a good discriminability measure is that maximizing it amounts to maximizing the overall data entropy while minimizing the within-class entropies. In essence, this is a generalization of the Fisher discriminant principle to non-Gaussian distributions. Consequently, it is not crucial to use Shannon's definition of mutual information to this end. A broad family of discriminability criteria that is suitable for determining *optimal* subspace projections are given by Renyi's entropy

DIVERGENCE, ONE CAN THINK OF IT AS A GENERALIZATION OF ALGEBRAIC DISTANCE MEASURES (SUCH AS THE EUCLIDEAN NORM) TO PROBABILITY DISTRIBUTION SPACES.

$$J(\mathbf{w}) = H_\alpha(\mathbf{y}) - \sum_c p_c H_\alpha(\mathbf{y}\,|c). \tag{51}$$

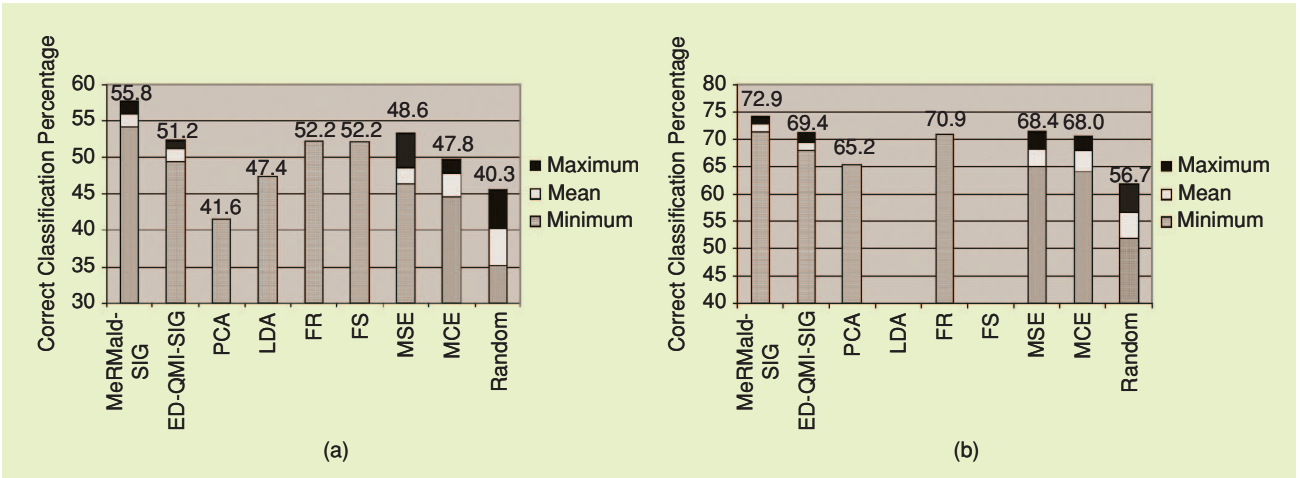
In (51), $\mathbf{y} = \mathbf{g}(\mathbf{x}, \mathbf{w})$ is the projected feature, $\mathbf{x}$ is the original high-dimensional feature, p_c is the class prior, and $\mathbf{g}(., \mathbf{w})$ is the parametric projection topology. The necessary entropy terms can be easily estimated nonparametrically from samples: the first term is calculated using all samples and (26), while the conditional entropy terms are estimated using only the samples from the corresponding class. The optimization can again be achieved using a gradient-based algorithm. Using SIG will also improve computational efficiency. An important consideration in multidimensional projections is to maintain mutual independence of the individual projections as much as possible. For example, in the case of linear projections, this can be achieved by enforcing an orthogonality constraint as in PCA.

A comparison of various methods including wrapper and filter approaches as well as traditional and information theoretic techniques have been performed over three benchmark UCI datasets: Pima, Landsat, Letter [65]. Using the following methods, features from all three datasets are projected to lower dimensionalities (from one to the original data dimensionality) and classification errors are averaged over Monte Carlo runs and datasets: criterion in (51) for Renyi's quadratic entropy optimized using SIG (MeRMaId-SIG), Euclidean-distance based mutual information measure as proposed by Torkkola optimized using SIG (ED-QMI-SIG) [64]; PCA, LDA, feature ranking individually based on their classification error (FR); feature selection considering all combinations in a brute-force manner (FS), wrapper approach based on classifier MSE with class labels (MSE) [8]; wrapper approach based on classification error (MCE) [35]; and random projection weights in $[-1, 1]$ (random).

The summarized results in Figure 4 demonstrate that the discriminability measure in (51) works effectively in determining the linear projection direction that maximizes classification performance. The qualitative ordering of algorithms is also preserved for projections to higher dimensionalities.

AN INFORMATION THEORETIC FRAMEWORK FOR SPECTRAL CLUSTERING

Clustering is an important problem in machine learning that has various applications in a number of data-oriented disciplines. Most researchers agree that the current state-of-the-art in clustering is *spectral clustering*, a special case of pairwise affinity-based clustering that is based on the eigendecomposition of an affinity graph constructed nonparametrically using pairwise interactions of samples. While Fiedler is perhaps the first to notice that clustering could be achieved using the eigenvectors of the Laplacian of the connectivity graph [20], more recent results demonstrate clearly that the definition of



[FIG4] (a) Correct classification rates for projections to one dimension and (b) average correct classification rate over all projection dimensionalities. Note that FS and FR are identical for projections to a single dimension.

the Laplacian could be relaxed by defining affinity matrices in different ways according to context and clustering can be achieved by determining the optimal graph cut [44], [52], [54]. Most notably, it has been shown that the normalized cut algorithm is shown to be related to Markov random walks [41], [54] and there are theoretical conditions under which these algorithms will separate the clusters correctly. The construction of the affinity matrix through the use of kernel functions is the general trend, inspired by the intuition offered regarding the operation of the kernel machines (e.g., support vector machines). Kernel machines operate **linearly** by transforming the data to a high-dimensional feature space where inner product operations are sufficient to solve the task. Perhaps the most widely used spectral clustering algorithm is based on this principle [52], [43]. A current practical problem to which a theoretical answer is not known is what is a good kernel for the given problem. The current state of the art is basically to solve the problem with a variety of kernel choices and then select the best performing one.

The clustering problem is really one of discrimination. Consequently, a natural criterion for discriminability is the separation between the distributions of the clusters. Information theoretic or algebraic density divergence/distance measures provide a wide range of possibilities to measure cluster separability. Suppose that we assume the CSD of (17). Assuming a two-cluster problem for illustration purposes, we denote the current estimates of the two cluster distributions by $p_1(\mathbf{x})$ and $p_2(\mathbf{x})$. Given a sample set $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, using the current membership assignments we create two $N \times 1$ membership vectors $\mathbf{m}_1$ and $\mathbf{m}_2$ for the clusters, such that $\mathbf{m}_c$ has ones at entries corresponding to samples assigned to cluster c and zeros otherwise. Now, employing a weighted kernel density estimator for each cluster using the samples assigned to that cluster $\{\mathbf{x}_1^c, \ldots, \mathbf{x}_{N_1}^c\}$, we obtain the estimates

$$p_c(\mathbf{x}) = \frac{1}{\Omega_c} \sum_{t=1}^{N_c} \alpha_t^c K_\sigma \left(\mathbf{x} - \mathbf{x}_t^c\right) \tag{52}$$

where $\Omega_c = \alpha_1^c + \cdots + \alpha_{N_c}^c$. The weights introduced to the estimator determine how much each sample contributes to the density estimate, and they can be interpreted as the Lagrange coefficients in support vector machines. Their significance in spectral clustering will become apparent shortly. Substituting (52) in (17) and simplifying the expression with the standard calculations used in all of the previous examples, we obtain [34]

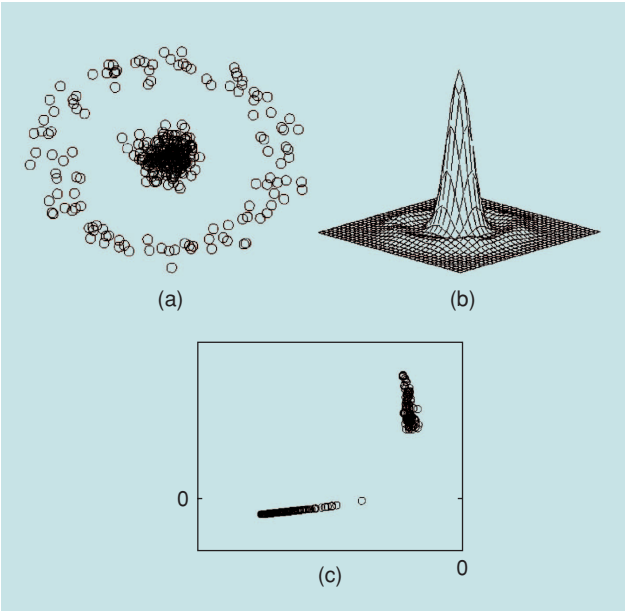
$$D_{CS}(p_1, p_2) \approx -\log \frac{\mathbf{m}_1^T \mathbf{K}_\alpha \mathbf{m}_2}{\sqrt{(\mathbf{m}_1^T \mathbf{K}_\alpha \mathbf{m}_1)(\mathbf{m}_1^T \mathbf{K}_\alpha \mathbf{m}_2)}} \tag{53}$$

where $\mathbf{K}_\alpha$ is the overall data affinity matrix with entries $\mathbf{K}_{\alpha,ts} = \alpha_t K_{\sqrt{\sigma_1^2+\sigma_2^2}}(\mathbf{x}_t, \mathbf{x}_s)\alpha_s$ for $s, t = 1, \ldots, N$.

A typical spectral clustering algorithm will utilize the eigenvectors of the affinity matrix (which might be constructed in different ways). In particular, we are interested in the algorithms that use the eigenvectors as the projections to a high dimensional feature space, as in the kernel-machine context [43], [52], [54]. Proceeding as usual, the eigendecomposition of the affinity matrix is determined $\mathbf{K}_\alpha = \Phi_\mathbf{x}^T \Lambda \Phi_\mathbf{x}$, where $\Phi_\mathbf{x}^T$ is the orthonormal eigenvector matrix and Λ is the diagonal eigenvalue matrix. The columns of $\Phi_\mathbf{x}$ correspond to the feature-transformed data points in accordance with standard spectral clustering and kernel-machine practice (see Figure 5). Substituting the eigendecomposition in (53) and defining the cluster mean vectors in the feature space as $\mu_c = (1/N_c)\Phi_\mathbf{x}\mathbf{m}_c$, the equivalent optimality criterion in the kernel induced feature space for this clustering problem becomes (denoting by $\langle \bullet, \bullet \rangle_\Lambda$ the weighted inner product as in the Mahalanobis distance)

$$\begin{aligned} D_{\mathrm{CS}}(p_1, p_2) &\approx -\log \frac{\mu_1^T \Lambda \mu_2}{\sqrt{(\mu_1^T \Lambda \mu_1)\,(\mu_2^T \Lambda \mu_2)}} \\ &= -\log \frac{< \mu_1, \mu_2 >_\Lambda}{||\mu_1||_\Lambda \cdot ||\mu_2||_\Lambda}. \end{aligned} \tag{54}$$

Since the goal is to assign memberships to samples that maximize separability, the vectors $\mathbf{m}_c$ must be selected such that (53) and (54) are maximized. This expression explains why the normalization of the samples to unit norm is crucial in the proposed heuristic approaches that employ, for example C-means [43] on the transformed data. In the density divergence framework, normalization comes naturally from the criterion. This is also why the algorithm of Shi and Malik [54] works extremely well: it corresponds to a normalized density distance measure in the original data space. In fact, in the above illustration, if the sample weights of (52) are selected to



[FIG5] Illustration of spectral clustering: (a) synthetic data set, (b) estimated data probability distribution, and (c) transformed data in the feature space (projected to the first two principal eigenvectors).

be $\alpha_s = p^{-1/2}(\mathbf{x}_s)$, where $p(\mathbf{x})$ is the overall data distribution evaluated using traditional kernel density estimation, $p(\mathbf{x}_s) \approx (1/N)\sum_{t=1}^{N} K_\sigma(\mathbf{x}_s, \mathbf{x}_t) \triangleq d_s$ and the diagonal matrix $\mathbf{D} = \mathrm{diag}\{d_1, \ldots, d_N\}$ is defined, the affinity matrix $\mathbf{K}_\alpha$ becomes $\mathbf{K}_f = \mathbf{D}^{-1/2}\mathbf{K}\mathbf{D}^{-1/2}$, where the matrix $\mathbf{K}$ is simply constructed using the entries $K_\sigma(\mathbf{x}_s, \mathbf{x}_t)$ for its (s, t) entry. The resultant affinity matrix $\mathbf{K}_f$ is equivalent to the graph Laplacian that is used in earlier work on spectral clustering [43]. Under the density divergence framework, this leads to an interesting interpretation of the Bayesian risk function that this particular spectral clustering algorithm tries to optimize. It is straightforward to demonstrate that, for this choice of the weights the asymptotic Bayesian risk function of the criterion in (53) is approximately

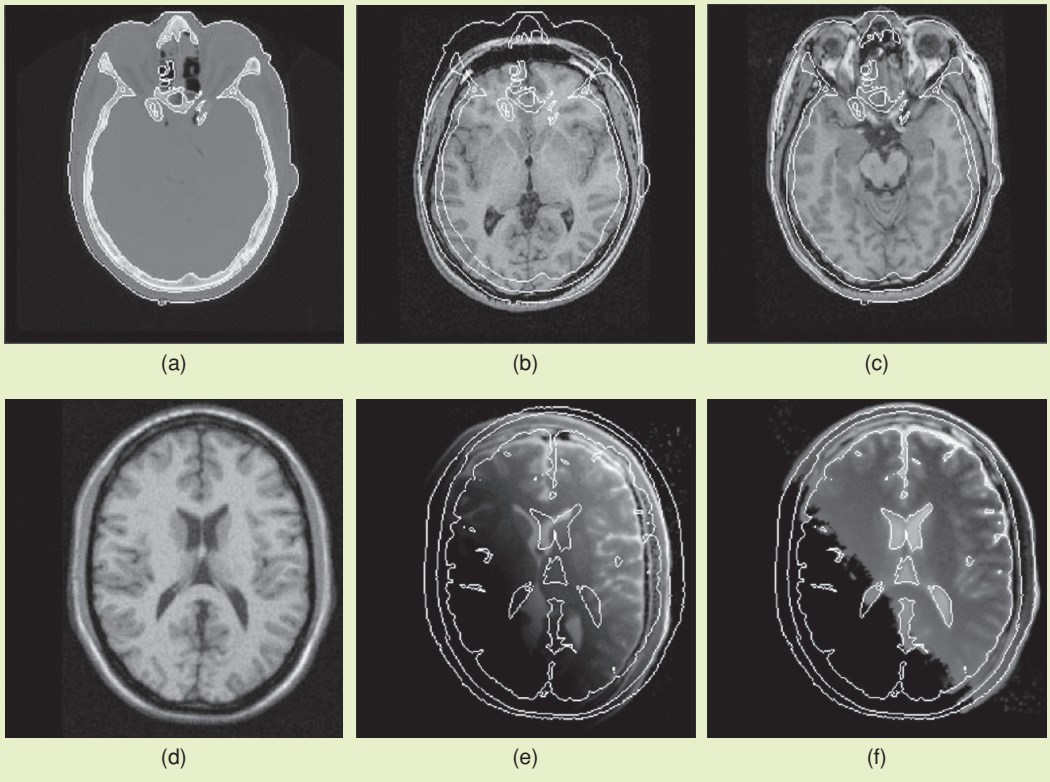
$$\begin{aligned}\text{Risk} = &\sqrt{\frac{q_2}{q_1}}P(\text{Decide } 2 \mid \text{True } 1) \\ &+ \sqrt{\frac{q_1}{q_2}}P(\text{Decide } 1 \mid \text{True } 2)\end{aligned} \tag{55}$$

where q_c denotes the cluster prior probabilities [34]. Therefore, the clustering algorithms based on the graph Laplacian penalize clustering errors in samples belonging to less-likely clusters more.

MULTIMODAL IMAGE REGISTRATION

In a variety of signal processing problems, measurements come from different sensors working under measurement modalities, thus creating the problem of how to fuse the information collected from these various sensors. For example, in video-audio processing, image frames are measured synchronously with accompanying audio, and it is reasonable to assume that the movements of some physical objects in the video are related to some temporal aspects of the audio being recorded [21]. In medical image registration, on the other hand, measurements are taken using different imaging modalities (e.g., MR, CT), where the alignment of the images requires exploiting the *correlated* spatial behavior between the images, such as edges.

In recent years, there has been an increasing amount of interest in employing ITL techniques for solving these multimodal signal processing problems. This interest is backed up by the formal representation of physical processes as Markov chains of cause-effect relationships. The introduction of probabilistic points of views is naturally accompanied by the use of information theoretic tools. An important preprocessing step in multimodal signal processing is the extraction of appropriate features for all measurement modalities, and this selection must be done in a principled manner. Maximizing the mutual information between the features obtained from different



[FIG6] Image registration using the normalized entropy criterion. CT-MR registration: (a) CT image, (b) MR image with contours of CT overlaid, and (c) affine registration of CT and MR using the normalized entropy criterion between edge features. MR-MR registration: (d) reference MR image, (e) MR image with strong bias field and contours of reference MR overlaid, and (f) affine registration of the two MR images using the normalized entropy criterion between intensity features. Reprinted with permission from Elsevier, Copyright (2005).

modalities seem to be the natural choice [21], [63]; however, certain robustness requirements might impose normalization constraints on the optimality criterion for feature extraction. For example, in medical image registration, the immediate use of mutual information was observed to be problematic, leading to the normalized entropy criterion [62], [63] (not to be confused with (41) in model order selection)

$$\mathrm{NE}(A,B) = \frac{H_S(A)+H_S(B)}{H_S(A,B)} = \frac{I_S(A,B)}{H_S(A,B)} + 1 \qquad (56)$$

where A and B are random variables corresponding to images from the two modalities. Maximizing (56) corresponds to maximizing mutual information between the features, while keeping their joint entropy low. Therefore, it could be regarded as introducing a *feature efficiency* constraint to the optimal feature extraction problem. Other normalization terms are also possible, such as the sum of marginal feature entropies instead of their joint entropy, as well as measures that utilize Renyi's definitions to control the emphasis that one puts on different density regions in each component. Under the traditional Gaussianity assumption, feature efficiency coefficients of the type presented in (56) will clearly reduce to measures based on covariance [49].

INFORMATION-THEORETIC TECHNIQUES HAVE FOUND WIDESPREAD APPLICATION IN ADAPTIVE OPTIMIZATION OF MIXTURE DENSITY MODELS AS WELL AS SELF-ORGANIZING MAPS AND OTHER TOPOGRAPHIC MAPS.

To illustrate these concepts, we focus in the medical image registration example and present some results on alignment. Conventional medical image registration approaches based on mutual information or other information theoretic measures rely on the gray-scale intensity levels as the feature of choice in determining the correct alignment between two images [62], [68]. An alternative feature of choice would be the edge information, which could be estimated using standard edge-detection techniques, for example. These two features are compared in CT-MR image registration using the local gradient as a measure of *edginess* [63], where it has been shown that while registration based on the mutual information of intensity features fail to register the images correctly in the case of affine transformations, the edge features achieve correct registration with the optimization of the normalized entropy criterion in (56). The existence of strong magnetic field bias due to inhomogeneity in the MR measurements also causes problems for registration algorithms [58] and in mutual information based registration using intensity features this might create catastrophic results if the bias distortions are strong enough to mask the actual image. The conventional approach to resolve this difficulty is to correct the bias problem with a minimum entropy criterion [42], [68] before proceeding with the registration (which is typically followed by segmentation). In fact, the so-called normalized entropy criterion (which could be renamed as the normalized mutual information criterion) can achieve both goals simultaneously. As seen in Figure 6, the normalized entropy criterion can be used to successfully register multimodal and unimodal images under both scenarios with appropriate features.

CONCLUSIONS

Adaptive filtering techniques have been an integral part of optimal signal processing. Especially online adaptation rules, such as the LMS algorithm, are extensively used in optimal real-time signal processing where a priori design of such filters are not possible or feasible. An increasingly larger number of contemporary such signal processing applications demand for more advanced filter topologies and optimality criteria that extract information more efficiently from measured signals. Higher-order statistics and especially information theoretic optimality measures attract ever-increasing attention to this end, since they provide a natural framework where the information content of available data can be assessed. These information theoretic measures also provide the basis of a unified adaptive information filtering methodology that improves performance and facilitates online implementation.

In this article, we reviewed the ITL framework based on the smooth Parzen window estimates of entropy and divergence. The ITL methodology allows simple gradient-based learning rules to be constructed for the optimization of nonlinear filter topologies under information theoretic optimization criteria under extremely general conditions. The performance of the filters designed using information theoretic criteria is, in almost all realistic scenarios, better than the traditional solutions offered by second-order statistical criteria. Here we presented applications of ITL to density estimation, nonlinear system identification, blind source separation, data dimensionality reduction, clustering, and multimodal information fusion. While these sample applications were selected for illustrative purposes, the ITL framework is applicable to any signal processing that requires the data-oriented optimal design of filter topologies.

ACKNOWLEDGMENTS

The authors would like to thank Elsevier and IEEE for granting permission to reprint Figures 1–3 and 6, Kenneth E. Hild II for providing Figure 4, Robert Jenssen for providing Figure 5, and to the reviewers and editors for their valuable comments. This work was supported by NSF under grants ECS-9900394, ECS-0300340, and ECS-0524835.

AUTHORS

Deniz Erdogmus (derdogmus@ieee.org) received B.S. degrees in electrical and electronics engineering and in mathematics in 1997, an M.S. degree in electrical and electronics engineering in 1999 from the Middle East Technical University, Turkey, and a

Ph.D. degree in electrical and computer engineering from the University of Florida in 2002. He was a research engineer at TUBITAK-SAGE, Turkey, and a research assistant and a postdoctoral research associate at the University of Florida. Currently, he is an assistant professor in the Departments of Computer Science and Electrical Engineering and Biomedical Engineering at the Oregon Health and Science University. His research focuses on information theoretic adaptive signal processing and its applications to biomedical signal processing problems, including brain interfaces. He has over 40 journal articles, book chapters, and conference papers. He is also an associate editor and guest editor for various journals, participates in various conference organization and scientific committees, and is a member of Tau Beta Pi, Eta Kappa Nu, and IEE. He is a Member of the IEEE. He received the IEEE-SPS 2003 Best Young Author Paper and 2004 INNS Young Investigator Awards.

THE GOAL OF THIS ARTICLE IS EXACTLY TO OUTLINE THE FRAMEWORK AND THE ALGORITHMS NEEDED TO MOVE FROM QUADRATIC TO INFORMATION COSTS, WHICH LEAD TO ADAPTIVE INFORMATION FILTERING.

Jose C. Principe is a distinguished professor of electrical and biomedical engineering at the University of Florida. He joined the University of Florida in 1987, after eight years as a professor at the University of Aveiro, Portugal. He received degrees in electrical engineering from the University of Porto (B.S.), Portugal, University of Florida (M.S. and Ph.D.), and a Laurea Honoris Causa degree from the Universita Mediterranea, Italy. His interests lie in nonlinear non-Gaussian optimal signal processing and modeling and in biomedical engineering. He created the Computational NeuroEngineering Laboratory. He is a Fellow of the IEEE, past president of the International Neural Network Society, and editor in chief of *IEEE Transactions on Biomedical Engineering* as well as a former member of the Advisory Science Board of the FDA. He holds five patents and has submitted seven more. He was supervisory committee chair of 47 Ph.D. and 61 master's students, and he is author of more than 400 refereed publications.

REFERENCES

[1] J. Beirlant, E.J. Dudewicz, L. Gyorfi, and E.C. van der Meulen, "Nonparametric entropy estimation: An overview," *Int. J. Math. Statist. Sci.*, vol. 6, no. 1, pp. 17–39, 1997.

[2] J.F. Bercher and C. Vignat, "Estimating the entropy of a signal with applications," *IEEE Trans. Signal Processing*, vol. 48, no. 6, pp. 1687–1694, 2000.

[3] A. Bell and T. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Comput.*, vol. 7, pp. 1129–1159, 1995.

[4] C. Biernacki, G. Celeux, and G. Govaert, "An improvement of the NEC criterion for assessing the number of clusters in a mixture model," *Pattern Recognit. Lett.*, vol. 20, no. 3, pp. 267–272, 1999.

[5] J.F. Cardoso and A. Souloumiac, "Blind beamforming for non-Gaussian signals," *Proc. Inst. Elect. Eng.*, pt. F, vol. 140, no. 6, pp. 362–370, 1993.

[6] J.F. Cardoso, "Blind signal separation: Statistical principles," *Proc. IEEE*, vol. 86, no. 10, pp. 2009–2025, 1998.

[7] G. Celeux and G. Soromenho, "An entropy criterion for assessing the number of clusters in a mixture model," *J. Classification*, vol. 13, pp. 195–212, 1996.

[8] C. Chatterjee and V. Roychowdhury, "Statistical risk analysis for classification and feature extraction by multilayer perceptrons," in *Proc. ICNN '96*, 1996, pp. 1610–1615.

[9] P. Comon, "Independent component analysis: A new concept?" *Signal Processing*, vol. 36, no. 3, pp. 287–314, 1994.

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

[11] I. Csiszar and J. Korner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. New York: Academic, 1981.

[12] L. Devroye and G. Lugosi, *Combinatorial Methods in Density Estimation*. New York: Springer, 2001.

[13] R.O. Duda, P.E. Hart, and D.G. Stork, *Pattern Classification*, 2nd ed. New York: Wiley, 2000.

[14] R.P.W. Duin, "On the choice of the smoothing parameters for parzen estimators of probability density functions," *IEEE Trans. Comput.*, vol. 25, no. 11, pp. 1175–1179, 1976.

[15] A. Elgammal, R. Duraiswami, and L.S. Davis, "Efficient kernel density estimation using the fast Gauss transform with applications to color modeling and tracking," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 25, no. 11, pp. 1499–1504, 2003.

[16] D. Erdogmus and J.C. Principe, "An error-entropy minimization algorithm for supervised training of nonlinear adaptive systems," *IEEE Trans. Signal Processing*, vol. 50, no. 7, pp. 1780–1786, 2002.

[17] D. Erdogmus and J.C. Principe, "Generalized information potential criterion for adaptive system training," *IEEE Trans. Neural Networks*, vol. 13, no. 5, pp. 1035–1044, 2002.

[18] D. Erdogmus, "Information theoretic learning: Renyi's entropy and its applications to adaptive system training," Ph.D. dissertation, Univ. Florida, Gainsville, 2002.

[19] D. Erdogmus, J.C. Principe, and K.E. Hild II, "On-line entropy manipulation: Stochastic information gradient," *IEEE Signal Processing Lett.*, vol. 10, no. 8, pp. 242–245, 2003.

[20] M. Fiedler, "Algebraic connectivity in graphs," *Czechoslovak Math. J.*, vol. 23, no. 98, pp. 298–305, 1973.

[21] J.W. Fisher III, T. Darrell, W.T. Freeman, and P. Viola, "Learning joint statistical models for audio-visual fusion and segregation," in *Proc. Advances in NIPS*, 2000, pp. 772–778.

[22] A. Globerson and N. Tishby, "Sufficient dimensionality reduction," *J. Mach. Learning Res.*, vol. 3, pp. 1307–1331, 2003.

[23] L. Greengard, *The Rapid Evaluation of Potential Fields in Particle Systems*. Cambridge, MA: MIT Press, 1988.

[24] L. Greengard and J. Strain, "The fast Gauss transform," *SIAM J. Sci. Statist. Comput.*, vol. 12, no. 1, pp. 79–94, 1991.

[25] S. Haykin, *Neural Networks: A Comprehensive Foundation*, 2nd ed. Upper Saddle River, NJ: Prentice Hall, 1999.

[26] S. Haykin, Ed., *Unsupervised Adaptive Filtering, Volume 1: Blind Source Separation*. New York: Wiley, 2000.

[27] S. Haykin, Ed., *Unsupervised Adaptive Filtering, Volume 2: Blind Deconvolution*. New York: Wiley, 2000.

[28] M.E. Hellman and J. Raviv, "Probability of error, equivocation and the Chernoff bound," *IEEE Trans. Inform. Theory*, vol. 16, pp. 368–372, 1970.

[29] A.O. Hero III, B. Ma, O.J.J. Michel, and J. Gorman, "Applications of entropic spanning graphs," *IEEE Signal Processing Mag.*, vol. 19, no. 5, pp. 85–95, 2002.

[30] K.E. Hild II, D. Erdogmus, and J.C. Principe, "Blind source separation using renyi's mutual information," *IEEE Signal Processing Lett.*, vol. 8, no. 6, pp. 174–176, 2001.

[31] K.E. Hild II, "Blind source separation of convolutive mixtures using Renyi's divergence," Ph.D. dissertation, Univ. of Florida, Gainsville, 2003.

[32] K.E. Hild II, D. Erdogmus, and J.C. Principe, "An analysis of entropy estimators for blind source separation," *Signal Process.*, vol. 86, no. 1, pp. 182–194, 2006.

[33] A. Hyvarinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. New York: Wiley, 2001.

[34] R. Jenssen, D. Erdogmus, J.C. Principe, and T. Eltoft, "The Laplacian PDF distance: A cost function for clustering in a kernel feature space," in *Proc. Advances in NIPS*, 2005, pp. 360–367.

[35] B.H. Juang and S. Katagiri, "Discriminative learning for minimum error classification," *IEEE Trans. Signal Processing*, vol. 40, no. 12, pp. 3043–3054, 1992.

[36] T. Kohonen, *Self Organizing Maps*. New York: Springer, 1995.

[37] A.N. Kolmogorov, "Interpolation and extrapolation of stationary random sequences," *SSSR Math.*, vol. 5, pp. 3–14, 1941.

[38] D. Koller and M. Sahami, "Toward optimal feature selection," in *Proc. ICML*, 1996, pp. 284–292.

[39] L.F. Kozachenko and N.N. Leonenko, "Sample estimate of the entropy of random vector," *Probl. Inform. Transmission*, vol. 23, no. 2, pp. 95–101, 1987.

[40] J.Q. Li and A.R. Barron, "Mixture density estimation," in *Proc. Advances in NIPS*, 1999.

[41] M. Meila and J. Shi, "Learning segmentation by random walks," in *Proc. Advances in NIPS*, 2000, pp. 873–897.

[42] E.G. Learned-Miller and P. Ahammad, "Joint MRI bias removal using entropy minimization across images," in *Proc. Advances in NIPS*, 2005.

[43] A.Y. Ng, M. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Proc. Advances in NIPS*, 2001, pp. 849–856.

[44] A. Pothen, H.D. Simon, and K.P. Liou, "Partitioning sparse matrices with eigenvectors of graphs," *SIAM J. Matrix Anal. Applicat.*, vol. 11, no. 3, pp. 430–452, 1990.

[45] J.C. Principe, J.W. Fisher, and D. Xu, "Information theoretic learning," in *Unsupervised Adaptive Filtering*, S. Haykin, Ed. New York: Wiley, 2000, pp. 265–319.

[46] J.C. Principe, Y.N. Rao, and D. Erdogmus, "Error whitening Wiener filters: Theory and algorithms," in *Least-Mean-Square Adaptive Filters*, S. Haykin and B. Widrow, Eds. New York: Wiley, 2003.

[47] Y.N. Rao, D. Erdogmus, G.Y. Rao, and J.C. Principe, "Stochastic error whitening algorithm for linear filter estimation with noisy data," *Neural Networks*, vol. 16, no. 5–6, pp. 873–880, 2003.

[48] A. Renyi, *Probability Theory*. North Holland: Amsterdam, 1970.

[49] A. Roche, G. Malandin, X. Pennec, and N. Ayache, "The correlation ratio as a new similarity measure for multimodal image registration," in *Proc. Medical Image Computing and Computer-Assisted Intervention*. Cambridge, MA, 1998, pp. 1115–1124.

[50] A. Sayed, *Fundamentals of Adaptive Filtering*. New York: Wiley, 2003.

[51] N. Schraudolph, "Gradient-based manipulation of nonparametric entropy estimates," *IEEE Trans. Neural Networks*, vol. 15, no. 4, pp. 828–837, 2004.

[52] G. Scott and H. Longuet-Higgins, "Feature grouping by relocalisation of eigenvectors of the proximity matrix," in *Proc. British Machine Vision Conf.*, 1990, pp. 103–108.

[53] C.E. Shannon and W. Weaver, *The Mathematical Theory of Communication*. Urbana, IL: Univ. of Illinois Press, 1964.

[54] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, 2000.

[55] B.W. Silverman, *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall, 1986.

[56] N. Slonim and N. Tishby, "Agglomerative information bottleneck," in *Proc. Advances in NIPS*, 1999.

[57] A.Smola and B. Schölkopf, "A tutorial on support vector regression," Royal Holloway College, Univ. of London, Tech. Rep. NC-TR-98-030, 1998.

[58] E. Solanas and J.P. Thiran, "Exploiting voxel correlation for automated MRI bias field," in *Proc. Medical Image Computing and Computer-Assisted Intervention*, Utrecht, Netherlands, 2001, pp. 1220–1221.

[59] P. Stoica and R. Moses, *Introduction to Spectral Analysis*. Upper Saddle River, NJ: Prentice-Hall, 1997.

[60] P. Stoica and Y. Selen, "Model-order selection: A review of information criterion rules," *IEEE Signal Processing Mag.*, vol. 21, no. 4, pp. 36–47, 2004.

[61] J. Strain, "The fast Gauss transform with variable scales," *SIAM J. Sci. Statist. Comput.*, vol. 12, no. 5, pp. 1131–1139, 1991.

[62] C. Studholme, D.J. Hawkes, and D.L.G. Hill, "An overlap invariant entropy measure of 3D medical image alignment," *Pattern Recognit.*, vol. 32, no. 1, pp. 71–86, 1999.

[63] T. Butz and J.P. Thiran, "From error probability to information theortic multimodal signal processing," *Signal Process.*, vol. 85, no. 5, pp. 875–902, 2005.

[64] K. Torkkola, "Feature extraction by non-parametric mutual information maximization," *J. Machine Learning Res.*, vol. 3, no. 3, pp. 1415–1438, 2003.

[65] UCI Machine Learning Repository [Online]. Available: http://www.ics.uci.edu/~mlearn/MLRepository.html

[66] M.M. Van Hulle, "Joint entropy maximization in kernel-based topographic maps," *Neural Comput.*, vol. 14, no. 8, pp. 1887–1906, 2002.

[67] O. Vasicek, "A test for normality based on the sample entropy," *J. Royal Statistical Soc. B*, vol. 38, no. 1, pp. 54–59, 1976.

[68] P. Viola and W.M. Wells III, "Alignment by maximization of mutual information," in *Proc. Int. Conf. Computer Vision*, 1995, pp. 16–23.

[69] N. Wiener, *Extrapolation, Interpolation, and Smoothing of Stationary Time Series with Engineering Applications*. Cambridge. MA: MIT Press, 1949.

[70] E. Wolsztynski, E. Thierry, and L. Pronzato, "Minimum entropy estimation in semiparametric models," *Signal Process.*, vol. 85, no. 5, pp. 937–949, 2005.

[71] C. Yang, R. Duraiswami, N.A. Gumerov, and L. Davis, "Improved fast Gauss transform and efficient kernel density estimation," in *Proc. ICCV '03*, 2003, pp. 464–471.

APPENDIX
PARAMETRIC ENTROPY ESTIMATION

The parametric approach relies on assuming a family of distributions (such as the Gaussian, beta, exponential) that parametrically describes each candidate distribution. The optimal pdf estimate for the data is then determined using Bayesian techniques, such as ML or MAP. For example, the ML estimate yields $p(x; \theta_{\mathrm{ML}})$ as the density estimate by solving

$$\theta_{\mathrm{ML}} = \arg\max_{\theta} \sum_{k=1}^{N} \log p(x_k ; \theta) \tag{A.1}$$

where $p(x ; \theta)$ is the selected parametric family of distributions. It can be shown that the ML density estimate asymptotically converges to the member of the parametric family that minimizes the KLD with the true underlying density. To observe this, suppose that the samples are generated by a density $q(x)$ and a parametric family $p(x ; \theta)$ is considered for data modeling

$$\begin{aligned}
\theta_{\mathrm{KLD}} &= \arg\min_{\theta} D_{\mathrm{KL}}(q ; p_\theta) \\
&= \arg\min_{\theta} \int q(x) \log[q(x)]/[p(x ; \theta)]dx \\
&= \arg\min_{\theta} - H_S(X) - E_X[\log p(X ; \theta)] \\
&= \arg\max_{\theta} E_X[\log p(X ; \theta)] \\
&= \lim_{N\to\infty} \theta_{\mathrm{ML}}.
\end{aligned} \tag{A.2}$$

The MAP estimate asymptotically converges to the same KLD-optimal parameters since the weight of the a priori parameter distribution asymptotically diminishes. The modeling capability of the parametric approach can be enhanced by allowing mixture models (such as mixture of Gaussians). Nevertheless, the parametric entropy estimation approach encounters two difficulties in ITL applications: each adaptation step requires solving one ML model fitting procedure and insufficient model complexity for general-purpose data modeling given the typically constrained parametric models.

Another interesting approach to entropy estimation is to use the duality between the pdf of a random variable and the power spectral density (PSD) of an associated stochastic process [2]. The samples of the random variable can be used to generate the samples from which the PSD of the associated process can be estimated using traditional and established spectral estimation techniques [59]. SP