Universal Estimation of Information Measures for Analog Sources

Universal Estimation of Information Measures for Analog Sources

Qing Wang

Credit Suisse Group
New York, NY 10010
USA
qingwang@Princeton.edu

Sanjeev R. Kulkarni

Princeton University
Princeton, NJ 08544
USA
Kulkarni@Princeton.edu

Sergio Verdú

Princeton University
Princeton, NJ 08544
USA
Verdu@Princeton.edu

the essence of knowledge

Boston – Delft

Foundations and Trends® in Communications and Information Theory

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
USA
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is Q. Wang, S. R. Kulkarni and S. Verdú, Universal Estimation of Information Measures for Analog Sources, Foundations and Trends® in Communications and Information Theory, vol 5, no 3, pp 265–353, 2008

ISBN: 978-1-60198-230-8

Foundations and Trends® in Communications and Information Theory
Volume 5 Issue 3, 2008

Editorial Scope

Foundations and Trends® in Communications and Information Theory will publish survey and tutorial articles in the following topics:

- Coded modulation
- Coding theory and practice
- Communication complexity
- Communication system design
- Cryptology and data security
- Data compression
- Data networks
- Demodulation and Equalization
- Denoising
- Detection and estimation
- Information theory and statistics
- Information theory and computer science
- Joint source/channel coding
- Modulation and signal design
- Multiuser detection
- Multiuser information theory
- Optical communication channels
- Pattern recognition and learning
- Quantization
- Quantum information processing
- Rate-distortion theory
- Shannon theory
- Signal processing for communications
- Source coding
- Storage and recording codes
- Speech and Image Compression
- Wireless Communications

Information for Librarians

Foundations and Trends® in Communications and Information Theory, 2008, Volume 5, 6 issues. ISSN paper version 1567-2190. ISSN online version 1567-2328. Also available as a combined paper and online subscription.

Foundations and Trends® in
Communications and Information Theory
Vol. 5, No. 3 (2008) 265–353

DOI: 10.1561/0100000021

Universal Estimation of Information Measures for Analog Sources

Qing Wang[1], Sanjeev R. Kulkarni[2] and Sergio Verdú[3]

[1] *Credit Suisse Group, 11 Madison Avenue, New York, NY 10010, USA, qingwang@Princeton.edu*
[2] *Department of Electrical Engineering, Princeton University, Princeton, NJ 08544, USA, Kulkarni@Princeton.edu*
[3] *Department of Electrical Engineering, Princeton University, Princeton, NJ 08544, USA, Verdu@Princeton.edu*

Abstract

This monograph presents an overview of universal estimation of information measures for continuous-alphabet sources. Special attention is given to the estimation of mutual information and divergence based on independent and identically distributed (i.i.d.) data. Plug-in methods, partitioning-based algorithms, nearest-neighbor algorithms as well as other approaches are reviewed, with particular focus on consistency, speed of convergence and experimental performance.

Contents

1

Introduction

Entropy, differential entropy and mutual information, introduced by Shannon [216] in 1948, arise in the study of the fundamental limits of data compression and data transmission. Divergence, used by Wald [258] in 1945, and often attributed to Kullback and Leibler [144], also plays a major role in information theory as well as in large deviations theory. Entropy, mutual information and divergence measure the randomness, dependence and dissimilarity, respectively of random objects. In addition to their prominent role in information theory, they have found numerous applications, among others, in probability theory [13, 19], ergodic theory [218], statistics [64, 142], convex analysis and inequalities [69], physics [25, 27, 147, 150], chemistry [79], molecular biology [270], ecology [138], bioinformatics [81, 83, 214], neuroscience [201, 232], machine learning [73], linguistics [26, 44], and finance [52, 53, 56]. Many of these applications require a universal estimate of information measures which does not assume knowledge of the statistical properties of the observed data. Over the past few decades, several non-parametric algorithms have been proposed to estimate information measures. This monograph aims to present a comprehensive survey of universal estimation of information measures for memoryless analog (real-valued or real vector-valued) sources with an emphasis on the estimation of mutual information and divergence and their applications. We review the consistency of the universal algorithms and the corresponding sufficient conditions as well as their speed of convergence.

The monograph is organized as follows. In the remainder of this section, we review the concepts of information measures, their applications in theory and practice, and we formulate the universal estimation problem. Section 2 introduces plug-in algorithms and discusses their performance. Section 3 presents partitioning-based methods, gives a consistency analysis and describes the most advanced version in this class of algorithms. Section 4 investigates the nearest-neighbor approach for information estimation and studies its convergence. Other methods based on density estimation and minimal spanning trees are reviewed in Section 5. Section 6 summarizes and provides experimental results that serve as an illustration of the relative merits of the various methods. Section 7 gives a brief discussion of the estimation of mutual information rate for processes with memory.

1.1 Entropy

The concept of entropy as an information measure was introduced by Shannon [216]. The entropy $H(X)$ of a discrete random variable X is defined in terms of its probability mass function $P_X(\cdot)$:

$$H(X) = \sum_{x \in \mathcal{X}} P_X(x) \log \frac{1}{P_X(x)}. \tag{1.1}$$

Throughout this monograph, the convention $0 \log 1/0 = 0$ is used. Entropy $H(X)$ quantifies the information or uncertainty associated with X. Entropy plays a key role in fundamental limits of lossless data compression. The entropy definition (1.1) as an information measure, however, is only applicable to discrete random sources.

Generally speaking, the information estimation for discrete data is at a more advanced stage than that for analog data. In the case of entropy estimation for discrete sources, most of the work is devoted to data with memory. Back in 1951, Shannon [217] considered the estimation of entropy of English via the number of trials to guess subsequent symbols in a given text. Cover and King [51] later proposed a gambling estimate of English entropy and proved its consistency for stationary ergodic data. The Lempel–Ziv string matching method was used in [267] and [134] for entropy estimation for stationary ergodic processes. Cai et al. [43] proposed algorithms based on the Burrows–Wheeler block sorting transform to estimate entropy for finite alphabet, finite memory sources. In addition, for memoryless sources, different approaches [139, 177, 184, 265] have been designed to overcome difficulties in the under-sampled regime. We next turn our attention to information measures that can be applied to analog sources, which is the focus of this monograph.

1.2 Differential Entropy

1.2.1 Definition

Differential entropy was proposed in 1948 simultaneously by Shannon [216] and Wiener [263]. It is only defined for continuous random variables (see [55] for its basic properties). Let X is a continuous random variable with a probability density function (pdf) p_X defined on $\mathbb{R}^d$. Its differential entropy $h(X)$ is given by

$$h(X) = \int_{\mathbb{R}^d} p_X(x) \log \frac{1}{p_X(x)} \mathrm{d}x. \tag{1.2}$$

The Gaussian distribution maximizes the differential entropy over all distributions with a given covariance matrix. The exponential distribution maximizes the differential entropy over all distributions with a given mean and supported on the positive half line. Among distributions supported on a given finite interval, the differential entropy is maximized by the uniform distribution. Various explicit expressions for differential entropies of univariate and multivariate probability densities can be found in [6, 66, 148].

1.2.2 Universal Estimation

Let X is a continuous random variable in $\mathbb{R}^d$ with density p_X. Suppose $\{X_1,\ldots,X_n\}$ are i.i.d. realizations of X. A universal estimator of the

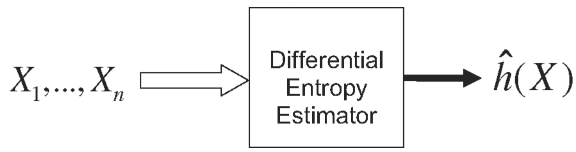

Fig. 1.1 Universal estimation of differential entropy.

differential entropy of X (see Figure 1.1) is an algorithm which outputs a consistent estimate, $\hat{h}(X)$, of $h(X)$ given only the observations $\{X_i\}$ and no knowledge of p_X. Beirlant et al. [23] provides a survey on non-parametric estimation of differential entropy for i.i.d. samples. In Sections 2, 3, 4, and 5, we review several algorithms for differential entropy estimation.

1.2.3 Applications

1.2.3.1 Quantization

Like entropy, differential entropy is closely related to data compression. For analog sources, as the quantizer becomes finer and finer, the entropy of the output behaves as the differential entropy plus the logarithm of the reciprocal of the quantization bin size. In particular, suppose $q_n(\cdot)$ is a uniform quantizer with infinitely many levels and step size $1/n$. In 1959, Rényi [199] showed that the entropy of the quantizer output $q_n(X)$ behaves as

$$H(q_n(X)) = h(X) + \log n + o(1) \tag{1.3}$$

See [29, 30, 60, 61, 91, 92, 102] for generalized results in the approximation of the quantizer output entropy via differential entropy. Those and other results on quantization are surveyed in the review by Gray and Neuhoff [98].

1.2.3.2 Asymptotic Equipartition Property

One of the most important roles of entropy arises in the asymptotic equipartition property (AEP) [56, Chapter 3], which characterizes the probability of typical sequences, namely those whose sample entropy is close to the entropy. Although not nearly as useful, a similar property holds for analog sources using differential entropy.

Definition 1.1. Let $x_1, x_2, \ldots, x_n$ is a sequence of random variables drawn i.i.d. according to the density p_X. For $\epsilon > 0$, $x_1, x_2, \ldots, x_n$ is an ϵ-typical sequence if

$$\left| \frac{1}{n} \log \frac{1}{p_X(x_1, \ldots, x_n)} - h(X) \right| \leq \epsilon, \tag{1.4}$$

where $p_X(x_1, x_2, \ldots, x_n) = \Pi_{i=1}^n p_X(x_i)$. For $\epsilon > 0$ and any n, the typical set $A_\epsilon^{(n)}$ is the collection of all sequences $x_1, x_2, \ldots, x_n$ which are ϵ-typical.

Let the volume $\mathrm{Vol}(A)$ of a set $A \subset \mathbb{R}^n$ be defined as

$$\mathrm{Vol}(A) = \int_A \mathrm{d}x_1 \mathrm{d}x_2 \cdots \mathrm{d}x_n. \tag{1.5}$$

The following theorem from [56, Chapter 8] characterizes the volume and probability of the typical set $A_\epsilon^{(n)}$ in terms of differential entropy.

Theorem 1.1. The typical set $A_\epsilon^{(n)}$ has the following properties:

(1) $\Pr\left(A_\epsilon^{(n)}\right) > 1 - \epsilon$ for n sufficiently large;
(2) $\mathrm{Vol}\left(A_\epsilon^{(n)}\right) \leq \exp\left(n(h(X) + \epsilon)\right)$ for all n;
(3) $\mathrm{Vol}\left(A_\epsilon^{(n)}\right) \geq (1 - \epsilon)\exp\left(n(h(X) - \epsilon)\right)$ for n sufficiently large.

Note that if $h(X) > 0$, then the volume of the typical set grows exponentially in the dimension. Conversely, if $h(X) < 0$, it shrinks exponentially.

1.2.3.3 Maximum Differential Entropy Principle

The principle of maximum entropy was proposed by Jaynes [120, 121, 122] in the context of thermodynamics (see also [219]). This principle is a general method to select probability distributions given partial information on their moments.

Theorem 1.2. (Maximum Differential Entropy Distribution). Let f is a probability density function supported on the set S. The unique solution to the following optimization problem:

$$\text{Maximize } h(f) \triangleq \int_S f(x) \log \frac{1}{f(x)} \mathrm{d}x,$$

subject to

$$\int_S f(x) r_i(x) \mathrm{d}x = \alpha_i, \quad \text{for } 1 \leq i \leq m. \tag{1.6}$$

is

$$f^*(x) = \exp\left(\lambda_0 + \sum_{i=1}^{m} \lambda_i r_i(x)\right), \quad x \in S, \tag{1.7}$$

where $\lambda_0, \ldots, \lambda_m$ are chosen such that the constraints (1.6) are satisfied.

The principle of maximum differential entropy has been applied to density estimation [39, 203, 235, 236] and spectral estimation [40, 48].

1.2.3.4 Entropy Power Inequalities and the Convolution of Densities

For a continuous random variable X in $\mathbb{R}^d$ with differential entropy $h(X)$, the entropy power of X is defined to be

$$N(X) = \frac{1}{2\pi e} \exp\left(\frac{2}{d} h(X)\right). \tag{1.8}$$

Entropy power inequalities relate the entropy power of the sum of independent random variables to the sums of entropy powers contained in subsets of the random variables, for an arbitrary collection of subsets. In particular, let $X_1, \ldots, X_n$ is independent random variables in $\mathbb{R}^d$, then

$$N(X_1 + \cdots + X_n) \geq \sum_{i=1}^{n} N(X_i), \tag{1.9}$$

where equality holds if and only if $X_1, \ldots, X_n$ are Gaussian random vectors with proportional covariance matrices. Since the density of the sum of independent random variables is given by the convolution of the individual densities, an alternative interpretation of (1.9) is that convolution increase entropy power.

The inequality (1.9) is put forth by Shannon [216] and proved by Stam [229] and has been strengthened in various ways in [13, 99, 159, 244]. These types of inequalities are useful for the examination of monotonicity in central limit theorems [13, 19, 159, 244] for independent random variables.

1.3 Divergence

1.3.1 Definition

While certain analogies exist between entropy and differential entropy, the differential entropy can be negative and is not invariant under invertible transformations. More useful and fundamental for the continuous case is the divergence, also known as Kullback–Leibler divergence or relative entropy, first used by Wald [258] and formally introduced by Kullback and Leibler [144] in 1951 as a measure of distance between distributions. The definition of divergence carries over directly from discrete to continuous distributions, and possesses the convenient property of being invariant under one-to-one transformations. Suppose P and Q are probability distributions defined on the same measurable space $(\Omega, \mathcal{F})$. The divergence between P and Q is defined as

$$D(P\|Q) = \int_\Omega \mathrm{d}P \log \frac{\mathrm{d}P}{\mathrm{d}Q}. \tag{1.10}$$

when P is absolutely continuous with respect to Q (denoted as $P \ll Q$, i.e. $P(A) = 0$ for any $A \in \mathcal{F}$ such that $Q(A) = 0$), and $+\infty$ otherwise. Since $P \ll Q$ implies that the Radon–Nikodym derivative $\mathrm{d}P/\mathrm{d}Q$ exists, an alternative definition of divergence is given by

$$D(P\|Q) = \int_\Omega \mathrm{d}Q \frac{\mathrm{d}P}{\mathrm{d}Q} \log \frac{\mathrm{d}P}{\mathrm{d}Q}. \tag{1.11}$$

Specifically, for distributions on a discrete alphabet $\mathcal{A}$, divergence becomes

$$D(P\|Q) = \sum_{a\in\mathcal{A}} P(a)\log\frac{P(a)}{Q(a)}. \tag{1.12}$$

where $0\log 0/0 = 0$ by convention. For continuous distributions on $\mathbb{R}^d$, if the densities of P and Q with respect to Lebesgue measure exist, denoted by $p(x)$ and $q(x)$, respectively, with $p(x) = 0$ for P-almost every x such that $q(x) = 0$, then

$$D(P\|Q) = D(p\|q) = \int_{\mathbb{R}^d} p(x)\log\frac{p(x)}{q(x)}\mathrm{d}x. \tag{1.13}$$

A useful list of explicit expressions of divergence between common pdf's is given in [188].

As a distance measure, divergence is always non-negative with $D(P\|Q) = 0$ if and only if $P = Q$. However, divergence is not symmetric and does not satisfy the triangle inequality and thus is not a metric. Other distance measures can be related to divergence by, for example, Pinsker's inequality [143, 190]:

$$D(P\|Q) \geq \frac{1}{2}V(P,Q)\log e, \tag{1.14}$$

where $V(P,Q)$ is the variational distance defined as

$$V(P,Q) = V(Q,P) = \sup_{\{A_i\}}\sum_i |P(A_i) - Q(A_i)|, \tag{1.15}$$

where the supremum is taken over all $\mathcal{F}$-measurable partitions $\{A_i\}$ of Ω. For more inequalities regarding divergence and related measures, see [34, 59, 85, 158, 238].

1.3.2 Universal Estimation

Suppose P and Q are probability distributions defined on the same Euclidean space $(\mathbb{R}^d, \mathcal{B}_{\mathbb{R}^d})$ and $P \ll Q$. Let p and q are probability density functions corresponding to P and Q, respectively. The problem

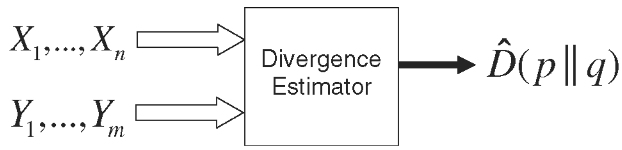

Fig. 1.2 Universal estimation of divergence $D(p\|q)$.

is to design a consistent estimate of $D(P\|Q)$ given i.i.d. samples $\{X_1,\ldots,X_n\}$ and $\{Y_1,\ldots,Y_m\}$ are drawn according to p and q, respectively (see Figure 1.2).[1] As before, in the construction of universal estimators, no knowledge is available about p and q.

1.3.3 Applications of Divergence

1.3.3.1 Mismatch Penalty in Data Compression

Assume that X is a discrete random variable drawn according to a distribution P. The average length of a prefix code is lower bounded by the entropy $H(P)$. This bound is achieved if

$$\ell(a) = \log\frac{1}{P(a)}. \tag{1.16}$$

are integers for all a, where the base of the logarithm is equal to the size of the code alphabet. On the other hand, the minimum average length of a prefix code is upper bounded by the entropy plus one.

In the case of mismatch where the choice of the code assumes a different distribution Q, the minimum average length is upper bounded by $H(P) + D(P\|Q) + 1$.

1.3.3.2 Chernoff–Stein Lemma

In binary hypothesis testing, if we fix one of the error probabilities and minimize the other probability of error, the Chernoff–Stein lemma shows that the latter will decay exponentially with exponent equal to the divergence between the two underlying distributions.

[1] Note that m and n are not required to be equal.

Theorem 1.3. (Chernoff–Stein Lemma) [56, 254].
Let $X_1, X_2, \dots, X_n \in \mathcal{A}^n$ is i.i.d. random variables distributed according to a distribution F. Consider a hypothesis testing problem:

$$\begin{aligned} H_0 : F &= P \\ H_1 : F &= Q, \end{aligned} \tag{1.17}$$

where $D(P\|Q) < \infty$. Let $D_n \subseteq \mathcal{A}^n$ be the decision region for hypothesis H_0. Let the probabilities of error be

$$\alpha_n = P^n(D_n^c), \quad \beta_n = Q^n(D_n). \tag{1.18}$$

For $0 < \epsilon < 1/2$, define

$$\beta_n^*(\epsilon) = \min_{D_n \subseteq \mathcal{A}^n, \alpha_n \le \epsilon} \beta_n. \tag{1.19}$$

Then

$$\lim_{n\to\infty} \frac{1}{n} \log \frac{1}{\beta_n^*(\epsilon)} = D(P\|Q). \tag{1.20}$$

1.3.3.3 A *Posteriori* Likelihood Result

Divergence also characterizes the limit of the log-likelihood ratio [55] and is useful in maximum likelihood detection [256, Problem 3.6]. Suppose the hypothesis testing problem is as shown in (1.17) and the distributions P and Q satisfy that

$$D(P\|Q) < \infty \text{ and } D(Q\|P) < \infty. \tag{1.21}$$

By the weak law of large numbers, if P is the true distribution, we have

$$\frac{1}{n} \log \frac{P(X_1, X_2, \dots, X_n)}{Q(X_1, X_2, \dots, X_n)} \to D(P\|Q), \quad \text{in probability;} \tag{1.22}$$

and if Q is the true distribution,

$$\frac{1}{n} \log \frac{P(X_1, X_2, \dots, X_n)}{Q(X_1, X_2, \dots, X_n)} \to -D(Q\|P), \quad \text{in probability.} \tag{1.23}$$

1.3.3.4 Capacity of Non-Gaussian Additive Channels

Channel capacity is the tightest upper bound on the amount of information that can be reliably transmitted over a communications channel. For channels with additive-noise of fixed power, Gaussian noise is shown to be least favorable [216]. Specifically, assuming the same power constraints, the capacity of non-Gaussian channels is always greater than or equal to that of Gaussian channels. An upper bound on the capacity of additive non-Gaussian noise channels depends on the "non-Gaussianness" of the noise distribution, or equivalently, the divergence between the actual noise distribution and a Gaussian distribution with the same variance.

Theorem 1.4. [216, 119].
Consider a discrete-time additive-noise channel,

$$Y_i = X_i + N_i, \quad i = 1, \dots, n, \tag{1.24}$$

where X_i and N_i are i.i.d. and

- the noise $\{N_i\}$ has distribution P_N with variance σ^2 and is independent of the input $\{X_i\}$;
- The input signals $\{X_i\}$ satisfy the power constraint (individual or on average over codebook):

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 \leq P. \tag{1.25}$$

Then channel capacity is bounded by

$$\frac{1}{2}\log\left(1+\frac{P}{\sigma^2}\right) \leq C \leq \frac{1}{2}\log\left(1+\frac{P}{\sigma^2}\right) + D\left(P_N \| \mathcal{N}(0,\sigma^2)\right). \tag{1.26}$$

Pinsker et al. [189] studied a discrete channel where the additive noise is the sum of a dominant Gaussian noise and a relatively weak non-Gaussian contaminating noise. The behavior of the capacity of continuous-time power-constrained channels with additive non-Gaussian noise is investigated in [32, 33, 194, 195, 196] and upper and lower bounds are given in terms of the divergence between the noise process and the Gaussian process with the same covariance.

Analogously, Gaussian signals are the hardest to compress under a mean-square fidelity criterion. The rate-distortion function of a non-Gaussian source is upper bounded by the rate-distortion function of the Gaussian source minus its divergence with respect to a Gaussian source with identical variance.

1.3.3.5 Differential Entropy and Divergence

Note that differential entropy can be formulated as a special case of divergence. Let X is a random vector in $\mathbb{R}^d$ with mean μ and covariance matrix $\mathbf{\Sigma}$ and X_G is a Gaussian random vector with the same mean and the same covariance matrix. Then the differential entropy of X is

$$h(X) = \frac{1}{2}\log\left((2\pi e)^d \det(\mathbf{\Sigma})\right) - D(p_X \| p_{X_G}), \tag{1.27}$$

where p_X and p_{X_G} are the pdf's of X and X_G, respectively, and $D(p_X \| p_{X_G})$ gauges the non-Gaussianness of X.

1.3.3.6 Statistical Inference

Divergence has proven to be useful in various aspects of statistical inference [142], including density estimation, parameter estimation, and hypothesis testing.

Hall [107] studied divergence in the context of kernel density estimation. Let p is the true density and $\hat{p}$ the kernel density estimate. Then the divergence $D(p\|\hat{p})$ can be used as a measure of loss for the density estimate. It is shown in [107] that an appropriate choice of the kernel will lead to the minimization of the average divergence loss:

$$D(p\|\hat{p}) = \int p(x)\log\frac{p(x)}{\hat{p}(x)}\mathrm{d}x. \tag{1.28}$$

Divergence is also used in [167] to analyze the convergence speed of convolutions to the Gaussian distribution.

Given i.i.d. samples $\{X_1,\ldots,X_n\}$ and $\{Y_1,\ldots,Y_m\}$ are generated from densities $p(\cdot)$ and $q = p(\cdot - \theta)$, respectively, Bhattacharya [31] considered the estimation of the shift parameter θ. An efficient estimate is then given by

$$\hat{\theta} \triangleq \arg\min_{\theta} D_{n,m}(X_1^n \| Y_1^m - \theta), \tag{1.29}$$

where $D_{n,m}(X_1^n \| Y_1^m - \theta)$ denotes the empirical divergence estimate based on the samples $\{X_1, \ldots, X_n\}$ and $\{Y_1 - \theta, \ldots, Y_m - \theta\}$.

Menéndez et al. [166] studies parameter estimation of statistical models for categorical data. By formulating the estimation problem as a minimization of the divergence between theoretical and empirical vectors of means, they evaluate the asymptotic properties of the corresponding estimators.

For hypothesis testing, divergence estimation was applied by Ebrahimi et al. [78] to construct a test of fit for exponentiality by comparing the non-parametric divergence estimate to the parametric estimate assuming exponential distribution. Dasu et al. [68] and Krishnamurthy et al. [140] have used divergence estimates to detect changes in internet traffic and to determine stationarity in the data stream.

1.3.3.7 Pattern Recognition

Divergence is known to be an important measure of dissimilarity for pattern recognition. In the area of image processing, divergence estimates have been applied to texture classification [77, 163, 266], shape and radiance estimation [84], and face recognition [12, 215].

Audio and speech classification is another field where divergence proves to be useful. Speech signals are usually modelled as hidden Markov processes [82]. Silva and Narayanan [221, 222] proposed an upper bound on the divergence for hidden Markov models and discussed its applications to speech recognition (see [21, 35, 128, 155, 175, 268] for more literature on this subject).

Divergence can also be used to construct kernels in support vector machine (SVM) algorithms for machine learning. Moreno et al. [174] (see also [252]) proposed an SVM [251] algorithm with the kernel defined as

$$\phi(p,q) = \beta \mathrm{e}^{-\alpha(D(p\|q)+D(q\|p))}, \tag{1.30}$$

where $D(p\|q) + D(q\|p)$ is the symmetrized version of divergence between probability distributions p and q. This algorithm produces good results for multimedia classification.

1.4 Mutual Information

1.4.1 Definition

Mutual information is another important concept in information theory. It measures the statistical dependence between two random objects. Mutual information is defined as

$$I(X;Y) = D(P_{XY}\|P_X P_Y),$$

i.e., the divergence between the joint distribution and the product of the marginal distributions. As a special case of divergence, mutual information is non-negative and is zero if and only if the two random variables are independent. For discrete random variables X and Y with joint probability mass function P_{XY} and marginal probability mass functions P_X and P_Y, the mutual information between X and Y is

$$I(X;Y) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P_{XY}(x,y)\log\frac{P_{XY}(x,y)}{P_X(x)P_Y(y)} \tag{1.31}$$

$$= H(X) + H(Y) - H(X,Y), \tag{1.32}$$

where $\mathcal{X}$ and $\mathcal{Y}$ are the alphabets of X and Y, respectively.

If X and Y are continuous random variables with joint pdf p_{XY} and marginal pdf's p_X and p_Y, respectively, $I(X;Y)$ is given by

$$I(X;Y) = D(p_{XY}\|p_X p_Y)$$

$$= \int\int p_{XY}(x,y)\log\frac{p_{XY}(x,y)}{p_X(x)p_Y(y)}\mathrm{d}x\mathrm{d}y \tag{1.33}$$

$$= h(X) + h(Y) - h(X,Y). \tag{1.34}$$

Similar to (1.27), mutual information between analog random variables with finite second moments can be expressed in terms of non-Gaussianness. Let $\mathbf{\Sigma}_X$ and $\mathbf{\Sigma}_Y$ are the covariance matrices of X

and Y, respectively and $\mathbf{\Sigma}$ be the covariance matrix of (X,Y). Suppose (X_G,Y_G) are jointly Gaussian with covariance matrix $\mathbf{\Sigma}$. Then,

$$\begin{aligned} I(X;Y) = I(X_G;Y_G) &+ D(P_{XY}\|P_{X_GY_G}) \\ &- D(P_X\|P_{X_G}) - D(P_Y\|P_{Y_G}) \end{aligned} \tag{1.35}$$

where

$$I(X_G;Y_G) = \frac{1}{2}\log\frac{\det\mathbf{\Sigma}_X \det\mathbf{\Sigma}_Y}{\det\mathbf{\Sigma}}. \tag{1.36}$$

1.4.2 Universal Estimation

Estimators of mutual information for analog sources can be obtained from divergence estimates using the definition in (1.33) or from differential entropy estimates using the relationship (1.34).

Suppose $X \in \mathbb{R}^{d_X}$ is a d_X-dimensional random vector with density p_X and $Y \in \mathbb{R}^{d_Y}$ is a d_Y-dimensional random vector with density p_Y. Let $\{(X_1,Y_1),\ldots,(X_n,Y_n)\}$ is i.i.d. samples generated from the joint density p_{XY} of (X,Y). The estimation of mutual information can be formulated as the estimation of divergence, i.e.,

$$\hat{I}(X;Y) = \hat{D}(p_{XY}\|p_X p_Y). \tag{1.37}$$

The idea is to form independent pairs of X and Y by re-pairing the X and Y samples. For example, we may shift the Y sequence by half the sequence length. Then we could assume X_i and $Y_{i+\lfloor n/2\rfloor}$ to be approximately independent (in the index, the sum is $\bmod n$). Thus in lieu of estimating mutual information given samples $\{X_i,Y_i\}$, we estimate divergence between p_{XY} and $p_X \times p_Y$ based on samples $\{(X_i,Y_i)\}$ and $\{(X_i,Y_{i+\lfloor n/2\rfloor})\}$.

Alternatively, mutual information estimates can be derived from the estimates of differential entropies via (1.34):

$$\tilde{I}(X;Y) = \hat{h}(X) + \hat{h}(Y) - \hat{h}(X,Y). \tag{1.38}$$

As long as the entropy estimator is applicable to multi-dimensional data, we automatically obtain a mutual information estimate.

1.4.3 Applications

1.4.3.1 Channel Capacity

Mutual information plays a major role in the fundamental limits of channel coding and lossy compression. Shannon [216] introduced the concept of channel capacity (maximal information rate compatible with arbitrarily low error probability) and showed that for memoryless channels it is given by

$$C = \max_{P_X} I(X;Y), \tag{1.39}$$

where the maximum is taken over all possible input distributions P_X. Maximal mutual information also plays a role in the randomness required for system simulation, and in the fundamental limits of identification via channels [255].

1.4.3.2 Lossy Compression

Shannon [216] introduced the concept of rate-distortion function (minimal information rate compatible with reproduction of the source within a given distortion) and showed that for memoryless source P_X it is given by

$$R(D) = \min_{P_{Y|X}} I(X;Y), \tag{1.40}$$

where the minimum is taken over all possible conditional distributions that guarantee the required distortion level D.

1.4.3.3 Secrecy

Mutual information also plays a role in secure communications. Let $X^n = \{X_1,\ldots,X_n\}$ and $Y^n = \{Y_1,\ldots,Y_n\}$ are n i.i.d. realizations of correlated random variables X and Y. Alice and Bob observe the sequences X^n and Y^n, respectively. Assume that they can communicate with each other over an error-free public channel. Let V^n denote all the transmissions on the public channel. After the transmission, Alice generates a k-bit string S_A^n, based on (X^n,V^n), and Bob generates a k-bit string S_B^n, based on (Y^n,V^n). A bit string S^n is called a secret key if there exist S_A^n and S_B^n, such that

(1) $\lim_{n\to\infty} Pr(S^n = S_A^n = S_B^n) = 1$;
(2) $\lim_{n\to\infty} \frac{1}{n} I(S^n;V^n) = 0$;
(3) $H(S^n) = k$.

The largest secret key rate is [165, 4]

$$C_s = I(X;Y), \tag{1.41}$$

namely the mutual information rate between the observations available to Alice and Bob, respectively. Consequently, estimates of mutual information can be used to evaluate the efficiency of secrecy generation algorithms [264, 269].

1.4.3.4 Independence Test

Minimization of mutual information is widely used in independence tests. Robinson [202] examined mutual information in the context of testing for memory in random processes. Let $X_n, n = 1,2,\ldots$ is a stationary process. Assume that X_1 is a continuous random variable with pdf $h(x)$ and X_1 and X_2 have joint pdf $f(x,y)$. Under such assumptions, the null hypothesis

$$H_0 : f(x,y) = h(x)h(y) \tag{1.42}$$

is equivalent to memorylessness of the process. In [202], a hypothesis test is constructed using consistent estimates of mutual information as test statistics. Applications to testing the random walk hypothesis for exchange rate series and some other hypotheses of econometric interest are described as well. See [37, 75, 86, 88, 97, 133, 197, 225, 226, 237] for a sampling of the literature on this subject.

Mutual information is also used to identify independent components. Comon [50] studied independent component analysis (ICA) of a random vector. The concept of ICA may be seen as an extension of principal component analysis, which only imposes uncorrelatedness. The idea of ICA is to utilize mutual information as a measure of dependence and search for a linear transformation that minimizes the mutual information between the components of the vector. Further works on this topic are presented in [118, 127, 231].

1.4.3.5 Multimedia Processing

Mutual information has been used as a similarity measure for image registration because of its generality and high accuracy. Given a reference image modelled as a random vector U (e.g., a brain scan), a second image V needs to be put into the same coordinate system as the reference image. The estimated alignment is given by the transformation T^* on the image V that maximizes the mutual information between the image U and the transformed version of image V, namely:

$$T^* = \arg\max_T I(U;T(V)). \tag{1.43}$$

Image registration based on mutual information has been investigated for medical imaging in [38, 160, 164, 241, 248] with focus on different aspects of the registration process. A review of methodologies and specific applications is presented by Pluim et al. in [192]. More recent work [16, 131, 234, 240] employed mutual information in fMRI data analysis. For example, Tsai et al. [240] computed the brain activation map by quantifying the relationship between the fMRI temporal response of a voxel and the experimental protocol timeline using mutual information. Similar registration criteria are explored in [46] for remote sensing images. The estimation of mutual information between images is discussed in [18, 146, 153]. An upper bound is derived in [227] for the mutual information between a fixed image and a deformable template containing a fixed number of gray levels.

1.4.3.6 Computational Biology and Neuroscience

Adami [1] considers applications of mutual information in the study of the genetic code:

- Investigation of the information content of DNA binding site.
- Prediction of protein structure.
- Detection of protein–protein and DNA-protein interactions.
- Drug design by maximizing the mutual information between the protease and inhibitor library.

Aktulga et al. [7, 8] (see also Schneider [212]) demonstrated the use of mutual information in identifying statistically correlated segments of DNA or RNA.

Furthermore, since mutual information provides a general measure of dependence, there has been an increasing popularity in computational biology of using mutual information to cluster co-expressed genes [41, 168]. See [230] for a tutorial on this topic.

Information-theoretic methods have also been used in neuroscience to study the dependence between stimuli and neural response [36, 100, 176, 184, 232] and to classify neurons according to their functions [126, 213].

1.4.3.7 Machine Learning

Machine learning is concerned with the design and development of algorithms and techniques that allow automatic extraction of rules and patterns from massive data sets. The connection between information theory and machine learning has received much attention. For example, Kraskov et al. [136] designed a distance measure based on mutual information and applied this measure to hierarchical clustering. The hierarchical clustering consists of organizing data as a hierarchy of nested partitions by linking the two closest clusters, where the distance between the discrete random variables X and Y is defined as

$$D(X,Y) = 1 - \frac{I(X;Y)}{H(X,Y)}, \tag{1.44}$$

where $H(X,Y)$ is the joint entropy of X and Y.

Mutual information is also incorporated in boosting algorithms [15, 152, 156] to improve classification performance. An information theoretic interpretation of boosting is proposed by Kivinen et al. [132].

Another important application of information measures is in feature extraction [145, 152, 239], which is an important step in pattern recognition tasks often dictated by practical feasibility. In [145], a method is proposed for learning discriminative feature transforms using a criterion based on the mutual information between class labels and transformed features. Experiments show that this method is effective in reducing the dimensionality and leads to better classification results.

1.5 Rényi Entropy and Rényi Differential Entropy

Rényi entropy [200] is a generalization of Shannon entropy (1.1). Let X is a discrete random object with probability mass function $P_X(\cdot)$. The Rényi entropy of X of order α, where $\alpha \geq 0$ and $\alpha \neq 1$, is defined as

$$H_\alpha(X) = \frac{1}{1-\alpha} \log \left(\sum_{x \in \mathcal{X}} P_X^\alpha(x) \right). \tag{1.45}$$

If we take the limit as $\alpha \to 1$, we obtain the entropy:

$$H(X) = \lim_{\alpha \to 1} H_\alpha(X). \tag{1.46}$$

In the limit as α approaches 0, H_α converges to the cardinality of the alphabet of X:

$$H_0(X) = \log |\mathcal{X}|, \tag{1.47}$$

which is also known as the Hartley entropy. It is also interesting to note that for $\alpha = 2$,

$$H_2(X) = -\log \left(\sum_{x \in \mathcal{X}} P_X^2(x) \right) = -\log P[X = Y], \tag{1.48}$$

where Y is a random variable independent of X but distributed identically to X. Relations between Shannon and Rényi entropies of integer orders are discussed in [272].

If X is equiprobable, $H_\alpha(X) = \log |\mathcal{X}|$. Otherwise the Rényi entropies are monotonically decreasing as a function of α.

Rényi entropy also satisfies several important properties of Shannon entropy including:

- *Continuity*: $H_\alpha(X)$ is a continuous function of the probabilities $P_X(x), x \in \mathcal{X}$;
- *Symmetry*: $H_\alpha(X)$ is a symmetric function of $P_X(x), x \in \mathcal{X}$. Namely $H_\alpha(X)$ remains unchanged if the the probabilities are reassigned to the outcomes $x \in \mathcal{X}$;
- *Additivity*: If Y is independent of X, we have

$$H_\alpha(X,Y) = H_\alpha(X) + H_\alpha(Y). \tag{1.49}$$

For analog sources, Rényi differential entropy generalizes the notion of differential entropy. For a continuous random variable X with probability density function p_X, the Rényi differential entropy h_α of order $\alpha \geq 0$, $\alpha \neq 1$, is defined as

$$h_\alpha(X) = \frac{1}{1-\alpha} \log \int_{\mathbb{R}^d} p_X^\alpha(x) \mathrm{d}x \tag{1.50}$$

Note that the differential entropy can be expressed as the limit of Rényi differential entropy

$$h(X) = \lim_{\alpha \to 1} h_\alpha(X). \tag{1.51}$$

As $\alpha \to 0$, the zeroth-order Rényi entropy gives the logarithm of the measure of the support set of the density p_X:

$$h_0(X) = \log\left(\lambda \left\{ x \in \mathbb{R}^{d_X} : p_X(x) > 0 \right\}\right). \tag{1.52}$$

For comparison, recall that differential entropy gives the logarithm of the effective volume of the typical sequences (Theorem 1.1).

Rényi differential entropy plays a fundamental role in several information theory problems. For example, in vector quantization, Rényi differential entropy characterizes the behavior of the rate-distortion function in the fine quantization regime [9, 91, 178, 185]. For simplicity, consider a one-dimensional quantization problem where X is a continuous random variable with pdf p_X and N is the number of levels. Algazi [9] used the rth power distortion measure and showed that for sufficiently large N the minimum distortion is given by

$$D_r(N) \approx \frac{1}{r+1} 2^{-r} \exp\left\{-r\left(\log N - h_{1/(1+r)}(X)\right)\right\}, \tag{1.53}$$

where $h_{1/(1+r)}(X)$ is the Rényi differential entropy of X of order $1/(1+r)$.

Rényi differential entropy is also useful for clustering and data classification. In [3, 94, 123], an information theoretic criterion is developed based on Rényi differential entropy to optimize the clustering results. In image registration, Rényi differential entropy is employed as a similarity metric [209, 210, 211].

1.6 f-Divergence

The f-divergence is a family of distance measures introduced by Csiszár [59, 62, 63] and independently by Ali and Silvey [10]. Its many properties are discussed in [183, 246, 247, 245]. Suppose P and Q are probability distributions defined on the same measurable space $(\Omega, \mathcal{F})$ and Q is absolutely continuous with respect to P with $\mathrm{d}Q/\mathrm{d}P$ being the Radon-Nikodym derivative. Let $f : [0, +\infty) \to \mathbb{R}$ is a convex function. The f-divergence between P and Q is defined as

$$D_f(P\|Q) = \int_\Omega f\left(\frac{\mathrm{d}Q}{\mathrm{d}P}\right)\mathrm{d}P. \tag{1.54}$$

For discrete distributions on an alphabet $\mathcal{A}$, f-divergence becomes

$$D_f(P\|Q) = \sum_{a\in\mathcal{A}} P(a) f\left(\frac{Q(a)}{P(a)}\right). \tag{1.55}$$

For continuous distributions with probability density functions p and q,

$$D_f(p\|q) = \int_{\mathbb{R}^d} p(x) f\left(\frac{q(x)}{p(x)}\right)\mathrm{d}x. \tag{1.56}$$

Various measures of distance between probability distributions are special cases of f-divergence (see [20, 183] for a longer list)

- (*Kullback–Leibler*) *divergence*:

$$D(P\|Q) = \int \mathrm{d}P \log\frac{\mathrm{d}P}{\mathrm{d}Q} = D_f(P\|Q), \tag{1.57}$$

  with $f(u) = -\log u$;
- *Variational distance*:

$$V(p,q) = \int |p(x) - q(x)|\mathrm{d}x = D_f(p\|q), \tag{1.58}$$

  with $f(u) = 1/2|1-u|$;
- *Hellinger distance*:

$$H(p,q) = \int \left|\sqrt{p(x)} - \sqrt{q(x)}\right|\mathrm{d}x = D_f(p\|q), \tag{1.59}$$

  with $f(u) = (\sqrt{x}-1)^2$;

- *Bhattacharyya distance*:

$$B(p,q) = \int \sqrt{p(x)q(x)}\mathrm{d}x = -D_f(p\|q), \tag{1.60}$$

  where $f(u) = -\sqrt{u}$.
- *Rényi divergence of order* α:

$$D_\alpha(P\|Q) = \frac{1}{\alpha - 1}\int p^\alpha(x)q^{1-\alpha}(x)\mathrm{d}x = \log D_f(p\|q), \tag{1.61}$$

  where $f(u) = \frac{1}{\alpha-1}u^{1-\alpha}$.

f-divergence is applicable in a number of problems. For instance, f-divergence parameterizes the Chernoff exponent governing the minimum probability of error in binary hypothesis testing [55]. Consider two hypotheses p and q for the underlying probability density function. Let the prior probabilities are α and $1 - \alpha$. The error probability of the optimal Bayes rule is:

$$\begin{aligned} P_e &= \int \min\{\alpha p(x), (1-\alpha)q(x)\}\,\mathrm{d}x \\ &= D_f(p\|q) + 1, \end{aligned} \tag{1.62}$$

with

$$f(u) = -\min\{u, 1-u\}. \tag{1.63}$$

f-divergence is useful in pattern recognition applications to identify independent components [17]. A correspondence between surrogate loss functions for classification and f-divergence has been shown in [179] f-divergence is also employed as a dissimilarity measure for image registration in [191] and [109, 157], and in the design of quantizers [193].

References

[1] C. Adami, "Information theory in molecular biology," *Physics of Life Reviews*, vol. 1, pp. 3–22, Online. Available: http://arxiv.org/abs/q-bio/0405004, 2004.

[2] T. M. Adams and A. B. Nobel, "On density estimation from ergodic processes," *The Annals of Probability*, vol. 26, no. 2, pp. 794–804, April 1998.

[3] M. Aghagolzadeh, H. Soltanian-Zadeh, and B. N. Araabi, "New information-based clustering method using Rényi's entropy and fuzzy C-means clustering," *Proceeding of Signal and Image Processing*, pp. 411–414, 2005.

[4] R. Ahlswede and I. Csiszár, "Common randomness in information theory and cryptography — Part I: Secrete sharing," *IEEE Transactions on Information Theory*, vol. 39, no. 4, pp. 1121–1132, July 1993.

[5] I. Ahmad and P. Lin, "A nonparametric estimation of the entropy for absolutely continuous distributions," *IEEE Transactions on Information Theory*, vol. 22, no. 3, pp. 372–375, May 1976.

[6] N. A. Ahmed and D. V. Gokhale, "Entropy expressions and their estimators for multivariate distributions," *IEEE Transactions on Information Theory*, vol. 35, no. 3, pp. 688–692, May 1989.

[7] H. M. Aktulga, I. Kontoyiannis, L. A. Lyznik, L. Szpankowski, A. Y. Grama, and W. Szpankowski, "Identifying statistical dependence in genomic sequences via mutual information estimates," *EURASIP Journal on Bioinformatics and Systems Biology*, no. 3, id. 14741, July 2007.

[8] H. M. Aktulga, I. Kontoyiannis, L. A. Lyznik, L. Szpankowski, A. Y. Grama, and W. Szpankowski, "Statistical dependence in biological sequences," *Proceedings of 2007 International Symposium on Information Theory* (*ISIT2007*), pp. 2676–2680, June 2007.

[9] V. R. Algazi, "Useful approximations to optimum quantization," *IEEE Transactions on Communications Technology*, vol. COM-14, pp. 297–301, 1966.

[10] S. M. Ali and S. D. Silvey, "A general class of coefficients of divergence of one distribution from another," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 28, no. 1, pp. 131–142, 1966.

[11] B. D. O. Anderson, J. B. Moore, and R. M. Hawkes, "Model approximation via prediction error identification," *Automatica*, vol. 14, pp. 615–622, November 1978.

[12] O. Arandjelovic, G. Shakhnarovich, J. Fisher, R. Cipolla, and T. Darrell, "Face recognition with image sets using manifold density divergence," *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, pp. 581–588, 2005.

[13] S. Artstein, K. Ball, F. Barthe, and A. Naor, "Solution of Shannon's problem on the monotonicity of entropy," *Journal of the American Mathematical Society*, vol. 17, pp. 975–982, 2004.

[14] S. Arya, "Nearest neighbor searching and applications," Ph.D. Thesis, no. CS-TR-3490, University of Maryland, June, 1995.

[15] J. A. Aslam, "Improving algorithms for boosting," *Proceedings of the 13th Annual Conference on Computational Learning Theory*, pp. 200–207, 2000.

[16] S. P. Awate, T. Tasdizen, R. T. Whitaker, and N. L. Foster, "Adaptive nonparametric Markov modeling for unsupervised MRI brain-tissue classification," *Medical Image Analysis*, vol. 10, no. 5, pp. 726–739, 2006.

[17] Y. Bao and H. Krim, "Rényi entropy based divergence measures for ICA," *Proceedings of 2003 IEEE Workshop on Statistical Signal Processing*, no. 28, pp. 565–568, 2003.

[18] A. Bardera, M. Feixas, I. Boada, and M. Sbert, "Compression-based image registration," *Proceedings of the 2006 IEEE International Symposium on Information Theory (ISIT2006)*, pp. 436–440, July 2006.

[19] A. R. Barron, "Entropy and the central limit theorem," *Annals of Probability*, vol. 14, pp. 336–342, 1986.

[20] M. Basseville, "Distance measures for signal processing and pattern recognition," *Signal Processing*, vol. 18, no. 4, pp. 349–369, December 1989.

[21] H. Beigi, S. Maes, and J. Sorensen, "A distance measure between collections of distributions and its application to speaker recognition," *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP'98)*, vol. 2, pp. 753–756, 1998.

[22] J. Beirlant, "Limit theory for spacing statistics from general univariate distributions," *Publications de l'Institut de Statistique de l'Université de Paris XXXI fasc*, vol. 1, pp. 27–57, 1986.

[23] J. Beirlant, E. J. Dudewicz, L. Györfi, and E. van der Meulen, "Nonparametric entropy estimation: An overview," *The International Journal of Mathematics and Statistical Sciences*, vol. 6, no. 1, pp. 17–39, June 1997.

[24] J. Beirlant and M. C. A. Zuijlen, "The empirical distribution function and strong laws for functions of order statistics of uniform spacings," *Journal of Multivariate Analysis*, vol. 16, pp. 300–317, 1985.

[25] J. D. Bekenstein, "Information in the holographic universe," *Scientific American*, pp. 59–65, August 2003.

[26] D. Benedetto, E. Caglioti, and V. Loreto, "Language trees and zipping," *Physical Review Letters*, vol. 88, no. 4, id. 048702, January 2002.

[27] C. H. Bennett and P. Shor, "Quantum information theory," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2724–2742, October 1998.

[28] J. L. Bentley, "Multidimensional binary search trees used for associative searching," *Communications of the ACM*, vol. 18, no. 9, pp. 509–517, 1975.

[29] T. Berger, "Optimum quantizers and permutation codes," *IEEE Transactions on Information Theory*, vol. 18, pp. 759–765, November 1972.

[30] T. Berger, "Minimum entropy quantizers and permutation codes," *IEEE Transactions on Information Theory*, vol. 28, pp. 149–157, March 1982.

[31] P. K. Bhattacharya, "Efficient estimation of a shift parameter from grouped data," *The Annals of Mathematical Statistics*, vol. 38, no. 6, pp. 1770–1787, December 1967.

[32] J. Binia, "On the capacity of certain additive non-Gaussian channels," *IEEE Transactions on Information Theory*, vol. 25, pp. 448–452, July 1979.

[33] J. Binia, "New bounds on the capacity of certain infinite-dimensional additive non-Gaussian channels," *IEEE Transactions on Information Theory*, vol. 51, pp. 1218–1221, March 2005.

[34] J. Binia, "On divergence-power inequalities," *IEEE Transactions on Information Theory*, vol. 53, no. 3, pp. 1179–1182, March 2007.

[35] I. Bocharov and P. Luki, "Application of Kullback-Leibler metric to speech recognition," Nizhny Novgorod Linguistic University, Russia, Technical Report, 2003.

[36] A. Borst and F. E. Theunissen, "Information theory and neural coding," *Nature Neuroscience*, vol. 2, pp. 947–957, 1999.

[37] D. R. Brillinger and A. Guha, "Mutual information in the frequency domain," *Journal of Statistical Planning and Inference*, vol. 137, no. 3, pp. 1076–1084, March 2007.

[38] L. G. Brown, "A survey of image registration techniques," *ACM Computing Surveys*, vol. 24, pp. 325–376, December 1992.

[39] P. W. Buchen and M. Kelly, "The maximum entropy eistribution of an asset inferred from option prices," *The Journal of Financial and Quantitative Analysis*, vol. 31, pp. 143–159, March 1996.

[40] J. P. Burg, "Maximum entropy spectral analysis," Ph.D. thesis, Department of Geophysics, Stanford University, Stanford, CA, 1975.

[41] A. Butte and I. S. Kohane, "Mutual information relevance networks: Functional genomic clustering using pairwise entropy measurements," *Proceedings of the Pacific Symposium on Biocomputing*, vol. 5, pp. 415–426, 2000.

[42] T. Cacoullus, "Estimation of a multivariate density," *The Annals of the Institute of Statistical Mathematics*, vol. 18, p. 179, 1966.

[43] H. Cai, S. R. Kulkarni, and S. Verdú, "Universal entropy estimation via block sorting," *IEEE Transactions on Information Theory*, vol. 50, no. 7, pp. 1551–1561, July 2004.

[44] H. Cai, S. R. Kulkarni, and S. Verdú, "Universal divergence estimation for finite-alphabet sources," *IEEE Transactions on Information Theory*, vol. 52, no. 8, pp. 3456–3475, August 2006.

[45] A. Carbonez, L. Györfi, and E. C. van der Meulen, "Nonparametric entropy estimation based on randomly censored data," *Problems of Control and Inforrmation Theory*, vol. 20, pp. 441–451, 1991.

[46] H. Chen, P. K. Varshney, and M. K. Arora, "Performance of mutual information similarity measure for registration of multi-temporal remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 41, no. 11, part. 1, pp. 2445–2454, November 2003.

[47] B. C. Cheng and P. M. Robinson, "Density estimation in strongly dependent non-lineartime series," *Statistica Sinica*, vol. 1, pp. 335–359, 1991.

[48] B. S. Choi and T. M. Cover, "An information-theoretic proof of Burg's maximum entropy spectrum," *Proceedings of IEEE*, vol. 72, pp. 1094–1095, 1984.

[49] R. L. Cilibrasi and P. M. Vitányi, "Clustering by compression," *IEEE Transactions on Information Theory*, vol. 51, no. 4, pp. 1523–1545, April 2005.

[50] P. Comon, "Independent component analysis, a new concept?," *Signal Processing*, vol. 36, pp. 287–314, 1994.

[51] T. Cover and R. King, "A convergent gambling estimate of the entropy of English," *IEEE Transactions on Information Theory*, vol. 24, no. 4, pp. 413–421, July 1978.

[52] T. M. Cover, "Universal portfolios," *Mathematical Finance*, vol. 1, no. 1, pp. 1–29, January 1991.

[53] T. M. Cover, "Shannon and investment," *IEEE Information Theory Newsletter* (*Special Golden Jubilee Issue*), pp. 10–11, June 1998.

[54] T. M. Cover and P. E. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, January 1967.

[55] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Wiley Series in Telecommunications, 1991.

[56] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Wiley Series in Telecommunications, Second ed., 2005.

[57] H. Cramér, "On some classes of series used in mathematical statistics," *Proceedings of the 6th Scandinavian Congress of Mathematicians*, pp. 399–425, 1925.

[58] N. Cressie, "On the logarithms of higher order spacings," *Biometrika*, vol. 63, pp. 343–355, 1976.

[59] I. Csiszár, "Information-type measures of difference of probability distributions and indirect observations," *Studia Scientiarum Mathematicarum Hungarica*, vol. 2, pp. 229–318, 1967.

[60] I. Csiszár, "On generalized entropy," *Studia Scientiarum Mathematicarum Hungarica*, vol. 4, pp. 401–419, 1969.

[61] I. Csiszár, "Generalized entropy and quantization problems," *Transactions of the 6th Praque Conference on Information Theory, Statistical Decision Functions, Random Processes* (*Academia, Prague*), 1973.

[62] I. Csiszár, "I-divergence geometry of probability distributions and minimization problems," *The Annals of Probability*, vol. 3, no. 1, pp. 146–158, February 1975.

[63] I. Csiszár and J. Korner, *Information Thoery: Coding Theorems for Discrete Memoryless Systems*. New York: Academic Press, 1982.

[64] I. Csiszár and P. C. Shields, "Information theory and statistics: A tutorial," *Foundations and Trends in Communications and Information Theory*, vol. 1, no. 4, pp. 417–528, December 2004.

[65] G. A. Darbellay and I. Vajda, "Estimation of the information by an adaptive partitioning of the observation space," *IEEE Transactions on Information Theory*, vol. 45, no. 4, pp. 1315–1321, May 1999.

[66] G. A. Darbellay and I. Vajda, "Entropy expressions for multivariate continuous distributions," *IEEE Transactions on Information Theory*, vol. 46, no. 2, pp. 709–712, March 2000.

[67] B. V. Dasarathy, *Nearest Neighbor Pattern Classification Techniques.* IEEE Computer Society Press, 1991.

[68] T. Dasu, S. Krishnan, S. Venkatasubramanian, and K. Yi, "An information-theoretic approach to detecting changes in multi-dimensional data streams," *Proceedings of the 38th Symposium on the Interface of Statistics, Computing Science, and Applications* (*Interface '06*), Pasadena, CA, May 2006.

[69] A. Dembo, T. M. Cover, and J. A. Thomas, "Information theoretic inequalities," *IEEE Transactions on Information Theory*, vol. 37, no. 6, pp. 1501–1518, November 1991.

[70] L. Devroye, "A course in density estimation," in *Progress in Probability and Statistics*, vol. 14, Birkhäuser, 1987.

[71] L. Devroye and L. Györfi, *Nonparametric Density Estimation, the L_1 View.* Wiley Series in Probability and Mathematical Statistics, 1985.

[72] L. Devroye, L. Györfi, A. Krzyzak, and G. Lugosi, "On the strong universal consistency of nearest neighbor regression function estimates," *The Annals of Statistics*, vol. 22, no. 3, pp. 1371–1385, September 1994.

[73] L. Devroye, L. Györfi, and G. Lugosi, *A Probablistic Theory of Pattern Recognition.* Springer, 1996.

[74] L. P. Devroye and T. J. Wagner, "The strong uniform consistency of nearest neighbor density estimates," *The Annals of Statistics*, vol. 5, no. 3, pp. 536–540, May 1977.

[75] A. Dionísio, R. Menezes, and D. A. Mendes, "Entropy-based independence test," *Nonlinear Dynamics*, vol. 44, no. 1–4, pp. 351–357, June 2006.

[76] Y. G. Dmitriev and F. P. Tarasenko, "On the estimation of functionals of the probability density and its derivatives," *Thoery of Probability and Its Applications*, vol. 18, no. 3, pp. 628–633, 1974.

[77] M. N. Do and M. Vetterli, "Wavelet-based texture retrieval using generalized Gaussian density and Kullback-Leibler distance," *IEEE Transactions on Image Processing*, vol. 11, no. 2, pp. 146–158, February 2002.

[78] N. Ebrahimi, M. Habibullah, and E. S. Soofi, "Testing exponentiality based on Kullback-Leibler information," *Journal of the Royal Statistical Society. Series B* (*Methodological*), vol. 54, no. 3, pp. 739–748, 1992.

[79] K. Eckschlager, *Information Theory in Analytical Chemistry.* New York: Wiley, 1994.

[80] F. Y. Edgeworth, "The generalised law of error, or law of great numbers," *Journal of the Royal Statistical Society*, vol. 69, no. 3, pp. 497–539, September 1906.

[81] I. Eidhammer, I. Jonassen, and W. R. Taylor, *Protein Bioinformatics: An Algorithmic Approach to Sequence and Structure*. John Wiley and Sons, 2004.

[82] Y. Ephraim and N. Merhav, "Hidden Markov process," *IEEE Transactions on Information Theory*, vol. 48, pp. 1518–1569, June 2002.

[83] W. J. Ewens and G. R. Grant, *Statistical Methods in Bioinformatics: An Introduction*. Springer, 2005.

[84] P. Favaro and S. Soatto, "Shape and radiance estimation from the information divergence of blurred images," *Proceedings of the 6th European Conference on Computer Vision — Part I*, pp. 755–768, 2000.

[85] A. A. Fedotov, P. Harremoes, and F. Topsoe, "Refinements of Pinsker's inequality," *IEEE Transactions on Information Theory*, vol. 49, no. 6, pp. 1491–1498, June 2003.

[86] M. Fernandes, "Nonparametric entropy-based tests of independence between stochastic processes," Ph.D. dissertation, the Solvay Business School, Université Libre de Bruxelles, 2000.

[87] E. Fix and J. L. Hodges, "Discriminatory analysis, nonparametric discrimination: Consistency properties," USAF School of Aviation Medicine, Randolph Field, TX, USA, Technical Report 4, Project Number 21-49-004, 1951.

[88] A. M. Fraser and H. L. Swinney, "Independent coordinates or strange attractors from mutual information," *Physical Review A*, vol. 33, no. 2, pp. 1134–1140, February 1986.

[89] K. Fukunaga and L. D. Hostetler, "Optimization of k-nearest-neighbor density estimates," *IEEE Transactions on Information Theory*, vol. 20, no. 5, pp. 320–326, May 1973.

[90] K. Fukunaga and P. M. Nerada, "A branch and bound algorithm for computing k-nearest neighbors," *IEEE Transactions on Computers*, vol. 24, pp. 750–753, July 1975.

[91] A. Gersho, "Asymptotically optimal block quantization," *IEEE Transactions on Information Theory*, vol. 25, pp. 373–380, July 1979.

[92] H. Gish and J. N. Pierce, "Asymptotically efficient quantizing," *IEEE Transactions on Information Theory*, vol. 14, pp. 676–683, September 1968.

[93] M. Godavarti and A. Hero, "Convergence of differential entropies," *IEEE Transactions on Information Theory*, vol. 50, no. 2, pp. 171–176, January 2004.

[94] E. Gokcay and J. C. Principe, "Information theoretic clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 2, pp. 158–171, February 2002.

[95] M. H. Goldwasser, D. S. Johnson, and C. C. McGeoch, *Data Structures, Near Neighbor Searches, and Methodology: 5th and 6th DIMACS Implementation Challenge*. American Mathematical Society, 2002.

[96] M. N. Goria, N. N. Leonenko, V. V. Mergel, and P. L. Novi Inverardi, "A new class of random vector entropy estimators and its applications in testing statistical hypotheses," *Journal of Nonparametric Statistics*, vol. 17, no. 3, pp. 277–297, April 2005.

[97] C. Granger and J. L. Lin, "Using the mutual information coefficient to identify lags in nonlinear models," *The Journal of Time Series Analysis*, vol. 15, no. 4, pp. 371–384, 1994.

[98] R. M. Gray and D. L. Neuhoff, "Quantization," *IEEE Transactions on Information Theory*, vol. 44, pp. 1–63, October 1998.

[99] D. Guo, S. Shamai, and S. Verdú, "Proof of entropy power inequalities via MMSE," in *Proceedings of 2006 IEEE International Symposium on Information Theory*, pp. 1011–1015, Seattle, WA, July 2006.

[100] P. Gurzi, G. Biella, and A. Spalvier, "Estimate of mutual information carried by neuronal responses from small data samples," *Network: Computation in Neural Systems*, vol. 7, pp. 717–725, 1996.

[101] L. Györfi, W. Härdle, P. Sarda, and P. Vieu, *Nonparametric Curve Estimation from Time Series*. Berlin: Springer-Verlag, 1989.

[102] L. Györfi, T. Linder, and E. C. van der Meulen, "On the asymptotic optimality of quantizers," in *Proceedings of the 11th Symposium on Information Theory in the Benelux, Noordwijkerhout*, pp. 29–35, The Netherlands, October 1990.

[103] L. Györfi and G. Lugosi, "Kernel density estimation from ergodic sample is not universally consistent," *Computational Statistics and Data Analysis*, vol. 14, no. 4, pp. 437–442, November 1992.

[104] L. Györfi, G. Morvai, and S. Yakowitz, "Limits to consistent on-line forecasting for ergodic time series," *IEEE Transactions on Information Theory*, vol. 44, no. 2, pp. 886–892, June 1998.

[105] L. Györfi and E. C. van der Meulen, "Density-free convergence properties of various estimators of entropy," *Computational Statistics and Data Analysis*, vol. 5, no. 4, pp. 425–436, September 1987.

[106] P. Hall, "Limit theorems for sums of general functions of m-spacings," *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 96, pp. 517–532, 1984.

[107] P. Hall, "On Kullback-Leibler loss and density estimation," *The Annals of Statistics*, vol. 15, pp. 1491–1519, December 1987.

[108] P. Hall and S. Morton, "On the estimation of the entropy," *The Annals of the Institute of Statistical Mathematics*, vol. 45, pp. 69–88, 1993.

[109] A. O. Hero, J. Gorman, and O. J. J. Michel, "Image registration with minimum spanning tree algorithm," *Proceedings of the 2000 International Conference on Image Processing*, vol. 1, pp. 481–484, Vancouver, Canada, September 2000.

[110] A. O. Hero, B. Ma, O. J. J. Michel, and J. Gorman, "Applications of entropic spanning graphs," *Signal Processing Magazine*, vol. 19, no. 5, pp. 85–95, September 2003.

[111] A. O. Hero and O. J. J. Michel, "Robust entropy estimation strategies based on edge weighted random graphs," *Proceedings of Meeting of the International Society for Optical Engineering* (*SPIE*), pp. 250–261, San Diego, CA, July 1998.

[112] A. O. Hero and O. J. J. Michel, "Estimation of Rényi information divergence via pruned minimal spanning trees," *Proceedings of 1999 IEEE Workshop on Higher Order Statistics*, Caesaria, Israel, June 1999.

[113] A. O. Hero and O. J. J. Michel, "Asymptotic theory of greedy approximations to minimal k-point random graphs," *IEEE Transactions on Information Theory*, vol. 45, no. 6, pp. 1921–1938, September 1999.

[114] J. Hidalgo, "Non-parametric estimation with strongly dependent time multivariate time series," *Journal of Time Series Analysis*, vol. 18, no. 2, pp. 95–122, March 1997.

[115] A. Hinneburg, C. C. Aggarwal, and D. A. Keim, "What is the nearest neighbor in high dimensional spaces?," in *Proceedings of the 26th VLDB Conference*, pp. 506–515, Cairo, Egypt, 2000.

[116] H. C. Ho, "On the strong uniform consistency of density estimation for strongly dependent sequences," *Statistics and Probability Letters*, vol. 22, no. 2, pp. 149–156, February 1995.

[117] A. Hyvärinen, "New approximations of differential entropy for independent component analysis and projection pursuit," in *Advances in Neural Information Processing System* (*NIPS'97*), vol. 10, pp. 273–279, MIT Press, 1998.

[118] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. John Wiley & Sons, 2001.

[119] S. Ihara, "On the capacity of channels with additive non-Gaussian noise," *Information and Control*, vol. 37, pp. 34–39, April 1978.

[120] E. T. Jaynes, "Information theory and statistical mechanics," *Physical Review*, vol. 106, pp. 620–630, May 1957.

[121] E. T. Jaynes, "Information theory and statistical mechanics ii," *Physical Review*, vol. 108, pp. 171–190, October 1957.

[122] E. T. Jaynes, "Information theory and statistical mechanics," in *Statistical Physics*, pp. 181–218, Brandeis Summer Institute 1962, New York, NY: W. A. Benjamin, Inc., 1963.

[123] R. Jenssen, K. E. Hild, D. Erdogmus II, J. C. Principe, and T. Eltoft, "Clustering using Rényi's entropy," *Proceeding of the 2003 International Joint Conference on Neural Networks*, vol. 1, pp. 523–528, July 2003.

[124] H. Joe, "On the estimation of entropy and other functionals of a multivariate density," *The Annals of the Institute of Statistical Mathematics*, vol. 41, no. 4, pp. 683–697, 1989.

[125] H. Joe, "Relative entropy measures of multivariate dependence," *Journal of the American Statistical Association*, vol. 84, no. 405, pp. 171–176, March 1989.

[126] D. H. Johnson, C. M. Gruner, K. Baggerly, and C. Seshagiri, "Information-theoretic analysis of neural coding," *Journal of Computational Neuroscience*, vol. 10, no. 1, pp. 47–69, January 2001.

[127] M. C. Jones and R. Sibson, "What is projection pursuit?," *Journal of the Royal Statistical Society. Series A* (*General*), vol. 150, pp. 1–37, 1987.

[128] B. Juang and L. R. Rabiner, "A probabilistic distance measure for hidden Markov models," *AT&T Technical Journal*, vol. 64, pp. 391–408, February 1985.

[129] A. Kaltchenko, "Algorithms for estimating information distance with application to bioinformatics and linguistics," *Proceedings of Canadian Conference on Electrical and Computer Engineering*, vol. 4, pp. 2255–2258, May 2004.

[130] D. Kazakos and P. Papantoni-Kazakos, "Spectral distance measures between Gaussian processes," *IEEE Transactions on Automatic control*, vol. AC-25, no. 5, pp. 950–959, 1980.

[131] J. Kim, J. W. Fisher, A. Tsai, C. Wible, A. S. Willsky, and W. M. Wells, "Incorporating spatial priors into an information theoretic approach for fMRI data analysis," *Proceedings of the 3rd International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 62–71, October 2000.

[132] J. Kivinen and M. K. Warmuth, "Boosting as entropy projection," *Proceedings of the 12th Annual Conference on Computational Learning Theory*, pp. 133–144, 1999.

[133] I. Kojadinovic, "On the use of mutual information in data analysis: an overview," *Proceedings of 11th International Symposium on Applied Stochastic Models and Data Analysis* (*ASMDA'05*), pp. 738–747, May 2005.

[134] I. Kontoyiannis, P. H. Algoet, Y. M. Suhov, and A. J. Wyner, "Nonparametric entropy estimation for stationary processes and random fields, with applications to English text," *IEEE Transactions on Information Theory*, vol. 44, no. 3, pp. 1319–1327, May 1998.

[135] L. F. Kozachenko and N. N. Leonenko, "Sample estimate of the entropy of a random vector," *Problems of Information Transmission*, vol. 23, pp. 95–101, 1987.

[136] A. Kraskov, H. Stögbauer, R. G. Andrzejak, and P. Grassberger, "Hierarchical clustering using mutual information," *Europhysics Letters*, vol. 70, no. 2, pp. 278–284, April 2005.

[137] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Physical Review E*, vol. 69, id. 066138, 2004.

[138] C. J. Krebs, *Ecological Methodology.* Addison Wesley Longman, 1999.

[139] R. E. Krichevsky and V. K. Trofimov, "The performance of universal encoding," *IEEE Transactions on Information Theory*, vol. 27, no. 2, pp. 199–207, March 1981.

[140] B. Krishnamurthy, H. V. Madhyastha, and S. Venkatasubramanian, "On stationarity in internet measurements through an information-theoretic lens," *Proceedings of the 1st IEEE Workshop on Networking and Database*, 2000.

[141] S. R. Kulkarni, S. E. Posner, and S. Sandilya, "Data-dependent k_n-nn and kernel estimators consistent for arbitrary processes," *IEEE Transactions on Information Theory*, vol. 48, no. 10, pp. 2785–2788, October 2002.

[142] S. Kullback, *Information Theory and Statistics.* New York: John Wiley and Sons, 1959.

[143] S. Kullback, "A lower bound for discrimination information in terms of variation (corresp.)," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 126–127, January 1967.

[144] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, March 1951.

[145] N. Kwak and C. H. Choi, "Input feature selection by mutual information based on Parzen window," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 12, 2002.

[146] J. Kybic, "High-dimensional mutual information estimation for image registration," *Proceedings of the 2004 International Conference on Image Processing* (*ICIP '04*), vol. 3, pp. 1779–1982, October 2004.

[147] R. Landauer, "Information is physical," *Physics Today*, vol. 44, no. 5, pp. 23–29, May 1991.
[148] A. V. Lazo and P. Rathie, "On the entropy of continuous probability distributions (corresp.)," *IEEE Transactions on Information Theory*, vol. 24, no. 1, pp. 120–122, January 1978.
[149] E. Learned-Miller, "Hyperspacings and the estimation of information theoretic quantities," UMass Amherst, Technical Report, 04-104, 2004.
[150] H. S. Leff and A. F. Rex, *Maxwells Demon: Entropy, Information, Computing.* Princeton University Press, 2002.
[151] J.-J. Lin, N. Saito, and R. A. Levine, "Edgeworth expansions of the Kullback-Leibler information," Division of Statistics, University of California, Davis, US, Technical Report, 1999.
[152] C. Liu and H.-Y. Shum, "Kullback-Leibler boosting," *Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, no. 18–20, pp. I–587–I–594, June 2003.
[153] J. Liu and P. Moulin, "Information-theoretic analysis of interscale and intrascale dependencies between image wavelet coefficients," *IEEE Transactions on Image Processing*, vol. 10, no. 11, pp. 1647–1658, November 2001.
[154] D. O. Loftsgaarden and C. P. Quesenberry, "A nonparametric estimate of a multivariate density function," *The Annals of Mathematical Statistics*, vol. 36, no. 3, pp. 1049–1051, June 1965.
[155] B. Logan and A. Salomon, "A music similarity function based on signal analysis," *IEEE International Conference on Multimedia and Expo*, pp. 745–748, August 2001.
[156] S. Lyu, "Infomax boosting," *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 533–538, June 2005.
[157] B. Ma, "Parametric and non-parametric approaches for multisensor data fusion," Ph.D. dissertation, EECS Department, University of Michigan, Ann Arbor, MI, 2001.
[158] M. Madiman and A. Barron, "Generalized entropy power inequalities and monotonicity properties of information," *IEEE Transactions on Information Theory*, vol. 53, no. 7, pp. 2317–2329, July 2007.
[159] M. Madiman and A. R. Barron, "Generalized entropy power inequalities and monotonicity properties of information," *IEEE Transactions on Information Theory*, vol. 53, no. 7, pp. 2317–2329, July 2007.
[160] F. Maes, D. Vandermeulen, and P. Suetens, "Medical image registration using mutual information," *Proceedings of the IEEE*, vol. 91, no. 10, pp. 1699–1722, October 2003.
[161] N. Mars and G. van Arragon, "Time delay estimation in non-linear systems," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 20, no. 3, pp. 619–621, June 1981.
[162] N. Mars and G. van Arragon, "Time delay estimation in non-linear systems using average amount of mutual information analysis," *Signal Processing*, vol. 4, no. 2–3, pp. 139–153, 1982.
[163] J. R. Mathiassen, A. Skavhaug, and K. Bø, "Texture similarity measure using Kullback-Leibler divergence between Gamma distributions," *Proceedings of*

the 7th European Conference on Computer Vision — Part III, pp. 133–147, 2002.

[164] C. R. Maurer and J. M. Fitzpatrick, "A review of medical image registration," in *Interactive Image-Guided Neurosurgery*, (R. J. Maciunas, ed.), pp. 17–44, Park Ridge, IL: American Association of Neurological Surgeons, 1993.

[165] U. Maurer, "Secret key agreement by public discussion," *IEEE Transactions on Information Theory*, vol. 39, no. 3, pp. 733–742, May 1993.

[166] M. Menendez, D. Morales, L. Pardo, and I. Vajda, "Divergence-based estimation and testing of statistical models of classification," *Journal of Multivariate Analysis*, vol. 54, no. 2, pp. 329–354, 1995.

[167] M. E. Meyer and D. V. Gokhale, "Kullback-Leibler information measure for studying convergence rates of densities and distributions," *IEEE Transactions on Information Theory*, vol. 39, pp. 1401–1404, July 1993.

[168] G. Michaels, D. Carr, M. Askenazi, S. Fuhrman, X. Wen, and R. Somogyi, "Cluster analysis and data visualization," *Proceedings of Pacific Symposium Biocomputing*, vol. 3, pp. 42–53, 1989.

[169] E. G. Miller, "A new class of entropy estimators for multi-dimensional densities," *Proceedings of 2003 IEEE International Conference on Acoustics, Speech and Signal* (*ICASSP'03*), vol. 3, pp. 297–300, April 2003.

[170] R. Moddemeijer, "On estimation of entropy and mutual information of continuous distributions," *Signal Processing*, vol. 16, no. 3, pp. 233–246, March 1989.

[171] R. Moddemeijer, "A statistic to estimate the variance of the histogram based mutual information estimator based on dependent pairs of observations," *Signal Processing*, vol. 75, no. 1, pp. 51–63, March 1999.

[172] A. Mokkadem, "Estimation of the entropy and information of absolutely continuous random variables," *IEEE Transactions on Information Theory*, vol. 35, no. 1, pp. 193–196, January 1989.

[173] Y. Moon, B. Rajagopalan, and U. Lall, "Estimation of mutual information using kernel density estimators," *Physical Review E*, vol. 52, no. 3, pp. 2318–2321, September 1995.

[174] P. J. Moreno, P. P. Ho, and N. Vasconcelos, "A Kullback-Leibler divergence based kernel for SVM classification in multimedia applications," HPL-2004-4, Cambridge, MA, Technical Report, 2004.

[175] P. J. Moreno and R. Rifkin, "Using the Fisher kernel method for web audio classification," HP Lab, Cambridge, MA, Technical Report, 2000.

[176] I. Nemenman, W. Bialek, and R. de Ruyter van Steveninck, "Entropy and information in neural spike trains: Progress on the sampling problem," *Physical Review E*, vol. 69, id. 056111, 2004.

[177] I. Nemenman, F. Shafee, and W. Bialek, "Entropy and inference: Revisited," in *Advances in Neural Information Processing Systems*, (T. G. Dietterich, S. Becker, and Z. Ghahramani, eds.), Cambridge, MA: MIT Press, 2002.

[178] D. N. Neuhoff, "On the asymptotic distribution of the errors in vector quantization," *IEEE Transactions on Information Theory*, vol. 42, no. 2, pp. 461–468, March 1996.

[179] X. Nguyen, M. J. Wainwright, and M. I. Jordan, "Divergence Measures, Surrogate Loss Functions and Experimental Design," in *NIPS*, Cambridge, MA: MIT Press, 2005.
[180] M. Nilsson and W. B. Kleijn, "On the estimation of differential entropy from data located on embedded manifolds," *IEEE Transactions on Information Theory*, vol. 53, no. 7, pp. 2330–2341, July 1996.
[181] A. B. Nobel, "Limits to classification and regression estimation from ergodic processes," *The Annals of Statistics*, vol. 27, no. 1, pp. 262–273, February 1999.
[182] A. Okabe, B. Boots, K. Sugihara, and S. N. Chiu, *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams.* Wiley Series in Probability and Statistics, Second ed., 1992.
[183] F. Österreicher, "Csiszár's f-divergences — basic properties," Austria, 2002.
[184] L. Paninski, "Estimation of entropy and mutual information," *Neural Computation*, vol. 15, no. 6, pp. 1191–1253, June 2003.
[185] P. F. Panter and W. Dite, "Quantization in pulse-count modulation with nonuniform spacing of levels," *Proceedings of IRE*, vol. 39, pp. 44–48, 1951.
[186] S. Panzeri and A. Treves, "Analytical estimates of limited sampling biases in different information measures," *Network: Computation in Neural Systems*, vol. 7, pp. 87–107, 1996.
[187] E. Parzen, "On estimation of a probability density function and mode," *The Annals of Mathematical Statistics*, vol. 33, no. 3, pp. 1065–1076, September 1962.
[188] W. D. Penny, "KL-divergences of Normal, Gamma, Dirichlet and Wishart densities," Wellcome department of cognitive neurology, University College London, Technical Report, March 2001. [Online]. Available: www.fil.ion.ucl.ac.uk/wpenny/publications/densities.ps.
[189] M. Pinsker, V. V. Prelov, and S. Verdú, "Sensitivity of channel capacity," *IEEE Transactions on Information Theory*, vol. 33, pp. 405–422, November 1989.
[190] M. S. Pinsker, *Information and Information Stability of Random Variables and Processes.* Izd. Akad. Nauk, 1960, translated by A. Feinstein, 1964.
[191] J. P. Pluim, J. Maintz, and M. A. Viergever, "f-information measures in medical image registration," *Proceedings of the Meeting of the International Society Optical Engineering* (*SPIE*), vol. 4322, pp. 579–587, 2001.
[192] J. P. W. Pluim, J. B. A. Maintz, and M. A. Viergever, "Mutual information based registration of medical images: A survey," *IEEE Transactions on Medical Imaging*, vol. 22, no. 8, pp. 986–1004, August 2003.
[193] H. V. Poor and J. B. Thomas, "Applications of Ali-Silvey distance measures in the design of generalized quantizers for binary decision systems," *IEEE Transactions on Communications*, vol. 25, no. 9, pp. 893–900, September 1977.
[194] V. V. Prelov, "Asymptotic behavior of a continuous channel with small additive noise," *Problems of Information and Transmission*, vol. 4, pp. 31–37, 1968.
[195] V. V. Prelov, "Asymptotic behavior of the capacity of a continuous channel with a large amount of noise," *Problemy Peredachi Informastsii*, vol. 6, pp. 40–57, April–June 1970.

[196] V. V. Prelov, “Communication channel capacity with almost Gaussian noise,” *Theory of Probability and its Applications*, vol. 33, pp. 405–422, 1989.

[197] R. Q. Quiroga, J. Arnhold, K. Lehnertz, and P. Grassberger, “Kullback-Leibler and renormalized entropies: Applications to electroencephalograms of epilepsy patients,” *Physical Review E*, vol. 62, no. 6, pp. 8380–8386, December 2000.

[198] B. L. S. P. Rao, *Nonparametric Functional Estimation.* Academic Press, 1983.

[199] A. Rényi, “On the dimension and entropy of probability distributions,” *Acta Mathematica Academiac Scientiarum Hungaricae*, vol. 10, no. 1–2, pp. 193–215, March 1959.

[200] A. Rényi, “On measures of entropy and information,” *Proceedings of the 4th Berkeley Symposium on Mathematics, Statistics and Probability*, vol. 1, pp. 547–561, 1961.

[201] F. Rieke, D. Warland, R. de Ruyter van Steveninck, and W. Bialek, *Spikes: Exploring the Neural Code.* Cambridge, MA, USA: MIT Press, 1999.

[202] P. M. Robinson, “Consistent nonparametric entropy-based testing,” *The Review of Economic Studies, Special Issue: The Econometrics of Financial Markets*, vol. 58, no. 3, pp. 437–453, May 1991.

[203] C. C. Rodrigez and J. Van Ryzin, “Large sample properties of maximum entropy histograms,” *IEEE Transactions on Information Theory*, vol. 32, pp. 751–759, November 1986.

[204] M. Rosenblatt, “A central limit theorem and a strong mixing condition,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 42, pp. 43–47, 1956.

[205] M. Rosenblatt, “Remarks on some nonparametric estimates of a density function,” *The Annals of Mathematical Statistics*, vol. 27, no. 3, pp. 832–837, September 1956.

[206] M. Rosenblatt, “Density estimates and Markov sequences,” in *Nonparametric Techniques in Statistical Inference*, (M. Puri, ed.), pp. 199–210, Cambridge University Press, London, 1970.

[207] G. Roussas, “Nonparametric estimation in Markov processes,” *Annals of the Institute of Statistical Mathematics*, vol. 21, no. 1, pp. 73–87, December 1969.

[208] G. Roussas, “Nonparametric estimation of the transition distribution function of a Markov process,” *Annals of Mathematical Statistics*, vol. 40, no. 4, pp. 1386–1400, 1969.

[209] M. R. Sabuncu and P. J. Ramadge, “Spatial information in entropy based image registration,” *Proceedings of the 2003 International Workshop on Biomedical Image Registration*, 2003.

[210] M. R. Sabuncu and P. J. Ramadge, “Using spanning graphs for efficient image registration,” *IEEE Transactions on Image Processing*, vol. 17, no. 5, pp. 788–797, May 2008.

[211] P. K. Sahoo and G. Arora, “A thresholding method based on two dimensional Rényi’s entropy,” *Pattern Recognition*, vol. 37, no. 6, pp. 1149–1161, July 2004.

[212] T. D. Schneider, “Use of information theory in molecular biology,” *Proceedings of 1992 Workshop on Physics and Computation* (*PhysComp ’92*), pp. 102–110, October 1992.

[213] E. Schneidman, W. Bialek, and M. J. Berry II, "An information theoretic approach to the functional classification of neurons," in *Advances in Neural Information Processing*, (S. Becker, S. Thrun, and K. OberMayer, eds.), pp. 197–204, Cambridge, US: MIT Press, 2003.

[214] J. Seckbach and E. Rubin, *The New Avenues in Bioinformatics.* Springer, 2004.

[215] G. Shakhnarovich, J. W. Fisher, and T. Darrell, "Face recognition from long-term observations," *Proceedings of the 7th European Conference on Computer Vision — Part III*, pp. 851–868, 2002.

[216] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, pp. 379–423, 623–656, July, October 1948.

[217] C. E. Shannon, "Prediction and entropy of printed English," *Bell System Technical Journal*, vol. 30, no. 1, pp. 50–64, 1951.

[218] P. C. Shields, "The interactions between ergodic theory and information theory," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2079–2093, October 1998.

[219] J. Shore and R. Johnson, "Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy," *IEEE Transactions on Information Theory*, vol. 26, pp. 26–37, January 1980.

[220] R. H. Shumway and A. N. Unger, "Linear discriminant functions for stationary time series," *Journal of the American Statistical Association*, vol. 69, no. 348, pp. 948–956, December 1974.

[221] J. Silva and S. S. Narayanan, "Average divergence distance as a statistical discrimination measure for hidden Markov models," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, pp. 890–906, May 2006.

[222] J. Silva and S. S. Narayanan, "Upper bound Kullback-Leibler divergence for hidden Markov models with application as discrimination measure for speech recognition," *Proceedings of 2006 International Symposium on Information Theory* (*ISIT 2006*), pp. 2299–2303, 2006.

[223] J. Silva and S. S. Narayanan, "Universal consistency of data-driven partitions for divergence estimation," *Proceedings of 2007 IEEE International Symposium on Information Theory* (*ISIT2007*), June 2007.

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

[225] H. Skaug and D. Tjøstheim, "Nonparametric tests of serial independence," in *Developments in Time Series Analysis*, (S. Rao, ed.), pp. 207–229, Chapman and Hill, 1993.

[226] H. Skaug and D. Tjøstheim, "Testing for serial independence using measures of distance between densities," in *Athens Conference on Applied Probability and Time Series*, (P. R. M. Rosenblatt, ed.), Springer Lecture Notes in Statistics, Springer, 1996.

[227] M. B. Skouson, Q. J. Guo, and Z. P. Liang, "A bound on mutual information for image registration," *IEEE Transactions on Medical Imaging*, vol. 20, pp. 843–846, August 1987.

[228] N. Slonim, G. S. Atwal, G. Tkacik, and W. Bialek, "Estimating mutual information and multi-information in large networks," [Online]. Available: http://arxiv.org/abs/cs.IT/0502017, 2005.

[229] A. Stam, "Some inequalities satisfied by the quantities of information of Fisher and Shannon," *Information and Control*, vol. 2, pp. 101–112, June 1959.

[230] R. Steuer, J. Kurths, C. O. Daub, J. Weise, and J. Selbig, "The mutual information: Detecting and evaluating dependencies between variables," *Bioinformatics*, vol. 18, pp. S231–S240, October 2002.

[231] H. Stögbauer, A. Kraskov, S. A. Astakhov, and P. Grassberger, "Least-dependent-component analysis based on mutual information," *Physical Review E*, vol. 70, id. 066123, 2004.

[232] S. P. Strong, R. Koberle, R. R. de Ruyter van Steveninck, and W. Bialek, "Entropy and information in neural spike trains," *Physical Review Letter*, vol. 80, no. 1, pp. 197–200, January 1998.

[233] F. P. Tarasenko, "On the evaluation of an unknown probability density function, the direct estimation of the entropy from independent observations of a continuous random variable and the distribution-free entropy test of goodness-of-fit," *Proccedings of IEEE*, vol. 56, pp. 2052–2053, 1968.

[234] T. Tasdizen, S. P. Awate, R. T. Whitaker, and N. L. Foster, "MRI tissue classification with neighborhood statistics: A nonparametric, entropy-minimizing approach," *Proceedings of the 2005 International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI2005)*, vol. 3750, pp. 517–525, 2005.

[235] H. Theil, "The entropy of the maximum entropy distribution," *Economic Letters*, pp. 145–148, 1980.

[236] H. Theil and D. G. Fiebig, *Exploiting Continuity: Maximum Entropy Estimation of Continuous Distributions*. Cambridge, MA: Ballinger Publishing Company, 1984.

[237] D. Tjøstheim, "Measures and tests of independence: A survey," *Statistics*, vol. 28, pp. 249–284, 1996.

[238] F. Topsøe, "Some inequalities for information divergence and related measures of discrimination," *IEEE Transactions on Information Theory*, vol. 46, no. 4, pp. 1602–1609, July 2000.

[239] K. Torkkola, "Feature extraction by nonparametric mutual information maximization," *Journal of Machine Learning Research*, vol. 3, pp. 1415–1438, March 2003.

[240] A. Tsai, J. W. Fisher III, C. Wible, W. M. Wells III, J. Kim, and A. S. Willsky, "Analysis of functional MRI data using mutual information," *Proceedings of the 2nd International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 473–480, October 1999.

[241] A. Tsai, W. Wells, C. Tempany, E. Grimson, and A. Willsky, "Mutual information in coupled multi-shape model for medical image segmentation," *Medical Image Analysis*, vol. 8, no. 4, pp. 429–445, December 2004.

[242] A. B. Tsybakov and E. C. van der Meulen, "Root-n consistent estimators of entropy for densities with unbounded support," Technical Report, 1992.

[243] A. B. Tsybakov and E. C. van der Meulen, "Root-n consistent estimators of entropy for densities with unbounded support," *Scandinavian Journal of Statistics*, vol. 23, pp. 75–83, 1996.

[244] A. M. Tulino and S. Verdú, "Monotonic decrease of the non-Gaussianness of the sum of independent random variables: A simple proof," *IEEE Transactions on Information Theory*, vol. 52, no. 9, pp. 4295–4297, September 2006.

[245] I. Vajda, "Note on discrimination information and variation," *IEEE Transactions on Information Theory*, vol. 16, no. 6, pp. 771–773, November 1970.

[246] I. Vajda, "On the f-divergence and singularity of probability measures," *Periudicu Mathematica Hiingarrricu*, vol. 2, pp. 223–234, 1972.

[247] I. Vajda, *Theory of Statistical Inference and Information*. Dordrecht-Boston: Kluwer, 1989.

[248] P. A. van den Elsen, E.-J. D. Pol, and M. A. Viergever, "Medical image matching – A review with classification," *IEEE Engineering in Medicine and Biology Magazine*, pp. 26–38, March 1993.

[249] B. Van Es, "Estimating functionals related to a density by a class of statistics based on spacings," *Scandinavian Journal of Statistics*, vol. 19, pp. 61–72, 1992.

[250] M. M. Van Hulle, "Edgeworth approximation of multivariate differential entropy," *Neural Computation*, vol. 17, pp. 1903–1910, September 2005.

[251] V. Vapnik, *The Nature of Statistical Learning Theory*. Springer-Verlag, 1999.

[252] N. Vasconcelos, P. Ho, and P. Moreno, "The Kullback-Leibler kernel as a framework for discriminant and localized representations for visual recognition," *Proceedings of 2004 European Conference on Computer Vision* (*ECCV 2004*), pp. 430–441, May 2004.

[253] O. Vasicek, "A test for normality based on sample entropy," *Journal of the Royal Statistical Society, Series B* (*Methodological*), vol. 38, no. 1, pp. 54–59, 1976.

[254] S. Verdú, "Lectures notes on information theory," unpublished.

[255] S. Verdú, "Fifty years of Shannon theory," *IEEE Transactions on Information Theory*, vol. 44, pp. 2057–2078, October 1998.

[256] S. Verdú, *Multiuser Detection*. Cambridge University Press, 1998.

[257] J. D. Victor, "Binless strategies for estimation of information from neural data," *Physical Review E*, vol. 66, id. 051903, 2002.

[258] A. Wald, "Sequential tests of statistical hypotheses," *Annals of Mathematical Statistics*, vol. 16, no. 2, pp. 117–186, 1945.

[259] D. L. Wallace, "Asymptotic approximations to distributions," *The Annals of Mathematical Statistics*, vol. 29, no. 3, pp. 635–654, September 1958.

[260] Q. Wang, S. R. Kulkarni, and S. Verdú, "Divergence estimation of continuous distributions based on data-dependent partitions," *IEEE Transactions on Information Theory*, vol. 51, no. 9, pp. 3064–3074, September 2005.

[261] Q. Wang, S. R. Kulkarni, and S. Verdú, "A nearest-neighbor approach to estimating divergence between continuous random vectors," *Proceedings of 2006 IEEE International Symposium on Information Theory* (*ISIT2006*), July 2006.

[262] Q. Wang, S. R. Kulkarni, and S. Verdú, "Divergence estimation for multi-dimensional distributions via k-nearest-neighbor distances," *IEEE Transactions on Information Theory*, vol. 55, no. 5, pp. 2392–2405, May 2009.

[263] N. Wiener, *Cybernetics: Control and Communication in the Animal and the Machine*. Cambridge, MA: MIT Press, 1948.

[264] R. Wilson, D. Tse, and R. A. Scholtz, "Channel identification: Secret sharing using reciprocity in ultrawideband channels," *IEEE Transactions on Information Forensics and Security*, vol. 2, pp. 364–375, September 2007.

[265] D. Wolpert and D. Wolf, "Estimating functions of probability distributions from a finite set of samples," *Physical Review E*, vol. 52, pp. 6841–6854, 1995.

[266] Y. Wu, K. L. Chan, and Y. Huang, "Image texture classification based on finite Gaussian mixture models," *Proceedings of The 3rd International Workshop on Texture Analysis and Synthesis* (*In Conjunction with ICCV2003*), pp. 107–112, October 2003.

[267] A. D. Wyner and J. Ziv, "Some asymptotic properties of the entropy of a stationary ergodic data source with applications to data-compression," *IEEE Transactions on Information Theory*, vol. 35, no. 6, pp. 1250–1258, November 1989.

[268] H. Yang, S. van Vuuren, and H. Hermansky, "Relevancy of time-frequency features for phonetic classification measured by mutual information," *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing* (*ICASSP'99*), vol. 1, pp. 225–228, 1999.

[269] C. Ye, A. Reznik, and Y. Shah, "Extracting secrecy from jointly Gaussian random variables," *Proceedings of 2006 IEEE International Symposium on Information Theory (ISIT2006)*, pp. 2593–2597, July 2006.

[270] H. P. Yockey, *Information theory and molecular biology.* New York: Cambridge University Press, 1992.

[271] J. Ziv and N. Merhav, "A measure of relative entropy between individual sequences with application to universal classification," *IEEE Transactions on Information Theory*, vol. 39, no. 4, pp. 1270–1279, July 1993.

[272] K. Życzkowski, "Rényi extrapolation of Shannon entropy," *Open Systems and Information Dynamics*, vol. 10, pp. 297–310, 2003.