Review

Divergence measures for statistical data processing—An annotated bibliography

Michèle Basseville *,1

IRISA, Campus de Beaulieu, 35042 Rennes Cedex, France

ARTICLE INFO

Article history:
Received 19 April 2012
Received in revised form 17 July 2012
Accepted 5 September 2012
Available online 14 September 2012

Keywords:
Divergence
Distance
Information
f-Divergence
Bregman divergence
Learning
Estimation
Detection
Classification
Recognition
Compression
Indexing

This paper provides an annotated bibliography for investigations based on or related to divergence measures for statistical data processing and inference problems.

© 2012 Elsevier B.V. All rights reserved.

Contents

1. Introduction ................................................................. 621
2. f-Divergences ............................................................. 622
3. Bregman divergences .................................................... 623
4. α-Divergences .............................................................. 624
5. Handling more than two distributions ......................... 625
6. Inference based on entropy and divergence criteria ............ 626
7. Spectral divergence measures ........................................ 627
8. Miscellanea ............................................................... 627
Acknowledgment ........................................................ 627
References ................................................................. 627

1. Introduction

Distance or divergence measures are of key importance in a number of theoretical and applied statistical inference and data processing problems, such as estimation, detection,
classification, compression, and recognition [28,111,114],
and more recently indexing and retrieval in databases
[21,58,123,203], and model selection [37,140,184,277].

The literature on such types of issues is wide and has considerably expanded in the recent years. In particular,
following the set of books published during the second half
of the eighties [8,43,66,112,156,167,231,268], a number of
books have been published during the last decade or so
[14,20,35,63,67,85,113,143,173,206,220,256,278,279,282].

The purpose of this paper is to provide an annotated
bibliography for a wide variety of investigations based on
or related to divergence measures for theoretical and
applied inference problems. The bibliography is presented
under the form of a soft classification with some text to
derive the addressed issues instead of a hard classification
made of lists of reference numbers.

The paper is organized as follows. Section 2 is devoted to
f-divergences and Section 3 is focussed on Bregman diver-
gences. The particular and important case of α-divergences
is the topic of Section 4. How to handle divergences between
more than two distributions is addressed in Section 5. Section
6 concentrates on statistical inference based on entropy and
divergence criteria. Divergence measures for multivariable
(Gaussian) processes, including spectral divergence measures,
are reported in Section 7. Section 8 addresses some miscellaneous
issues.

2. f-Divergences

f-Divergences between probability densities\(^2\) \(p(x)\) and
\(q(x)\) are defined as

\[
I_f(p,q) = \int q(x) f \left( \frac{p(x)}{q(x)} \right) \, dx
\]

with \(f\) a convex function satisfying \(f(1) = 0, f'(1) =
0, f''(1) = 1\). They have been introduced in the sixties inde-
dependently by Ali and Silvey [6], by Csiszár [68,69] and by
Morimoto [193], and then again in the seventies by Akaike
[5] and by Ziv and Zalkai [284]. Kullback–Leibler divergence

\[
K(p,q) = \int p(x) \ln \left( \frac{p(x)}{q(x)} \right) \, dx
\]

Hellinger distance, \(\chi^2\)-divergence, Csiszár α-divergence dis-
cussed in Section 4, and Kolmogorov total variation distance
are some well known instances of f-divergences. Other
instances may be found in [151,152,169,236].

f-Divergences enjoy some invariance properties investi-
gated in [167,168] (see also [29]), among which the two follow-
ing properties:

\[
\begin{align*}
&\text{for } f(u) = f(u) + \gamma u + \delta, \quad I_f(p,q) = I_f(p,q) + \gamma + \delta \\
&\text{for } f(u) = uf \left( \frac{1}{u} \right), \quad I_f(p,q) = I_f(p,q)
\end{align*}
\]

are used in the sequel. They also enjoy a universal mon-
onotonicity property known under the name of general-
ized data processing theorem [187,284]. Their topological
properties are investigated in [70]. An extension of the
family of f-divergences to squared metric distances is
introduced in [270].

Non-asymptotic variational formulations of f-divergences
have been recently investigated in [48,199,232,233]. Early
results on that issue obtained for Kullback–Leibler diver-
gence trace back to [90,112]. Such variational formulations
are of key interest for the purpose of studying the properties
of f-divergences or designing algorithms based on duality.
The application of variational formulation to estimating
divergence functionals and the likelihood ratio is addressed
in [200]. Other methods for estimating divergences have
been proposed in [219,274].

f-Divergences can usefully play the role of surrogate
functions, that are functions majorizing or minorizing
the objective or the risk functions at hand. For example,
f-divergences are used for defining loss functions that
yield Bayes consistency for joint estimation of the discri-
miminant function and the quantizer in [199], as surrogate
functions for independence and ICA in [183], and the
α-divergence in [14] is used in [182] as a surrogate function
for maximizing a likelihood in an EM-type algorithm. Bounds
on the minimax risk in multiple hypothesis testing and estimation
problems are expressed in terms of the f-divergences in
[41,116], respectively. An f-divergence estimate is exploited
for deriving a two-sample homogeneity test in [139].

f-Divergences, used as general (entropic) distance-like
functions, allow a non-smooth non-convex optimization
formulation of the partitioning clustering problem, namely
the problem of clustering with a known number of classes,
for which a generic iterative scheme keeping the simplicity
of the k-means algorithm is proposed in [258,259].

f-Divergences also turn out useful for defining robust
projection pursuit indices [196]. Convergence results of
projection pursuit through f-divergence minimization
with the aim of approximating a density on a set with
very large dimension are reported in [263].

Nonnegative matrix factorization (NMF), of wide-
spread use in multivariate analysis and linear algebra
[89], is another topic that can be addressed with
f-divergences. For example, NMF is achieved with the
aid of Kullback divergence and alternating minimization
in [98]. Itakura–Saito divergence in [102,163], f-diverge-
ces in [61], or α-divergences in [62,154]; see also [63].

The maximizers of the Kullback–Leibler divergence
from an exponential family and from any hierarchical
log-linear model are derived in [229,185], respectively.

Other investigations include comparison of experiments
[262]; minimizing f-divergences on sets of signed mea-
ures [47]; minimizing multivariate entropy functionals
with application to minimizing f-divergences in both vari-
ables [78]; determining the joint range of f-divergences
[119]; or proving that the total variation distance is the
only f-divergence which is an integral probability metric
(IPM) used in the kernel machines literature [246].

Recent applications involving the use of f-divergences
concern feature selection in fuzzy approximation spaces
[175], the selection of discriminative genes from micro-
array data [174], speckle data acquisition [195], medical
image registration [218], or speech recognition [221], to
mention but a few examples. A modification of f-divergences

---

\(^2\) With respect to the Lebesgue measure.
3. Bregman divergences

Bregman divergences, introduced in [46], are defined for vectors, matrices, functions and probability distributions. The Bregman divergence between vectors is defined as

\[ D_{\phi}(x,y) = \phi(x) - \phi(y) - \langle x - y, \nabla \phi(y) \rangle \]

(5)

with \( \phi \) a differentiable strictly convex function \( \mathbb{R}^d \rightarrow \mathbb{R} \). The symmetrized Bregman divergence writes

\[ D_{\phi}(x,y) = \langle \nabla \phi(x) - \nabla \phi(y), x - y \rangle \]

(6)

The Bregman matrix divergence is defined as

\[ D_{\phi}(X,Y) = \phi(X) - \phi(Y) - \text{Tr}((\nabla \phi(Y))^T(X - Y)) \]

(7)

for \( X,Y \) real symmetric \( d \times d \) matrices, and \( \phi \) a differentiable strictly convex function \( \mathbb{R}^d \rightarrow \mathbb{R} \). Those divergences preserve rank and positive semi-definiteness [88].

For \( \phi(X) = \ln|X| \), the divergence (7) is identical to the distance between two positive matrices defined as the Kullback–Leibler divergence between two Gaussian distributions having those matrices as covariance matrices. According to [155], that distance is likely to trace back to [129]. For example, it has been used for estimating structured covariance matrices [54] and for designing residual generation criteria for monitoring [30].

The divergence (7), in the general case for \( \phi \), has been recently proposed for designing and investigating a new family of self-scaling quasi-Newton methods [137,138]. The Bregman divergence between probability densities \( p(x) \) and \( q(x) \) is defined as [73,134]

\[ D_{\phi}(p,q) = \int (\phi(p(x)) - \phi(q(x)) - (p(x) - q(x)) \phi'(q(x))) \, dx \]

(8)

for \( \phi \) a differentiable strictly convex function. A Bregman divergence can also be seen as the limit of a Jensen difference [31,201], namely

\[ D_{\phi}(p,q) = \lim_{\beta \to 0} \frac{1}{\beta^2} f_{\phi}^{(\beta)}(p,q) \]

(9)

where the Jensen difference \( f_{\phi}^{(\beta)} \) is defined for \( 0 < \beta < 1 \) as

\[ f_{\phi}^{(\beta)}(p,q) = \beta \phi(p) + (1-\beta)\phi(q) - \phi(\beta p + (1-\beta)q) \]

(10)

For \( \beta = 1/2 \), Jensen difference is the Burbea–Rao divergence [53]; see also [93,169]. The particular case where \( D \) and \( J \) are identical is of interest [228].

The Bregman divergence measures enjoy a number of properties useful for learning, clustering and many other inference [26,73,99] and quantization [24] problems. They have been recently used for density-ratio matching [252]. In the discrete case, an asymptotic equivalence with the f-divergence and the Burbea–Rao divergence is investigated in [208]. More generally the relationships between the f-divergences and the Bregman divergences have been described in [29,31].

One important instance of the use of Bregman divergences for learning is the case of inverse problems [134,160].

3 See also [244] for another investigation of regularization.

where convex duality is extensively used. Convex duality is also used for minimizing a class of Bregman divergences subject to linear constraints in [81], whereas a simpler proof using convex analysis is provided in [36], and the results are used in [64].

Bregman divergences have been used for a generalization of the LMS adaptive filtering algorithm [150], for Bayesian estimation of distributions [100] using functional divergences introduced for quadratic discriminant analysis [247], and for \( l_1 \)-regularized logistic regression posed as Bregman distance minimization and solved with non-linear constrained optimization techniques in [80]. Iterative \( l_1 \)-minimization with application to compressed sensing is investigated in [280]. Those divergences are also useful for solving clustering problems. Used as general (entropic) distance-like functions, they allow a non-smooth non-convex optimization formulation of the partitioning clustering problem, for which a generic iterative scheme keeping the simplicity of the \( k \)-means algorithm is proposed in [258,259]. Clustering with Bregman divergences unifying \( k \)-means, LBG and other information theoretic clustering approaches is investigated in [26], together with a connection with rate distortion theory.4 Scaled Bregman distances are used in [251] and compared with \( f \)-divergences for key properties: optimality of the \( k \)-means algorithm [26] and invariance w.r.t. statistically sufficient transformations. Quantization and clustering with Bregman divergences are investigated in [99] together with convergence rates. \( k \)-means and hierarchical classification algorithms w.r.t. Burbea–Rao divergences (expressed as Jensen–Bregman divergences) are studied in [201].

The interplay between Bregman divergences and boosting, in particular AdaBoost, has been the topic of a number of investigations, as can be seen for example in [64,149,158,157] for an earlier study. Some controversy does exist however, see for example [161] where understanding the link between boosting and ML estimation in exponential models does not require Bregman divergences. The extension of AdaBoost using Bregman divergences, their geometric understanding and information

4 Note that minimum cross-entropy classification was addressed as an extension of coding by vector quantization in [243].
geometry is investigated in [194], together with consistency and robustness properties of the resulting algorithms.

The problems of matrix learning, approximation, factorization can also usefully be addressed with the aid of Bregman divergences. Learning symmetric positive definite matrices with the aid of matrix exponentiated gradient updates and Bregman projections is investigated in [264]. Learning low-rank positive semi-definite (kernel) matrices for machine learning applications is also addressed with Bregman divergences in [155]. Matrix rank minimization is achieved with a Bregman iterative algorithm in [172]. Nonnegative matrix approximation (NMF) with Bregman divergences is addressed in [62]; see also [63]. The particular case of using the density power divergence and a surrogate function for NMF is investigated in [103].

Applications involving the use of Bregman divergences concern nearest neighbor retrieval [57], color image segmentation [201], 3D image segmentation and word alignment [257], cost-sensitive classification for medical diagnosis (UCI datasets) [238], magnetic resonance image analysis [271], semi-supervised clustering of high dimensional text benchmark datasets and low dimensional UCI datasets [276], content-based multimedia retrieval with efficient neighbor queries [203], efficient range search from a query in a large database [58].

4. \( \alpha \)-Divergences

A large number of divergence measures, parameterized by \( \alpha \) and possibly \( \beta \) and/or \( \gamma \), have been introduced in the literature using an axiomatic point of view [3]. This can be seen for example in [255], and the reader is referred to [2,72] for critical surveys. However, a number of useful \( \alpha \)-divergences measures have been proposed, tracing back to Chernoff [59] and Rao [223]. A recent synthesis can be found in [60]; see also [63].

The Csiszár \( t \)-divergences of order \( \alpha \), also called \( \chi^2 \)-divergence of order \( \alpha \), have been introduced in [69] as \( \alpha \)-divergences (1) associated with

\[
g_{\alpha}(u) = \begin{cases} 
\frac{1}{\alpha(1-\alpha)}(2u + 1 - 2u^\alpha), & \alpha \neq 0, 1 \\
u - 1 - \ln u, & \alpha = 0 \\
1 - u + u \ln u, & \alpha = 1 
\end{cases}
\]

namely

\[
l_{\alpha}(p,q) = \begin{cases} 
\frac{1}{\alpha(1-\alpha)}(1 - \int p^{\alpha+1} dx), & \alpha \neq 0, 1 \\
k(p,q), & \alpha = 0 \\
k(p,q), & \alpha = 1 
\end{cases}
\]

where \( K(p,q) \) is defined in (2). They could also be called Havrda–Charvát’s \( \alpha \)-divergences [121]. They are identical to Read–Cressie’s power divergence [170,231]. See also [3,167,267,268].

It is easily seen that, up to a transformation of \( \alpha \), they are identical to Amari’s \( \alpha \)-divergences [8,14]—see also [10], and to the Tsallis divergences [265]. Actually, the \( \alpha \)-divergences (1) associated with

\[
f_{\alpha}(u) = \begin{cases} 
\frac{4}{1 - \alpha^2}(u - u^{\alpha+1/2}), & \alpha \neq \pm 1 \\
u \ln u, & \alpha = \pm 1 \\
-\ln u, & \alpha = 0 
\end{cases}
\]

considered in [10] write

\[
l_{\beta}(p,q) = \begin{cases} 
\frac{4}{1 - \alpha^2}(1 - \int p^{1+\alpha/2} q^{1-\alpha/2} dx), & \alpha \neq 0, 1 \\
k(p,q), & \alpha = 0 \\
k(q,p), & \alpha = 1 
\end{cases}
\]

and it can be checked that, for \( \beta = (1+\alpha)/2 \), we have \( l_{\beta}(p,q) = l_{\alpha}(p,q) \).

Moreover the equivalence with the \( \alpha \)-divergences built on the function

\[
f_{\alpha}(u) = \begin{cases} 
\frac{4}{1 - \alpha^2}(1 - u^{\alpha+1/2} - 2(1 - u)), & \alpha \neq 0, 1 \\
u \ln u - (u - 1), & \alpha = 0 \\
-\ln(u + (u - 1)), & \alpha = 1 
\end{cases}
\]

often considered stands from the invariance property (3).

For \( \alpha = 1 \) the divergence (14) is nothing but the Kullback–Leibler information, and for \( \alpha = 0 \) it is the Hellinger distance up to a multiplicative constant.

The \( \alpha \)-divergence (14) has been used earlier by Chernoff for investigating the error exponents and asymptotic efficiency of statistical tests with independent observations [59].

Some applications of those divergences are described in particular for model integration in [10], in EM algorithms [180–182], and message-passing algorithms for complex Bayesian networks approximation in [189].

It has been recently proved that they form the unique class belonging to both the \( \alpha \)-divergences and the Bregman divergences classes [11]. This extends the result in [73] that Kullback–Leibler divergence is the only Bregman divergence which is an \( \alpha \)-divergence.

\( \alpha \)-Divergences based on Arimoto’s entropies [18] and introduced in [248] define \( \alpha \)-divergences different from the above ones.

A different class of \( \alpha \)-divergences, known under the name of density power divergences, have been introduced in [32] by Basu et al. as

\[
D_{\alpha}(p,q) = \begin{cases} 
\frac{1}{\alpha} \int (p^{\alpha+1} - (\alpha + 1)p q^\alpha + \alpha q^{\alpha+1}) dx, & \alpha > 0 \\
k(p,q), & \alpha = 0 
\end{cases}
\]

They can be seen as Bregman divergences \( D_{\psi_{\alpha}} \) (8) associated with

\[
\psi_{\alpha}(u) = \begin{cases} 
\frac{1}{\alpha}(u^{\alpha+1} - u), & \alpha > 0 \\
u \ln u, & \alpha = 0 
\end{cases}
\]
They have been used for robust blind source separation [188], analyzing mixtures ICA models [192], model selection [140,141,176,184], estimation of tail index of heavy-tailed distributions [148] and estimation in age-stratified Poisson regression models for cancer surveillance [178]. They have recently been handled in [92] for distributions with mass not necessarily equal to one, an extension useful for designing boosting methods.

The Rényi’s $\alpha$-divergences have been defined in [234] as
\[
D_\alpha(p,q) = \begin{cases} 
\frac{1}{\alpha(\alpha-1)} \ln \int p^{\alpha q^{1-\alpha}} \, dx, & \alpha \neq 1 \\
K(p,q), & \alpha = 1
\end{cases}
\]
although they may have been proposed earlier [40,239]. Those divergences exhibit direct links with the Chernoff distance and with the moment generating function of the likelihood ratio [29,123]. Their use for channel capacity [6] distance and with the moment generating function of the [39], whereas characterizations of maximum Rényi’s entropy distributions are provided in [120,272].

Defining divergences between more than two distributions is useful for discrimination [186] and taxonomy [224,225], where they may be more appropriate than pairwise divergences. They are often called generalized $f$-divergences.

Generalized $f$-divergences have been introduced under the name of $f$-dissimilarity in [117]. For a set of $n$ probability distributions $p_1, \ldots, p_n$ with normalized positive weights $\beta_1, \ldots, \beta_n$, respectively, generalized Jensen divergences are defined as
\[
f_\psi^{(b)}(p_1, \ldots, p_n) = \sum_{i=1}^n \beta_i (\psi(p_i)) - \psi \left( \sum_{i=1}^n \beta_i p_i \right)
\]
where $\psi$ is a convex function.

The case of Shannon entropy, where $\psi(x) = -H(x)$ with $H(x) = -x \ln x$, has been addressed in [169]. In this case, it is easy to show that $f_\psi^{(b)}$ writes as the weighted arithmetic mean of the Kullback distances between each of the $p_i$'s and the barycenter of all the $p_i$'s. This property has been applied to word clustering for text classification [86].

If in addition $n=2$, the latter property holds for more general convex functions $\psi$: Jensen divergence can be written as the arithmetic mean of Bregman divergences to the barycenter,
\[
f^{(b)}_{\psi}(p_1, p_2) = \beta D_{\psi}(p_1, \beta p_1 + (1-\beta)p_2) + (1-\beta)D_{\psi}(p_2, \beta p_1 + (1-\beta)p_2)
\]

Actually the interplay between divergence measures and the notions of entropy, information and generalized mean values is quite tight [29,31,234]. More precisely, mean values can be associated with entropy-based divergences in two different ways. The first way [3,234] consists in writing explicitly the generalized mean values $\psi^{-1}(\sum_{i=1}^n \beta_i (\psi(p_i)))$ underlying the $f$-divergences. The Rényi’s $\alpha$-divergences (17) correspond to $\psi(u) = u^\alpha$, and this results in the $\alpha$-mean values $\left(\sum_{i=1}^n \beta_i p_i^\alpha\right)^{1/\alpha}$. The second way [38] consists in defining mean values by $\arg\min_{u} \sum_{i=1}^n \beta_i f(v/u_i)$, namely as projections, in the sense of distance $d$, onto the half-line $u_1 = \cdots = u_n > 0$ [75]. When $d$ is a $f$-divergence $d(v,u_i) = u_i f(v/u_i)$, this gives the entropic means [38], which are characterized implicitly by $\sum_{i=1}^n \beta_i f(v/u_i) = 0$, and necessarily homogeneous (scale invariant). The class of entropic means includes all available integral means and, when applied to a random variable, contains most of the centrality measures (moments, quantiles). When $d$ is a Bregman distance $d_k(u,v) = h(u)-h(v)-(u-v)h'(v)$, the corresponding mean values are exactly the above generalized mean values (for $\psi = h'$), which are generally not homogenous.

Consequently some means other than the arithmetic mean may be used in the definition (18) of generalized divergences. For example, the information radius introduced in [245] is the generalized mean of Rényi’s divergences between each of the $p_i$’s and the generalized mean of all the $p_i$’s, which boils down to [29]:
\[
S^{(b)}_{\alpha}(p_1, \ldots, p_n) = \frac{\alpha}{\alpha-1} \ln \int \left( \sum_{i=1}^n \beta_i p_i^\alpha(x) \right)^{1/\alpha} \, dx
\]
See also [74].

Mean values, barycenters, centroids have been widely investigated. The barycenter of a set of probability measures is studied in [214]. Barycenters in a dually flat space are introduced as minimizers of averaged Amari’s divergences in [212]. Left-sided, right-sided and symmetrized centroids are introduced as minimizers of averaged Bregman divergences in [202], whereas Burbea–Rao centroids are the topic of [201] with application to color image segmentation.

Geometric means of symmetric positive definite matrices are investigated in [19]. A number of Fréchet means are discussed in [91] with application to diffusion tensor imaging. Quasi-arithmetic means of multiple positive matrices by symmetrization from the mean of two matrices are investigated in [216]. Riemannian metrics on space of matrices are addressed in [215]. The relation of the symmetrized Kullback–Leibler divergence with the Riemannian distance between positive definite matrices is addressed in [86].
The Riemannian distance between positive definite matrices turns out to be a useful tool for analyzing signal processing problems such as the convergence of Riccati [45,159,162].

6. Inference based on entropy and divergence criteria

The relevance of an information theoretic view of basic problems in statistics has been known for a long time [235]. Whereas maximum likelihood estimation (MLE) minimizes the Kullback–Leibler divergence $\text{KL}(p, p^*)$ between an empirical and a true (or reference) distributions, minimizing other divergence measures turns out useful for a number of inference problems, as many examples in the previous sections suggested. Several books exist on such an approach to inference [35,67,167,173,206,268], and the field is highly active.

A number of point estimators based on the minimization of a divergence measure have been proposed. Power divergence estimates, based on the divergence (16) and written as M-estimates, are investigated in [32] in terms of consistency, influence function, equivariance, and robustness; see also [33,170,209]. Iteratively reweighted estimating equations for robust minimum distance estimation are proposed in [34] whereas a bootstrap root search is discussed in [177]. Recent investigations of the power divergence estimates include robustness to outliers and a local learning property [92] and Bahadur efficiency [118]. An application to robust blind source separation is described in [188].

The role of duality when investigating divergence minimization for statistical inference is addressed in [7,48]; see also [115,283]. A further investigation of the minimum divergence estimates introduced in [48] can be found in [260], which addresses the issues of influence function, asymptotic relative efficiency, and empirical performances. See also [213] for another investigation.

A comparison of density-based minimum divergence estimates is presented in [135]. A recent comparative study of four types of minimum divergence estimates is reported in [50], in terms of consistency and influence curves: this includes the power divergence estimates [32], the so-called power superdivergence estimates [47,168,269], the power subdivergence estimates [48,269], and the Rényi pseudo-distance estimates [164,269].

The efficiency of estimates based on a Havrda–Charvát’s $\alpha$-divergence, or equivalently on the Kullback–Leibler divergence with respect to a distorted version of the true density, is investigated in [97].

Robust LS estimates with a Kullback–Leibler divergence constraint are introduced and investigated in [165] where a connection with risk-sensitive filtering is established.

Hypothesis testing may also be addressed within such a framework [42]. The asymptotic distribution of a generalized entropy functional and the application to hypothesis testing and design of confidence intervals are studied in [94]. The asymptotic distribution of tests statistics built on divergence measures based on entropy functions is derived in [206,207]. This includes extensions of Burbea–Rao’s $J$-divergences [52,53] and of Sibson’s information radius [245].

Tests statistics based on entropy or divergence of hypothetical distributions with ML estimated values of the parameter have been recently investigated in [48–50]. Robustness properties of these tests are proven in [260]. The issue of which $f$-divergence should be used for testing goodness of fit is to be studied with the aid of the results in [119].

The key role of Kullback–Leibler divergence (2) for hypothesis testing in the i.i.d. case is outlined in Chernoff’s results about the error exponents [59]. These results have been extended to Gaussian detection [23,65,253] and to the detection of Markov chains [15,197]. The role of Kullback–Leibler divergence for composite hypothesis testing has been outlined and investigated by Hoeffding in the multinomial case [126]. Those universal (asymptotically optimal) tests have been extended to the exponential family of distributions in [84].

Maximum entropy, minimum divergence and Bayesian decision theory are investigated in [115] using the equilibrium theory of zero-sum games. Maximizing entropy is shown to be dual of minimizing worst-case expected loss. An extension to arbitrary decision problems and loss functions is provided, maximizing entropy is shown to be identical to minimizing a divergence between distributions, and a generalized redundancy-capacity theorem is proven. The existence of an equilibrium in the game is rephrased as a Pythagorean property of the related divergence.

The extension of the properties of the K–L divergence minimization [242] to Tsallis divergence is investigated in [266].

Generalized minimizers of convex integral functionals are investigated in detail in [79], extending the results obtained for the Shannon case in [77] to the general case.

Other learning criteria have been investigated in specific contexts. Whereas minimizing the Kullback–Leibler divergence (ML learning) turns out to be difficult and/or slow to compute with MCMC methods for complex high dimensional distributions, contrastive divergence (CD) learning [124] approximately follows the gradient of the difference of two divergences:
\[
\text{CD}_n = \text{KL}(p_0, p^*) - \text{KL}(p_n, p^*)
\]
and provides estimates with typically small bias. Fast CD learning can thus be used to get close to the ML estimate, and then slow ML learning helps refining the CD estimate [56,190]. The convergence properties of contrastive divergence learning are analyzed in [254]. The application to fast learning of deep belief nets is addressed in [125].

On the other hand, score matching consists in minimizing the expected square distance between the model score function and the data score function:
\[
J(\theta) = 1/2 \int_{\xi \in \mathbb{R}^d} p_x(\xi) ||\psi_\theta(\xi) - \psi_\xi(\xi)||^2 \, d\xi
\]
with $\psi_\theta(\xi) \triangleq \nabla_\xi \ln p_\theta(\xi)$ and $\psi_\xi(\xi) \triangleq \nabla_\xi \ln p_\xi(\cdot)$. This objective function turns out to be very useful for estimating non-normalized statistical models [127,128].
7. Spectral divergence measures

Spectral distance measures for scalar signal processing have been thoroughly investigated in [111,114]. Spectral distances between vector Gaussian processes have been studied in [144–146,240,241]; see also [147, Chap. 11] for general stochastic processes.

Kullback–Leibler divergence has been used for approximating Gaussian variables and Gaussian processes and outlining a link with subspace algorithm for system identification [249]. A distance based on mutual information for Gaussian processes is investigated in [44].

Kullback–Leibler and/or Hellinger distances have been used for spectral interpolation [55,104,142], spectral approximation [96,107,210], spectral estimation [222], and ARMA modeling [108].

Differential-geometric structures for prediction and smoothing problems for spectral density functions are introduced in [105,130,131]. This work has been pursued in [281] for the comparison of dynamical systems with the aid of the Kullback–Leibler rate pseudo-metric.

An axiomatic approach to metrics for power spectra can be found in [106].

The geometry of maximum entropy problems is addressed in [211].

The author is indebted to the Associate Editor and the Reviewers, whose careful reading and detailed comments helped in improving an earlier version of the paper.

Acknowledgment

The author is indebted to the Associate Editor and the Reviewers, whose careful reading and detailed comments helped in improving an earlier version of the paper.

References


