Signal Processing 93 (2013) 621–633

Contents lists available at SciVerse ScienceDirect

Signal Processing

journal homepage: www.elsevier.com/locate/sigpro

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

ABSTRACT

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

Contents

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,

* Tel.: +33 299 847 100; fax: +33 299 847 171.
E-mail address: michele.basseville@irisa.fr
[1] This author is with CNRS.

0165-1684/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.sigpro.2012.09.003

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 describe 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 divergences. 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

$$I_f(p,q) = \int q(x) f\left(\frac{p(x)}{q(x)}\right) dx \tag{1}$$

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

$$K(p,q) = \int p(x) \ln \frac{p(x)}{q(x)} dx, \tag{2}$$

Hellinger distance, χ^2-divergence, Csiszár α-divergence discussed 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 investigated in [167,168] (see also [29]), among which the two following properties:

$$\text{for } \hat{f}(u) = f(u) + \gamma u + \delta, \quad I_{\hat{f}}(p,q) = I_f(p,q) + \gamma + \delta \tag{3}$$

$$\text{for } \check{f}(u) = u f\left(\frac{1}{u}\right), \quad I_{\check{f}}(p,q) = I_f(q,p) \tag{4}$$

are used in the sequel. They also enjoy a universal *monotonicity* property known under the name of generalized 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 divergence 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 discriminant 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 widespread 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-divergences 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 measures [47]; minimizing multivariate entropy functionals with application to minimizing f-divergences in both variables [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 microarray 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.

for finite measures turns out to be useful for right-censored observations [250].

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

$$D_{\varphi}(x,y) = \varphi(x) - \varphi(y) - (x-y)^T \nabla \varphi(y) \tag{5}$$

with φ a differentiable strictly convex function $\mathbb{R}^d \longrightarrow \mathbb{R}$. The symmetrized Bregman divergence writes

$$\overline{D}_{\varphi}(x,y) = (\nabla \varphi(x) - \nabla \varphi(y))^T (x-y) \tag{6}$$

The Bregman matrix divergence is defined as

$$D_{\phi}(X,Y) = \phi(X) - \phi(Y) - \mathrm{Tr}((\nabla \phi(Y))^T (X-Y)) \tag{7}$$

for X,Y real symmetric $d \times d$ matrices, and ϕ a differentiable strictly convex function $\mathbb{S}^d \longrightarrow \mathbb{R}$. 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 ϕ, 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]

$$D_{\varphi}(p,q) = \int (\varphi(p(x)) - \varphi(q(x)) - (p(x)-q(x))\varphi'(q(x)))\, dx \tag{8}$$

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

$$D_{\varphi}(p,q) = \lim_{\beta \to 0} \frac{1}{\beta} J_{\varphi}^{(\beta)}(p,q), \tag{9}$$

where the Jensen difference $J_{\varphi}^{(\beta)}$ is defined for $0 < \beta < 1$ as

$$J_{\varphi}^{(\beta)}(p,q) \triangleq \beta \varphi(p) + (1-\beta)\varphi(q) - \varphi(\beta p + (1-\beta) q) \tag{10}$$

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] 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 nonlinear constrained optimization techniques in [80]. Iterative l_1-minimization with application to compressed sensing is investigated in [280]. Mixture models are estimated using Bregman divergences within a modified EM algorithm in [101]. Learning continuous latent variable models with Bregman divergences and alternating minimization is addressed in [275]. The use of Bregman divergences as *surrogate loss functions* for the design of minimization algorithms for learning that yield guaranteed convergence rates under weak assumptions is discussed in [205].

Learning structured models, such as Markov networks or combinatorial models, is performed in [257] using large-margin methods, convex–concave saddle point problem formulation and dual extra-gradient algorithm based on Bregman projections. Proximal minimization schemes handling Bregman divergences that achieve message-passing for graph-structured linear programs (computation of MAP configurations in Markov random fields) are investigated in [230].

There are also many instances of application of Bregman divergences 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

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

[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 with low rank matrices is discussed in [87], whereas matrix approximation based on the minimum Bregman information principle (generalization to all Bregman loss functions of MaxEnt and LS principles) is the topic of [25]. Nonnegative matrix factorization (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[5] [276], content-based multimedia retrieval with efficient neighbor queries [203], efficient range search from a query in a large database [58].

4. α-Divergences

A large number of divergence measures, parameterized by α and possibly β and/or γ, 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 α-divergence 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 I-divergences of order* α, also called χ^2-divergence of order α, have been introduced in [69] as *f-divergences* (1) associated with

$$g_\alpha(u) = \begin{cases} \dfrac{1}{\alpha(1-\alpha)}(\alpha u + 1 - \alpha - u^\alpha), & \alpha \neq \{0,1\} \\ u - 1 - \ln u, & \alpha = 0 \\ 1 - u + u \ln u, & \alpha = 1 \end{cases} \tag{11}$$

namely

$$I_{g_\alpha}(p,q) = \begin{cases} \dfrac{1}{\alpha(1-\alpha)}\left(1 - \int p^\alpha q^{1-\alpha}\, dx\right), & \alpha \neq \{0,1\} \\ K(q,p), & \alpha = 0 \\ K(p,q), & \alpha = 1 \end{cases} \tag{12}$$

where $K(p,q)$ is defined in (2). They could also be called Havrda–Charvát's α-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 α, they are identical to *Amari's α-divergences* [8,14]—see also [10], and to the *Tsallis divergences* [265]. Actually, the *f*-divergences (1) associated with

$$f_\alpha(u) = \begin{cases} \dfrac{4}{1-\alpha^2}(u - u^{(1+\alpha)/2}), & \alpha \neq \pm 1 \\ u \ln u, & \alpha = +1 \\ -\ln u, & \alpha = -1 \end{cases} \tag{13}$$

considered in [10] write

$$I_{f_\alpha}(p,q) = \begin{cases} \dfrac{4}{1-\alpha^2}\left(1 - \int p^{(1+\alpha)/2} q^{(1-\alpha)/2}\, dx\right), & \alpha \neq \pm 1 \\ K(p,q), & \alpha = +1 \\ K(q,p), & \alpha = -1 \end{cases} \tag{14}$$

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

Moreover the equivalence with the *f*-divergences built on the function

$$\tilde{f}_\alpha(u) = \begin{cases} \dfrac{4}{1-\alpha^2}(1 - u^{(1+\alpha)/2}) - \dfrac{2}{1-\alpha}(u-1), & \alpha \neq \pm 1 \\ u \ln u - (u-1), & \alpha = +1 \\ -\ln u + (u-1), & \alpha = -1 \end{cases}$$

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 α-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 in [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 *f*-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 *f*-divergence.

f -Divergences based on Arimoto's entropies [18] and introduced in [248] define α-divergences different from the above ones.

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

$$D_\alpha(p,q) = \begin{cases} 1/\alpha \int (p^{\alpha+1} - (\alpha+1) p q^\alpha + \alpha q^{\alpha+1})\, dx, & \alpha > 0 \\ K(p,q), & \alpha = 0 \end{cases} \tag{15}$$

They can be seen as *Bregman divergences* D_{φ_α} (8) associated with

$$\varphi_\alpha(u) = \begin{cases} 1/\alpha(u^{\alpha+1} - u), & \alpha > 0 \\ u \ln u, & \alpha = 0 \end{cases} \tag{16}$$

[5] For which a non-parametric approach to learning φ in (6) in the form $\varphi(x) = \sum_{i=1}^N \beta_i h(x_i^T x)$ with h a strictly convex function $\mathbb{R} \longrightarrow \mathbb{R}$ is used for distance metric learning.

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 α-divergences* have been defined in [234] as

$$D_\alpha(p,q) = \begin{cases} \dfrac{1}{\alpha(\alpha-1)} \ln \int p^\alpha q^{1-\alpha}\, dx, & \alpha \neq 1 \\ K(p,q), & \alpha = 1 \end{cases} \tag{17}$$

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] and their link with cutoff rates are addressed in [74]. Their involvement in estimation and coding problems is also address in [16,17,198]. Scale and concentration invariance properties have been investigated in [153].

A surrogate function for Rényi's α-divergence is the α-Jensen difference [123]. Entropy functionals derived from Rényi's divergences have been recently studied in [39], whereas characterizations of maximum Rényi's entropy distributions are provided in [120,272].

In recent years, Rényi's α-divergences have been used for robust image registration [122], differentiating brain activity [22], for feature classification, indexing, and retrieval in image and other databases [123], and detecting distributed denial-of-service attacks [166]. They have been shown to be possibly irrelevant for blind source separation [217,273].

A *two-parameter family of divergences* associated with a generalized mean, and reducing to Amari's α-divergences or Jensen difference in some cases, is investigated in [283], together with convex inequalities and duality properties related to divergence functionals.

5. Handling more than two distributions

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 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

$$J_\varphi^{(\beta)}(p_1,\ldots,p_n) \triangleq \sum_{i=1}^n \beta_i \varphi(p_i) - \varphi\left(\sum_{i=1}^n \beta_i p_i\right) \tag{18}$$

where φ is a convex function.

The case of Shannon entropy, where $\varphi(x) = -H(x)$ with $H(x) \triangleq -x \ln x$, has been addressed in [169]. In this case, it is easy to show that $J_\varphi^{(\beta)}$ 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 φ: Jensen divergence can be written as the arithmetic mean of Bregman divergences to the barycenter,

$$J_\varphi^{(\beta)}(p_1,p_2) = \beta D_\varphi(p_1,\beta p_1+(1-\beta)p_2)+(1-\beta)D_\varphi(p_2,\beta p_1+(1-\beta)p_2) \tag{19}$$

Actually the interplay between divergence measures and the notions of entropy, information and *generalized mean values* is quite tight [2,29,31,234]. 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 $\phi^{-1}(\sum_{i=1}^n \beta_i \phi(p_i))$ underlying the *f*-divergences. The Rényi α-divergences (17) correspond to $\phi(u) = u^\alpha$, and this results in the α-mean values $(\sum_{i=1}^n \beta_i p_i^\alpha)^{1/\alpha}$. The second way [38] consists in defining mean values by $\arg\min_v \sum_{i=1}^n \beta_i d(v,u_i)$, 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 $d_h(u,v) = h(u) - h(v) - (u-v)h'(v)$, the corresponding mean values are exactly the above generalized mean values (for $\phi = h'$), which are generally not homogeneous.

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]:

$$\mathcal{S}_\alpha^{(\beta)}(p_1,\ldots,p_n) = \frac{\alpha}{\alpha-1} \ln \int \left(\sum_{i=1}^n \beta_i p_i^\alpha(x)\right)^{1/\alpha} dx \tag{20}$$

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 defined 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

[6] In [18], another α-information is used for channel capacity.

[191]. 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 $KL(\hat{p},p^{\star})$ 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 boostrap 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 α-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:

$$CD_n = KL(p_0,p^{\star}) - KL(p_n,p^{\star}) \tag{21}$$

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:

$$J(\theta) = 1/2\int_{\xi \in \mathbb{R}^n} p_x(\xi) \|\psi_\theta(\xi) - \psi_x(\xi)\|^2 \, d\xi$$

with $\psi_\theta(\xi) \triangleq \nabla_\xi \ln p_\theta(\xi)$ and $\psi_x(\cdot) \triangleq \nabla_\xi \ln p_x(\cdot)$. 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].

8. Miscellanea

Information geometry, which investigates information for statistical inference based on differential geometry, has been studied by Csiszár [71], Rao [51,227], Amari [8,9,13,14] and Kass [143]. A recent book [20] explores neighborhoods of randomness and independence as well as a number of different applications. The tight connections with algebraic statistics are explored in several chapters of the recent book [109]. The role of information geometry in asymptotic statistical theory is discussed in [27]. The information geometric structure of the generalized empirical likelihood method based on minimum divergence estimates is investigated in [204]. A geometric interpretation of conjugate priors, based on MLE and MAP expressed as Bregman barycenters, is provided in [4]. An introduction to information geometry may be found in [12].

Information theoretic *inequalities* are investigated in [82,83,95,110,132,136,171]. Inequalities involving *f*-divergences are studied in [110,119,261].

The *axiomatic characterization* of information and divergence measures is addressed in [3,73,133,179,226,234,268, Chapter 10]; see also [1]. Additional references are provided in [29]. More recent treatments may be found in [76,237,106].

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

[1] J. Aczél, Lectures on Functional Equations and Their Applications, Mathematics in Science and Engineering, vol. 19, Academic Press, 1966.

[2] J. Aczél, Measuring information beyond communication theory—Why some generalized information measures may be useful, others not, Aequationes Mathematicae 27 (1984) 1–19.

[3] J. Aczél, Z. Daròczy, On Measures of Information and Their Characterizations, Mathematics in Science and Engineering, vol. 115, Academic Press, 1975.

[4] A. Agarwal, H. Daume III, A geometric view of conjugate priors, Machine Learning 81 (October (1)) (2010) 99–113.

[5] H. Akaike, A new look at the statistical model identification, IEEE Transactions on Automatic Control 19 (December (6)) (1974) 716–723.

[6] S.M. Ali, S.D. Silvey, A general class of coefficients of divergence of one distribution from another, Journal of the Royal Statistical Society—Series B Methodological 28 (1) (1966) 131–142.

[7] Y. Altun, A. Smola, Unifying divergence minimization and statistical inference via convex duality, in: G. Lugosi, H.U. Simon (Eds.), Proceedings of the 19th Annual Conference on Learning Theory (COLT'06), Pittsburgh, PA, USA, June 22–25, 2006 Lecture Notes in Computer Science, vol. 4005, Springer-Verlag, Berlin, Heidelberg, 2006, pp. 139–153.

[8] S.-I. Amari, Differential–Geometrical Methods in Statistics, Lecture Notes In Statistics, vol. 28, Springer-Verlag, New York, NY, USA, 1985.

[9] S.-I. Amari, Information geometry on hierarchy of probability distributions, IEEE Transactions on Information Theory 47 (July (5)) (2001) 1701–1711.

[10] S.-I. Amari, Integration of stochastic models by minimizing α-divergence, Neural Computation 19 (October (10)) (2007) 2780–2796.

[11] S.-I. Amari, α-divergence is unique belonging to both *f*-divergence and Bregman divergence classes, IEEE Transactions on Information Theory 55 (November (11)) (2009) 4925–4931.

[12] S.-I. Amari, Information geometry and its applications: convex function and dually flat manifold, in: Emerging Trends in Visual Computing - LIX Colloquium, November 2008, Lecture Notes in Computer Science, vol. 5416, Springer-Verlag, 2009, pp. 75–102.

[13] S.-I. Amari, Information geometry derived from divergence functions, in: 3rd International Symposium on Information Geometry and its Applications, Leipzig, FRG, August 2–6, 2010.

[14] S.-I. Amari, H. Nagaoka, Methods of Information Geometry, Translations of Mathematical Monographs, vol. 191, American Mathematical Society, Oxford University Press, Oxford, UK, 2000.

[15] V. Anantharam, A large deviations approach to error exponents in source coding and hypothesis testing, IEEE Transactions on Information Theory 36 (July (4)) (1990) 938–943.

[16] E. Arikan, An inequality on guessing and its application to sequential decoding, IEEE Transactions on Information Theory 42 (January (1)) (1996) 99–105.

[17] S. Arimoto, Information-theoretical considerations on estimation problems, Information and Control 19 (October (3)) (1971) 181–194.

[18] S. Arimoto, Information measures and capacity of order α for discrete memoryless channels, in: Topics in Information Theory—2nd Colloquium, Keszthely, HU, 1975, Colloquia Mathematica Societatis János Bolyai, vol. 16, North Holland, Amsterdam, NL, 1977, pp. 41–52.

[19] V. Arsigny, P. Fillard, X. Pennec, N. Ayache, Geometric means in a novel vector space structure on symmetric positive-definite matrices, SIAM Journal on Matrix Analysis and Applications 29 (1) (2007) 328–347.

[20] K.A. Arwini, C.T.J. Dodson, Information Geometry - Near Randomness and Near Independence, Lecture Notes in Mathematics, vol. 1953, Springer, 2008.

[21] J.A. Aslam, V. Pavlu, Query hardness estimation using Jensen–Shannon divergence among multiple scoring functions, in: G. Amati, C. Carpineto, G. Romano (Eds.), Advances in Information Retrieval—29th European Conference on IR Research, ECIR'07, Rome, Italy, Lecture Notes in Computer Science, vol. 4425, Springer-Verlag, Berlin Heidelberg, FRG, April 2–5, 2007, pp. 198–209.

[22] S. Aviyente, L. Brakel, R. Kushwaha, M. Snodgrass, H. Shevrin, W. Williams, Characterization of event related potentials using information theoretic distance measures, IEEE Transactions on Biomedical Engineering 51 (May (5)) (2004) 737–743.

[23] R.K. Bahr, Asymptotic analysis of error probabilities for the nonzero-mean Gaussian hypothesis testing problem, IEEE Transactions on Information Theory 36 (May (3)) (1990) 597–607.
[24] A. Banerjee, I. Dhillon, J. Ghosh, S. Merugu, An information theoretic analysis of maximum likelihood mixture estimation for exponential families, in: C.E. Brodley (Ed.), Proceedings of the 21st International Conference on Machine Learning (ICML'04), Banff, Alberta, Canada, ACM International Conference Proceeding Series, vol. 69, New York, NY, USA, July 4–8, 2004.
[25] A. Banerjee, I. Dhillon, J. Ghosh, S. Merugu, D.S. Modha, A generalized maximum entropy approach to Bregman co-clustering and matrix approximation, Journal of Machine Learning Research 8 (August) (2007) 1919–1986.
[26] A. Banerjee, S. Merugu, I.S. Dhillon, J. Ghosh, Clustering with Bregman divergences, Journal of Machine Learning Research 6 (October) (2005) 1705–1749.
[27] O.E. Barndorff-Nielsen, D.R. Cox, N. Reid, The role of differential geometry in statistical theory, International Statistical Review 54 (April (1)) (1986) 83–96.
[28] M. Basseville, Distance measures for signal processing and pattern recognition, Signal Processing 18 (December (4)) (1989) 349–369.
[29] M. Basseville, Information: entropies, divergences et moyennes. Research Report 1020, IRISA, ⟨hal.archives-ouvertes.fr/inria-00490399/⟩, May 1996 (in French).
[30] M. Basseville, Information criteria for residual generation and fault detection and isolation, Automatica 33 (May (5)) (1997) 783–803.
[31] M. Basseville, J.-F. Cardoso, On entropies, divergences, and mean values, in: Proceedings of the IEEE International Symposium on Information Theory (ISIT'95), Whistler, British Columbia, Canada, September 1995, p. 330.
[32] A. Basu, I.R. Harris, N. Hjort, M.C. Jones, Robust and efficient estimation by minimising a density power divergence, Biometrika 85 (September (3)) (1998) 549–559.
[33] A. Basu, B.G. Lindsay, Minimum disparity estimation for continuous models: efficiency, distributions and robustness, Annals of the Institute of Statistical Mathematics 46 (December (4)) (1994) 683–705.
[34] A. Basu, B.G. Lindsay, The iteratively reweighted estimating equation in minimum distance problems, Computational Statistics and Data Analysis 45 (March (2)) (2004) 105–124.
[35] A. Basu, H. Shioya, C. Park, Statistical Inference: The Minimum Distance Approach, Chapman & Hall/CRC Monographs on Statistics & Applied Probability, CRC Press, Boca Raton, FL, 2011.
[36] H.H. Bauschke, Duality for Bregman projections onto translated cones and affine subspaces, Journal of Approximation Theory 121 (March (1)) (1983) 1–12.
[37] M. Bekara, L. Knockaert, A.-K. Seghouane, G. Fleury, A model selection approach to signal denoising using Kullback's symmetric divergence, Signal Processing 86 (July (7)) (2006) 1400–1409.
[38] A. Ben-Tal, A. Charnes, M. Teboulle, Entropic means, Journal of Mathematical Analysis and Applications 139 (May (2)) (1989) 537–551.
[39] J.-F. Bercher, On some entropy functionals derived from Rényi information divergence, Information Sciences 178 (June (12)) (2008) 2489–2506.
[40] A. Bhattacharyya, On a measure of divergence between two statistical populations defined by their probability distributions, Bulletin of the Calcutta Mathematical Society 35 (1943) 99–109.
[41] L. Birgé, A new lower bound for multiple hypothesis testing, IEEE Transactions on Information Theory 51 (April (4)) (2005) 1611–1615.
[42] R.E. Blahut, Hypothesis testing and information theory, IEEE Transactions on Information Theory 20 (July (4)) (1974) 405–417.
[43] R.E. Blahut, Principles and Practice of Information Theory, Series in Electrical and Computer Engineering, Addison Wesley Publishing Co, 1987.
[44] J. Boets, K. De Cock, B. De Moor, A mutual information based distance for multivariate Gaussian processes, in: A. Chiuso, A. Ferrante, S. Pinzoni (Eds.), Modeling, Estimation and Control, Festschrift in Honor of Giorgio Picci on the Occasion of his Sixty-Fifth Birthday, Lecture Notes in Control and Information Sciences, vol. 364, Springer-Verlag, Berlin, FRG, October 2007, pp. 15–33.
[45] P. Bougerol, Kalman filtering with random coefficients and contraction, SIAM Journal on Control and Optimization 31 (July (4)) (1993) 942–959.
[46] L.M. Bregman, The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming, USSR Computational Mathematics and Mathematical Physics 7 (3) (1967) 200–217.
[47] M. Broniatowski, A. Keziou, Minimization of φ-divergences on sets of signed measures, Studia Scientiarum Mathematicarum Hungarica 43 (December (4)) (2006) 403–442.
[48] M. Broniatowski, A. Keziou, Parametric estimation and tests through divergences and the duality technique, Journal of Multivariate Analysis 100 (January (1)) (2009) 16–36.
[49] M. Broniatowski, A. Keziou, Divergences and duality for estimation and test under moment condition models, Journal of Statistical Planning and Inference 142 (September (9)) (2012) 2554–2573.
[50] M. Broniatowski, I. Vajda, Several applications of divergence criteria in continuous families. Kybernetika 48, arXiv:0911.0937, in press.
[51] J. Burbea, C.R. Rao, Entropy differential metric, distance and divergence measures in probability spaces: a unified approach, Journal of Multivariate Analysis 12 (December (4)) (1982) 575–596.
[52] J. Burbea, C.R. Rao, On the convexity of higher order Jensen differences based on entropy functions, IEEE Transactions on Information Theory 28 (November (6)) (1982) 961–963.
[53] J. Burbea, C.R. Rao, On the convexity of some divergence measures based on entropy functions, IEEE Transactions on Information Theory 28 (May (3)) (1982) 489–495.
[54] J. Burg, D. Luenberger, D. Wenger, Estimation of structured covariance matrices, Proceedings of the IEEE 70 (September (9)) (1982) 963–974.
[55] C.I. Byrnes, T.T. Georgiou, A. Lindquist, A generalized entropy criterion for Nevanlinna–Pick interpolation with degree constraint, IEEE Transactions on Automatic Control 46 (June (6)) (2001) 822–839.
[56] M.A. Carreira-Perpiñán, G.E. Hinton, On contrastive divergence learning, in: R. Cowell, Z. Ghahramani (Eds.), Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics (AISTATS'05), Barbados, UK, January 6–8, 2005, pp. 59–66.
[57] L. Cayton, Fast nearest neighbor retrieval for Bregman divergences, in: W.W. Cohen, A. McCallum, S.T. Roweis (Eds.), Proceedings of the 25th International Conference on Machine Learning (ICML'08), Helsinki, Finland, June 2008, pp. 112–119.
[58] L. Cayton, Efficient Bregman range search, in: Y. Bengio, D. Schuurmans, J. Lafferty, C.K.I. Williams, A. Culotta (Eds.), Advances in Neural Information Processing Systems 22, Vancouver, British Columbia, Canada, NIPS Foundation, December 7–10, 2009, pp. 243–251.
[59] H. Chernoff, A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations, Annals of Mathematical Statistics 23 (December (4)) (1952) 493–507.
[60] A. Cichocki, S.-I. Amari, Families of alpha- beta- and gamma-divergences: flexible and robust measures of similarities, Entropy 12 (June (6)) (2010) 1532–1568.
[61] A. Cichocki, R. Zdunek, S.-I. Amari, Csiszár's divergences for nonnegative matrix factorization: family of new multiplicative algorithm, in: J.P. Rosca, D. Erdogmus, J.C. Príncipe, S. Haykin (Eds.), Proceedings of the 6th International Conference on Independent Component Analysis and Blind Source Separation (ICA'06), Charleston, South Carolina, USA, Lecture Notes in Computer Science, vol. 3889, Springer-Verlag, Berlin Heidelberg, FRG, March 5–8, 2006, pp. 32–39.
[62] A. Cichocki, R. Zdunek, S.-I. Amari, Nonnegative matrix and tensor factorization, IEEE Signal Processing Magazine 25 (January (1)) (2008) 142–145.
[63] A. Cichocki, R. Zdunek, A. Phan, S.-I. Amari, Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation, John Wiley & Sons Ltd, 2009.
[64] M. Collins, R.E. Schapire, Y. Singer, Logistic regression, AdaBoost and Bregman distances, Machine Learning 48 (July (1–3)) (2002) 253–285.
[65] J. Coursol, D. Dacunha-Castelle, Sur la formule de Chernoff pour deux processus Gaussiens stationnaires, Comptes Rendus Hebdomadaires des Séances de l'Académie des Sciences 288 (May (Section A)) (1979) 769–770. (in French).
[66] T.M. Cover, J.A. Thomas, Elements of Information Theory, Wiley Series in Telecommunications, John Wiley & Sons Ltd, 1991.
[67] T.M. Cover, J.A. Thomas, Elements of Information Theory, second edition, John Wiley & Sons Ltd, aa, 2006.
[68] I. Csiszár, Eine informationstheoretische ungleichung und ihre anwendung auf den beweis der ergodizitat von markoffschen ketten, Magyar Tudományos Akadémia Matematikai Kutató Intezetenek Kozlemenyei 8 (1963) 85–108.

[69] I. Csiszár, Information-type measures of difference of probability distributions and indirect observation, Studia Scientiarum Mathematicarum Hungarica 2 (1967) 299–318.

[70] I. Csiszár, On topological properties of f-divergence, Studia Scientiarum Mathematicarum Hungarica 2 (1967) 329–339.

[71] I. Csiszár, I-divergence geometry of probability distributions and minimization problems, Annals of Probability 3 (Feb. (1)) (1975) 146–158.

[72] I. Csiszár, Information measures: a critical survey, in: J. Kozesnik (Ed.), Transactions of the 7th Conference on Information Theory, Statistical Decision Functions, Random Processes, Prague, vol. B, Academia, Prague, August 18–23, 1974, pp. 73–86.

[73] I. Csiszár, Why least squares and maximum entropy? An axiomatic approach to inference for linear inverse problems, Annals of Statistics 19 (December (4)) (1991) 2032–2066.

[74] I. Csiszár, Generalized cutoff rates and Renyi's information measures, IEEE Transactions on Information Theory 41 (January (1)) (1995) 26–34.

[75] I. Csiszár, Generalized projections for non-negative functions, Acta Mathematica Hungarica 68 (March (1–2)) (1995) 161–186.

[76] I. Csiszár, Axiomatic characterizations of information measures, Entropy 10 (September (3)) (2008) 261–273.

[77] I. Csiszár, F. Matus, Information projections revisited, IEEE Transactions on Information Theory 49 (June (6)) (2003) 1474–1490.

[78] I. Csiszár, F. Matus, On minimization of multivariate entropy functionals, in: V. Anantharam, I. Kontoyiannis (Eds.), Proceedings of the IEEE Information Theory Workshop on Networking and Information Theory (ITW'09), Volos, Greece, June 10–12, 2009, pp. 96–100.

[79] I. Csiszár, F. Matus, Generalized minimizers of convex integral functionals, Bregman distance, Pythagorean identities. ArXiv:1202.0666, February 2012.

[80] M. Das Gupta, T.S. Huang, Bregman distance to l_1 regularized logistic regression. ArXiv:1004.3814, April 2010.

[81] S. Della Pietra, V. Della Pietra, J. Lafferty, Duality and Auxiliary Functions for Bregman Distances, Technical Report Collection CMU-CS-01-109R, School of Computer Science, Carnegie Mellon University, February 2002.

[82] A. Dembo, Information inequalities and concentration of measure, Annals of Probability 25 (April (2)) (1997) 927–939.

[83] A. Dembo, T.M. Cover, J.A. Thomas, Information theoretic inequalities, IEEE Transactions on Information Theory 37 (November (6)) (1991) 1501–1518.

[84] A. Dembo, O. Zeitouni, second edition, Large Deviations Techniques and Applications, Applications of Mathematics, vol. 38, Springer-Verlag, Berlin, Heidelberg, DE, 1998.

[85] L. Devroye, L. Györfi, G. Lugosi, A Probabilistic Theory of Pattern Recognition, Stochastic Modelling and Applied Probability, vol. 31, Springer-Verlag, New York, USA, 1996.

[86] I.S. Dhillon, S. Mallela, R. Kumar, A divisive information-theoretic feature clustering algorithm for text classification, Journal of Machine Learning Research 3 (March) (2003) 1265–1287.

[87] I.S. Dhillon, S. Sra, Generalized nonnegative matrix approximations with Bregman divergences, in: Y. Weiss, B. Schölkopf, J. Platt (Eds.), Advances in Neural Information Processing Systems 18, Vancouver, British Columbia, Canada, December 5–8, 2005, MIT Press, Cambridge, MA, 2006, pp. 283–290.

[88] I.S. Dhillon, J.A. Tropp, Matrix nearness problems with Bregman divergences, SIAM Journal on Matrix Analysis and Applications 29 (4) (2008) 1120–1146.

[89] D. Donoho, V. Stodden, When does non-negative matrix factorization give a correct decomposition into parts?, in: S. Thrun, L. Saul, B. Schölkopf (Eds.), Advances in Neural Information Processing Systems 16, Vancouver, British Columbia, Canada, December 8–13, 2003, MIT Press, Cambridge, MA, 2004.

[90] M.D. Donsker, S. Varadhan, Asymptotic evaluation of certain Markov process expectations for large time, II, Communications on Pure and Applied Mathematics 28 (March (2)) (1975) 279–301.

[91] I.L. Dryden, A. Kolydenko, D. Zhou, Non-Euclidean statistics for covariance matrices, with applications to diffusion tensor imaging, Annals of Applied Statistics 3 (September (3)) (2009) 1102–1123.

[92] S. Eguchi, S. Kato, Entropy and divergence associated with power function and the statistical application, Entropy 12 (February (2)) (2010) 262–274.

[93] D.M. Endres, J.E. Schindelin, A new metric for probability distributions, IEEE Transactions on Information Theory 49 (July (7)) (2003) 1858–1860.

[94] M.D. Esteban, A general class of entropy statistics, Applications of Mathematics 42 (June (3)) (1997) 161–169.

[95] A.A. Fedotov, P. Harremoes, F. Topsoe, Refinements of Pinsker's inequality, IEEE Transactions on Information Theory 49 (June (6)) (2003) 1491–1498.

[96] A. Ferrante, M. Pavon, F. Ramponi, Hellinger versus Kullback–Leibler multivariable spectrum approximation, IEEE Transactions on Automatic Control 53 (May (4)) (2008) 954–967.

[97] D. Ferrari, Y. Yang, Maximum L_q-likelihood estimation, Annals of Statistics 38 (April (2)) (2010) 753–783.

[98] L. Finesso, P. Spreij, Nonnegative matrix factorization and I-divergence alternating minimization, Linear Algebra and its Applications 416 (July (2–3)) (2006) 270–287.

[99] A. Fischer, Quantization and clustering with Bregman divergences, Journal of Multivariate Analysis 101 (October (9)) (2010) 2207–2221.

[100] B.A. Frigyik, S. Srivastava, M.R. Gupta, Functional Bregman divergence and Bayesian estimation of distributions, IEEE Transactions on Information Theory 54 (November (11)) (2008) 5130–5139.

[101] Y. Fujimoto, N. Murata, A modified EM algorithm for mixture models based on Bregman divergence, Annals of the Institute of Statistical Mathematics 59 (March (1)) (2007) 3–25.

[102] C. Fé, N. Bertin, J.-L. Durrieu, Nonnegative matrix factorization with the Itakura–Saito divergence. With application to music analysis, Neural Computation 21 (March (3)) (2009) 793–830.

[103] C. Févotte, J. Idier, Algorithms for nonnegative matrix factorization with the β-divergence, Neural Computation 23 (September (9)) (2011) 2421–2456.

[104] T.T. Georgiou, Relative entropy and the multivariable multidimensional moment problem, IEEE Transactions on Information Theory 52 (March (3)) (2006) 1052–1066.

[105] T.T. Georgiou, Distances and Riemannian metrics for spectral density functions, IEEE Transactions on Signal Processing 55 (August (8)) (2007) 3995–4003.

[106] T.T. Georgiou, J. Karlsson, M.S. Takyar, Metrics for power spectra: an axiomatic approach, IEEE Transactions on Signal Processing 57 (March (3)) (2009) 859–867.

[107] T.T. Georgiou, A. Lindquist, Kullback–Leibler approximation of spectral density functions, IEEE Transactions on Information Theory 49 (November (11)) (2003) 2910–2917.

[108] T.T. Georgiou, A. Lindquist, A convex optimization approach to ARMA modeling, IEEE Transactions on Automatic Control 53 (June (5)) (2008) 1108–1119.

[109] P. Gibilisco, E. Riccomagno, M.P. Rogantin, H.P. Wynn (Eds.), Algebraic and Geometric Methods in Statistics, Cambridge University Press, Cambridge, UK, 2010.

[110] G.L. Gilardoni, On Pinsker's and Vajda's type inequalities for Csiszár's f-divergences, IEEE Transactions on Information Theory 56 (November (11)) (2010) 5377–5386.

[111] A.H.J. Gray, J.D. Markel, Distance measures for speech processing, IEEE Transactions on Acoustics, Speech, and Signal Processing 24 (October (5)) (1976) 380–391.

[112] R.M. Gray, Entropy and Information Theory, Springer-Verlag, New York, NY, USA, 1990, online corrected version, 2009, ⟨http://ee.stanford.edu/gray/it.html⟩.

[113] R.M. Gray, Entropy and Information Theory, second edition, Springer-Verlag, New York, NY, USA, 2010.

[114] R.M. Gray, A. Buzo, A.H.J. Gray, Y. Matsuyama, Distortion measures for speech processing, IEEE Transactions on Acoustics, Speech, and Signal Processing 28 (August (4)) (1980) 367–376.

[115] P.D. Grünwald, A.P. Dawid, Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory, Annals of Statistics 32 (August (4)) (2004) 1367–1433.

[116] A. Guntuboyina, Lower bounds for the minimax risk using f-divergences and applications, IEEE Transactions on Information Theory 57 (April (4)) (2011) 2386–2399.

[117] L. Györfi, T. Nemetz, f-Dissimilarity: a generalization of the affinity of several distributions, Annals of the Institute of Statistical Mathematics 30 (December (1)) (1978) 105–113.

[118] P. Harremoës, I. Vajda, On Bahadur efficiency of power divergence statistics. ArXiv:1002.1493, February 2010.

[119] P. Harremoës, I. Vajda, On pairs of f-divergences and their joint range, IEEE Transactions on Information Theory 57 (June (6)) (2011) 3230–3235.

[120] P. Harremoës, C. Vignat, Rényi entropies of projections, in: A. Barg, R.W. Yeung (Eds.), Proceedings of the IEEE International Symposium on Information Theory (ISIT'06), Seattle, WA, USA, July 9–14, 2006, pp. 1827–1830.

[121] M.E. Havrda, F. Charvát, Quantification method of classification processes: concept of structural α-entropy, Kybernetika 3 (1967) 30–35.

[122] Y. He, A.B. Hamza, H. Krim, A generalized divergence measure for robust image registration, IEEE Transactions on Signal Processing 51 (May (5)) (2003) 1211–1220.

[123] A.O. Hero, B. Ma, O. Michel, J. Gorman, Alpha-divergence for Classification, Indexing and Retrieval, Research Report CSPL-328, University of Michigan, Communications and Signal Processing Laboratory, May 2001.

[124] G.E. Hinton, Training products of experts by minimizing contrastive divergence, Neural Computation 14 (August (8)) (2002) 1771–1800.

[125] G.E. Hinton, S. Osindero, Y.-W. Teh, A fast learning algorithm for deep belief nets, Neural Computation 18 (July (7)) (2006) 1527–1554.

[126] W. Hoeffding, Asymptotically optimal tests for multinomial distributions, Annals of Mathematical Statistics 36 (April (2)) (1965) 369–401.

[127] A. Hyvárinen, Estimation of non-normalized statistical models by score matching, Journal of Machine Learning Research 6 (April) (2005) 695–709.

[128] A. Hyvárinen, Some extensions of score matching, Computational Statistics and Data Analysis 51 (February (5)) (2007) 2499–2512.

[129] W. James, C. Stein, Estimation with quadratic loss, in: J. Neyman (Ed.), Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, University of California Press, Berkeley, CA, USA, 1961, pp. 361–379.

[130] X. Jiang, Z.-Q. Luo, T.T. Georgiou, Geometric methods for spectral analysis, IEEE Transactions on Signal Processing 60 (March (3)) (2012) 1064–1074.

[131] X. Jiang, L. Ning, T.T. Georgiou, Distances and Riemannian metrics for multivariate spectral densities, IEEE Transactions on Automatic Control 57 (July (7)) (2012) 1723–1735.

[132] O. Johnson, A. Barron, Fisher information inequalities and the central limit theorem, Probability Theory and Related Fields 129 (July (3)) (2004) 391–409.

[133] R. Johnson, Axiomatic characterization of the directed divergences and their linear combinations, IEEE Transactions on Information Theory 25 (November (6)) (1979) 709–716.

[134] L.K. Jones, C.L. Byrne, General entropy criteria for inverse problems, with applications to data compression, pattern classification, and cluster analysis, IEEE Transactions on Information Theory 36 (January (1)) (1990) 23–30.

[135] M.C. Jones, N. Hjort, I.R. Harris, A. Basu, A comparison of related density-based minimum divergence estimators, Biometrika 88 (October (3)) (2001) 865–873.

[136] A. Kagan, T. Yu, Some inequalities related to the Stam inequality, Applications of Mathematics 53 (3) (2008) 195–205.

[137] T. Kanamori, A. Ohara, A Bregman extension of quasi-Newton updates II: convergence and robustness properties. ArXiv:1010.2846, October 2010.

[138] T. Kanamori, A. Ohara, A Bregman extension of quasi-Newton updates I: an information geometrical framework, Optimization Methods and Software 27, http://dx.doi.org/10.1080/10556788.2011.613073, in press.

[139] T. Kanamori, T. Suzuki, M. Sugiyama, *f*-divergence estimation and two-sample homogeneity test under semiparametric density-ratio models, IEEE Transactions on Information Theory 58 (February (2)) (2012) 708–720.

[140] A. Karagrigoriou, K. Mattheou, Measures of divergence in model selection, in: C.H. Skiadas (Ed.), Advances in Data Analysis. Statistics for Industry and Technology, Birkhäuser, Boston, MA, 2010, pp. 51–65.

[141] A. Karagrigoriou, T. Papaioannou, On measures of information and divergence and model selection criteria, in: F. Vonta, M. Nikulin, N. Limnios, C. Huber-Carol (Eds.), Statistical Models and Methods for Biomedical and Technical Systems. Statistics for Industry and Technology, Birkhäuser, Boston, MA, USA, 2008, pp. 503–518.

[142] J. Karlsson, T.T. Georgiou, A.G. Lindquist, The inverse problem of analytic interpolation with degree constraint and weight selection for control synthesis, IEEE Transactions on Automatic Control 55 (February (2)) (2010) 405–418.

[143] R.E. Kass, P.W. Vos, Geometrical Foundations of Asymptotic Inference, Series in Probability and Statistics, Wiley, 1997.

[144] D. Kazakos, On resolution and exponential discrimination between Gaussian stationary vector processes and dynamic models, IEEE Transactions on Automatic Control 25 (April (2)) (1980) 294–296.

[145] D. Kazakos, Spectral distance measures between continuous-time vector Gaussian processes, IEEE Transactions on Information Theory 28 (July (4)) (1982) 679–684.

[146] D. Kazakos, P. Papantoni-Kazakos, Spectral distance measures between Gaussian processes, IEEE Transactions on Automatic Control 25 (October (5)) (1980) 950–959.

[147] D. Kazakos, P. Papantoni-Kazakos, Detection and Estimation, Computer Science Press, 1990.

[148] M. Kim, S. Lee, Estimation of a tail index based on minimum density power divergence, Journal of Multivariate Analysis 99 (November (10)) (2008) 2453–2471.

[149] J. Kivinen, M.K. Warmuth, Boosting as entropy projection, in: Proceedings of the 12th Annual Conference on Computational Learning Theory (COLT'99), Santa Cruz, CA, USA, ACM, July 7–9, 1999, pp. 134–144.

[150] J. Kivinen, M.K. Warmuth, B. Hassibi, The *p*-norm generalization of the LMS algorithm for adaptive filtering, IEEE Transactions on Signal Processing 54 (May (5)) (2006) 1782–1793.

[151] L. Knockaert, A class of statistical and spectral distance measures based on Bose–Einstein statistics, IEEE Transactions on Signal Processing 41 (November (11)) (1993) 3171–3174.

[152] L. Knockaert, Statistical thermodynamics and natural *f*-divergences. unpublished paper ⟨users.ugent.be/lknockae/⟩, 1994.

[153] L. Knockaert, On scale and concentration invariance in entropies, Information Sciences 152 (June) (2003) 139–144.

[154] R. Kompass, A generalized divergence measure for nonnegative matrix factorization, Neural Computation 19 (March (3)) (2007) 780–791.

[155] B. Kulis, M.A. Sustik, I.S. Dhillon, Low-rank kernel learning with Bregman matrix divergences, Journal of Machine Learning Research 10 (February) (2009) 341–376.

[156] S. Kullback, J.C. Keegel, J.H. Kullback, Topics in Statistical Information Theory, Lecture Notes in Statistics, vol. 42, Springer-Verlag, New York, NY, USA, 1987.

[157] J.D. Lafferty, Statistical learning algorithms based on Bregman distances, in: Proceedings of the Canadian Workshop on Information Theory, Toronto, Canada, June 3–6, 1997, pp. 77–80.

[158] J.D. Lafferty, Additive models, boosting, and inference for generalized divergences, in: Proceedings of the 12th Annual Conference on Computational Learning Theory (COLT'99), Santa Cruz, CA, USA, ACM, July 7–9, 1999, pp. 125–133.

[159] J. Lawson, Y. Lim, A Birkhoff contraction formula with application to Riccati equations, SIAM Journal on Control and Optimization 46 (June (3)) (2007) 930–951.

[160] G. Le Besnerais, J.-F. Bercher, G. Demoment, A new look at entropy for solving linear inverse problems, IEEE Transactions on Information Theory 45 (July (5)) (1999) 1565–1578.

[161] G. Lebanon, J. Lafferty, Boosting and maximum likelihood for exponential models, in: T.G. Dietterich, S. Becker, Z. Ghahramani (Eds.), Advances in Neural Information Processing Systems 14, Vancouver, British Columbia, Canada, MIT Press, Cambridge, MA, December 3–8, 2001.

[162] H. Lee, H. Lim, Invariant metrics, contractions and nonlinear matrix equations, Nonlinearity 21 (April (4)) (2008) 857–878.

[163] A. Lefevre, F. Bach, C. Fevotte, Online algorithms for nonnegative matrix factorization with the Itakura–Saito divergence, in: Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'11), New Paltz, NY, USA, October 16–19, 2011, pp. 313–316.

[164] N. Leonenko, O. Seleznjev, Statistical inference for the ϵ-entropy and the quadratic Rényi entropy, Journal of Multivariate Analysis 101 (October (9)) (2010) 1981–1994.

[165] B. Levy, R. Nikoukhah, Robust least-squares estimation with a relative entropy constraint, IEEE Transactions on Information Theory 50 (January (1)) (2004) 89–104.

[166] K. Li, W. Zhou, S. Yu, Effective metric for detecting distributed denial-of-service attacks based on information divergence, IET Communications 3 (December (12)) (2009) 1851–1860.

[167] F. Liese, I. Vajda, Convex Statistical Distances, Texte zur Mathematick, vol. 95, Teubner, Leipzig, 1987.

[168] F. Liese, I. Vajda, On divergences and informations in statistics and information theory, IEEE Transactions on Information Theory 52 (October (10)) (2006) 4394–4412.

[169] J. Lin, Divergence measures based on the Shannon entropy, IEEE Transactions on Information Theory 37 (January (1)) (1991) 145–151.

[170] B.G. Lindsay, Efficiency versus robustness: the case for minimum Hellinger distance and related methods, Annals of Statistics 22 (June (2)) (1994) 1081–1114.

[171] E. Lutwak, D. Yang, G. Zhang, Cramér–Rao and moment-entropy inequalities for Rényi entropy and generalized Fisher information,

IEEE Transactions on Information Theory 51 (February (2)) (2005) 473–478.
[172] S. Ma, D. Goldfarb, L. Chen, Fixed point and Bregman iterative methods for matrix rank minimization, Mathematical Programming, Series A 128 (June (1–2)) (2011) 321–353.
[173] D. MacKay, Information Theory, Inference & Learning Algorithms, Cambridge University Press, Cambridge, UK, 2003.
[174] P. Maji, f-Information measures for efficient selection of discriminative genes from microarray data, IEEE Transactions on Biomedical Engineering 56 (April (4)) (2009) 1063–1069.
[175] P. Maji, S.K. Pal, Feature selection using f-information measures in fuzzy approximation spaces, IEEE Transactions on Knowledge and Data Engineering 22 (June (6)) (2010) 854–867.
[176] P. Mantalos, K. Mattheou, A. Karagrigoriou, An improved divergence information criterion for the determination of the order of an AR process, Communications in Statistics—Simulation and Computation 39 (5) (2010) 865–879.
[177] M. Markatou, A. Basu, B.G. Lindsay, Weighted likelihood equations with bootstrap root search, Journal of the American Statistical Association 93 (June (442)) (1998) 740–750.
[178] N. Martín, Y. Li, A new class of minimum power divergence estimators with applications to cancer surveillance, Journal of Multivariate Analysis 102 (September (8)) (2011) 1175–1194.
[179] A. Mathai, P. Rathie, Basic Concepts in Information Theory and Statistics, Wiley Eastern Ltd, New Delhi, India, 1975.
[180] Y. Matsuyama, Non-logarithmic information measures, α-weighted EM algorithms and speedup of learning, in: Proceedings of the IEEE International Symposium on Information Theory (ISIT'98), Cambridge, MA, USA, August 16–21, 1998, p. 385.
[181] Y. Matsuyama, The α-EM algorithm and its applications, in: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'00), vol. 1, Istanbul, Turkey, June 5–9, 2000, pp. 592–595.
[182] Y. Matsuyama, The α-EM algorithm: surrogate likelihood maximization using alpha-logarithmic information measures, IEEE Transactions on Information Theory 49 (March (3)) (2003) 692–706.
[183] Y. Matsuyama, N. Katsumata, S. Imahara, Convex divergence as a surrogate function for independence: the f-divergence, in: T.-W, Lee, T.-P. Jung, S. Makeig, T.J. Sejnowski, (Eds.), Proceedings of the 3rd International Conference on Independent Component Analysis and Blind Signal Separation, San Diego, CA, USA, December 2001, pp. 31–36.
[184] K. Mattheou, S. Lee, A. Karagrigoriou, A model selection criterion based on the BHHJ measure of divergence, Journal of Statistical Planning and Inference 139 (February (2)) (2009) 228–235.
[185] F. Matus, Divergence from factorizable distributions and matroid representations by partitions, IEEE Transactions on Information Theory 55 (December (12)) (2009) 5375–5381.
[186] K. Matusita, Discrimination and the affinity of distributions, in: T. Cacoullos (Ed.), Discriminant Analysis and Applications, Academic Press, New York, USA, 1973, pp. 213–223.
[187] N. Merhav, Data processing theorems and the second law of thermodynamics, IEEE Transactions on Information Theory 57 (August (8)) (2011) 4926–4939.
[188] M. Minami, S. Eguchi, Robust blind source separation by beta divergence, Neural Computation 14 (August (8)) (2002) 1859–1886.
[189] T. Minka, Divergence Measures and Message Passing, Technical Report MSR-TR-2005-173, Microsoft Research Ltd, 2005.
[190] A. Mnih, G. Hinton, Learning nonlinear constraints with contrastive backpropagation, in: D.V. Prokhorov (Ed.), Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN'05), vol. 2, Montréal, Québec, Canada, July 31–August 4, 2005, pp. 1302–1307.
[191] M. Moakher, P.G. Batchelor, Symmetric positive-definite matrices: from geometry to applications and visualization, in: J. Weickert, H. Hagen (Eds.), Visualization and Processing of Tensor Fields Mathematics—Visualization, vol. 17, Springer-Verlag, 2006, pp. 285–298.
[192] M.N.H. Mollah, M. Minami, S. Eguchi, Exploring latent structure of mixture ICA models by the minimum β-divergence method, Neural Computation 18 (January (1)) (2006) 166–190.
[193] T. Morimoto, Markov processes and the H-theorem, Journal of the Physical Society of Japan 18 (March (3)) (1963) 328–331.
[194] N. Murata, T. Takenouchi, T. Kanamori, S. Eguchi, Information geometry of U-Boost and Bregman divergence, Neural Computation 16 (July (7)) (2004) 1437–1481.
[195] A. Nascimento, R. Cintra, A. Frery, Hypothesis testing in speckled data with stochastic distances, IEEE Transactions on Geoscience and Remote Sensing 48 (January (1)) (2010) 373–385.
[196] G. Nason, Robust projection indices, Journal of the Royal Statistical Society—Series B Methodological 63 (3) (2001) 551–567.
[197] S. Natarajan, Large deviations, hypotheses testing, and source coding for finite Markov chains, IEEE Transactions on Information Theory 31 (May (3)) (1985) 360–365.
[198] P. Nath, On a coding theorem connected with Rényi's entropy, Information and Control 29 (November (3)) (1975) 234–242.
[199] X. Nguyen, M.J. Wainwright, M.I. Jordan, On surrogate loss functions and f-divergences, Annals of Statistics 37 (April (2)) (2009) 876–904.
[200] X. Nguyen, M.J. Wainwright, M.I. Jordan, Estimating divergence functionals and the likelihood ratio by convex risk minimization, IEEE Transactions on Information Theory 56 (November (11)) (2010) 5847–5861.
[201] F. Nielsen, S. Boltz, The Burbea–Rao and Bhattacharyya centroids, IEEE Transactions on Information Theory 57 (August (8)) (2011) 5455–5466.
[202] F. Nielsen, R. Nock, Sided and symmetrized Bregman centroids, IEEE Transactions on Information Theory 55 (June (6)) (2009) 2882–2904.
[203] F. Nielsen, P. Piro, M. Barlaud, Bregman vantage point trees for efficient nearest neighbor queries, in: Q. Sun, Y. Rui (Eds.), Proceedings of the IEEE International Conference on Multimedia and Expo (ICME'09), New York, NY, USA, June 28–July 3, 2009, pp. 878–881.
[204] T. Nishimura, F. Komaki, The information geometric structure of generalized empirical likelihood estimators, Communications in Statistics—Theory and Methods 37 (January (12)) (2008) 1867–1879.
[205] R. Nock, F. Nielsen, Bregman divergences and surrogates for learning, IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (November (11)) (2009) 2048–2059.
[206] L. Pardo, Statistical Inference Based on Divergence Measures, Chapman & Hall/CRC Monographs on Statistics & Applied Probability, CRC Press, Taylor & Francis Group, Boca Raton, FL, 2006.
[207] L. Pardo, M. Salicrú, M.L. Menéndez, D. Morales, Divergence measures based on entropy functions and statistical inference, Sankhyā: The Indian Journal of Statistics—Series B 57 (December (3)) (1995) 315–337.
[208] M. Pardo, I. Vajda, On asymptotic properties of information-theoretic divergences, IEEE Transactions on Information Theory 49 (July (7)) (2003) 1860–1867.
[209] R.K. Patra, A. Mandal, A. Basu, Minimum Hellinger distance estimation with inlier modification, Sankhyā: The Indian Journal of Statistics—Series B 70 (November (2)) (2008) 310–323.
[210] M. Pavon, A. Ferrante, On the Georgiou–Lindquist approach to constrained Kullback–Leibler approximation of spectral densities, IEEE Transactions on Automatic Control 51 (April (4)) (2006) 639–644.
[211] M. Pavon, A. Ferrante, On the geometry of maximum entropy problems. ArXiv:1112.5529, December 2011.
[212] B. Pelletier, Informative barycentres in statistics, Annals of the Institute of Statistical Mathematics 57 (December (4)) (2005) 767–780.
[213] B. Pelletier, Inference in ϕ-families of distributions, Statistics—A Journal of Theoretical and Applied Statistics 45 (3) (2011) 223–237.
[214] A. Perez, Barycenter of a set of probability measures and its application in statistical decision, in: T. Havranek, Z. Sidak, M. Novak (Eds.), Proceedings of the 6th Prague Symposium on Computational Statistics, COMPSTAT'84, Physica-Verlag, Wien, AUS, 1984, pp. 154–159.
[215] D. Petz, Monotone metrics on matrix spaces, Linear Algebra and its Applications 244 (September) (1996) 81–96.
[216] D. Petz, R. Temesi, Means of positive numbers and matrices, SIAM Journal on Matrix Analysis and Applications 27 (3) (2005) 712–720.
[217] D.-T. Pham, F. Vrins, M. Verleysen, On the risk of using Rényi's entropy for blind source separation, IEEE Transactions on Signal Processing 56 (October (10)) (2008) 4611–4620.
[218] J.P.W. Pluim, J.B.A. Maintz, M.A. Viergever, f-Information measures in medical image registration, IEEE Transactions on Medical Imaging 23 (December (12)) (2004) 1508–1516.
[219] B. Poczos, L. Xiong, J. Schneider, Nonparametric divergence estimation with applications to machine learning on distributions. ArXiv:1202.3758, February 2012.

[220] J.C. Principe, Information Theoretic Learning: Renyi's Entropy and Kernel Perspectives, Information Science and Statistics, Springer-Verlag, 2008.
[221] Y. Qiao, N. Minematsu, A study on invariance of f-divergence and its application to speech recognition, IEEE Transactions on Signal Processing 58 (July (7)) (2010) 3884–3890.
[222] F. Ramponi, A. Ferrante, M. Pavon, A globally convergent matricial algorithm for multivariate spectral estimation, IEEE Transactions on Automatic Control 54 (October (10)) (2009) 2376–2388.
[223] C.R. Rao, Information and accuracy attainable in the estimation of statistical parameters, Bulletin of the Calcutta Mathematical Society 37 (1945) 81–91.
[224] C.R. Rao, Diversity and dissimilarity coefficients: a unified approach, Theoretical Population Biology 21 (February (1)) (1982) 24–43.
[225] C.R. Rao, Diversity: its measurement, decomposition, apportionment and analysis, Sankhyā: The Indian Journal of Statistics—Series A 44 (1) (1982) 1–22.
[226] C.R. Rao, Rao's axiomatization of diversity measures, in: S. Kotz, N.L. Johnson (Eds.), Encyclopedia of Statistical Sciences, vol. 7, John Wiley & Sons Ltd, 1986, pp. 614–617.
[227] C.R. Rao, Differential metrics in probability spaces, in: S.-I. Amari, O.E. Barndorff-Nielsen, R.E. Kass, S.L. Lauritzen, C.R. Rao (Eds.), Differential Geometry in Statistical Inference, Lecture Notes—Monograph Series, vol. 10, Institute of Mathematical Statistics, Hayward, CA, USA, 1987, pp. 217–240.
[228] C.R. Rao, T. Nayak, Cross entropy, dissimilarity measures, and characterizations of quadratic entropy, IEEE Transactions on Information Theory 31 (September (5)) (1985) 589–593.
[229] J. Rauh, Finding the maximizers of the information divergence from an exponential family, IEEE Transactions on Information Theory 57 (June (6)) (2011) 3236–3247.
[230] P. Ravikumar, A. Agarwal, M.J. Wainwright, Message-passing for graph-structured linear programs: proximal methods and rounding schemes, Journal of Machine Learning Research 11 (March) (2010) 1043–1080.
[231] T. Read, N. Cressie, Goodness-of-Fit Statistics for Discrete Multivariate Data. Statistics, Springer, NY, 1988.
[232] M.D. Reid, R.C. Williamson, Composite binary losses, Journal of Machine Learning Research 11 (September) (2010) 2387–2422.
[233] M.D. Reid, R.C. Williamson, Information, divergence and risk for binary experiments, Journal of Machine Learning Research 12 (March) (2011) 731–817.
[234] A. Rényi, On measures of information and entropy, in: J. Neyman (Ed.), Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, University of California Press, Berkeley, CA, USA, 1961, pp. 547–561.
[235] A. Rényi, On some basic problems of statistics from the point of view of information theory, in: L.M. Le Cam, J. Neyman (Eds.), Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, University of California Press, Berkeley, CA, USA, 1967, pp. 531–543.
[236] A. Roman, S. Jolad, M.C. Shastry, Bounded divergence measures based on Bhattacharyya coefficient. ArXiv:1201.0418, January 2012.
[237] W. Sander, Measures of information, in: E. Pap (Ed.), Handbook of Measure Theory, vol. 2, North-Holland, Amsterdam, NL, 2002, pp. 1523–1565.
[238] R. Santos-Rodriguez, D. Garcia-Garcia, J. Cid-Sueiro, Cost-sensitive classification based on Bregman divergences for medical diagnosis, in: M.A. Wani (Ed.), Proceedings of the 8th International Conference on Machine Learning and Applications (ICMLA'09), Miami Beach, FL, USA, December 13–15, 2009, pp. 551–556.
[239] M.P. Schützenberger, Contribution aux applications statistiques de la théorie de l'information. Thèse d'État, Inst. Stat. Univ. Paris, 1953 (in French).
[240] F.C. Schweppe, On the Bhattacharyya distance and the divergence between Gaussian processes, Information and Control 11 (October (4)) (1967) 373–395.
[241] F.C. Schweppe, State space evaluation of the Bhattacharyya distance between two Gaussian processes, Information and Control 11 (September (3)) (1967) 352–372.
[242] J. Shore, R. Johnson, Properties of cross-entropy minimization, IEEE Transactions on Information Theory 27 (July (4)) (1981) 472–482.
[243] J.E. Shore, R.M. Gray, Minimum cross-entropy pattern classification and cluster analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence 4 (January (1)) (1982) 11–17.
[244] S. Si, D. Tao, B. Geng, Bregman divergence-based regularization for transfer subspace learning, IEEE Transactions on Knowledge and Data Engineering 22 (July (7)) (2010) 929–942.
[245] R. Sibson, Information radius, Probability Theory and Related Fields 14 (June (2)) (1969) 149–160.
[246] B.K. Sriperumbudur, A. Gretton, K. Fukumizu, G.R.G. Lanckriet, B. Schölkopf, On integral probability metrics, ϕ-divergences and binary classification. ArXiv:0901.2698, January 2009.
[247] S. Srivastava, M.R. Gupta, B.A. Frigyik, Bayesian quadratic discriminant analysis, Journal of Machine Learning Research 8 (June) (2007) 1277–1305.
[248] F. Österreicher, I. Vajda, A new class of metric divergences on probability spaces and its applicability in statistics, Annals of the Institute of Statistical Mathematics 55 (September (3)) (2003) 639–653.
[249] A.A. Stoorvogel, J.H. van Schuppen, Approximation problems with the divergence criterion for Gaussian variables and Gaussian processes, Systems and Control Letters 35 (November (4)) (1998) 207–218.
[250] W. Stummer, I. Vajda, On divergences of finite measures and their applicability in statistics and information theory, Statistics—A Journal of Theoretical and Applied Statistics 44 (April (2)) (2010) 169–187.
[251] W. Stummer, I. Vajda, On Bregman distances and divergences of probability measures, IEEE Transactions on Information Theory 58 (March (3)) (2012) 1277–1288.
[252] M. Sugiyama, T. Suzuki, T. Kanamori, Density-ratio matching under the Bregman divergence: a unified framework of density-ratio estimation. Annals of the Institute of Statistical Mathematics 64 (2) (2012), 1009–1044.
[253] Y. Sung, L. Tong, H.V. Poor, Neyman–Pearson detection of Gauss–Markov signals in noise: closed-form error exponent and properties, IEEE Transactions on Information Theory 52 (April (4)) (2006) 1354–1365.
[254] I. Sutskever, T. Tieleman, On the convergence properties of contrastive divergence, in: Y.W. Teh, M. Titterington (Eds.), Proceedings of the 13th International Workshop on Artificial Intelligence and Statistics (AISTATS'10), Chia Laguna, Sardinia, Italy, May 13–15, 2010, pp. 78–795.
[255] I.J. Taneja, On generalized information measures and their applications, Advances in Electronics and Electron Physics 76 (1989) 327–413.
[256] I.J. Taneja, Generalized Information Measures and Their Applications. ⟨www.mtm.ufsc.br/taneja/book/book.html⟩, 2001.
[257] B. Taskar, S. Lacoste-Julien, M.I. Jordan, Structured prediction, dual extragradient and Bregman projections, Journal of Machine Learning Research 7 (July) (2006) 1627–1653.
[258] M. Teboulle, A unified continuous optimization framework for center-based clustering methods, Journal of Machine Learning Research 8 (January) (2007) 65–102.
[259] M. Teboulle, P. Berkhin, I.S. Dhillon, Y. Guan, J. Kogan, Clustering with entropy-like k-means algorithms, in: J. Kogan, C. Nicholas, M. Teboulle (Eds.), Grouping Multidimensional Data—Recent Advances in Clustering, Springer-Verlag, Berlin, Heidelberg, 2006, pp. 127–160.
[260] A. Toma, M. Broniatowski, Dual divergence estimators and tests: robustness results, Journal of Multivariate Analysis 102 (January (1)) (2011) 20–36.
[261] F. Topsoe, Some inequalities for information divergence and related measures of discrimination, IEEE Transactions on Information Theory 46 (July (4)) (2000) 1602–1609.
[262] E. Torgersen, Comparison of Statistical Experiments, Encyclopedia of Mathematics and its Applications, vol. 36, Cambridge University Press, Cambridge, UK, 1991.
[263] J. Touboul, Projection pursuit through minimisation ϕ-divergence, Entropy 12 (June (6)) (2010) 1581–1611.
[264] K. Tsuda, G. Rätsch, M.K. Warmuth, Matrix exponentiated gradient updates for on-line learning and Bregman projection, Journal of Machine Learning Research 6 (June) (2005) 995–1018.
[265] M. Tsukada, H. Suyari, Tsallis differential entropy and divergences derived from the generalized Shannon–Khinchin axioms, in: Proceedings of the IEEE International Symposium on Information Theory (ISIT'09), Seoul, Korea, June 28–July 3, 2009, pp. 149–153.
[266] J. Vachery, A. Dukkipati, On Shore and Johnson properties for a special case of Csiszár f-divergences. ArXiv:1201.4285, January 2012.
[267] I. Vajda, χ^{α}-divergence and generalized Fisher's information, in: J. Kozesnik (Ed.), Transactions of the 6th Prague Conference on Information Theory, Statistical Decision Functions, Random Processes, Prague, September 19–25, 1971, Academia, Prague, 1973, pp. 873–886.

[268] I. Vajda, Theory of Statistical Inference and Information, Series B: Mathematical and Statistical Methods, vol. 11, Kluwer Academic Publishers, Dordrecht, 1989.
[269] I. Vajda, Modifications of Divergence Criteria for Applications in Continuous Families, Research Report 2230, Academy of Sciences of the Czech Republic, Institute of Information Theory and Automation, November 2008.
[270] I. Vajda, On metric divergences of probability measures, Kybernetika 45 (6) (2009) 885–900.
[271] B. Vemuri, M. Liu, S. Amari, F. Nielsen, Total Bregman divergence and its applications to DTI analysis, IEEE Transactions on Medical Imaging 30 (February (2)) (2011) 475–483.
[272] C. Vignat, A.O. Hero, J.A. Costa, A geometric characterization of maximum Rényi entropy distributions, in: Proceedings of the IEEE International Symposium on Information Theory (ISIT'06), Seattle, Washington, USA, July 2006, pp. 1822–1826.
[273] F. Vrins, D.-T. Pham, M. Verleysen, Is the general form of Renyi's entropy a contrast for source separation?, in: M.E. Davies, C.J. James, S.A. Abdallah, M.D. Plumbley (Eds.), Proceedings of the 7th International Conference on Independent Component Analysis and Blind Source Separation (ICA'07), London, UK, Lecture Notes in Computer Science, , Lecture Notes in Computer Science, vol. 4666, September 9–12, 2007, Springer-Verlag, Berlin, Heidelberg, FRG, 2007, pp. 129–136.
[274] Q. Wang, S.R. Kulkarni, S. Verdu, Divergence estimation for multidimensional densities via k-nearest-neighbor distances, IEEE Transactions on Information Theory 55 (May (5)) (2009) 2392–2405.
[275] S. Wang, D. Schuurmans, Learning continuous latent variable models with Bregman divergences, in: R. Gavaldà, K.P. Jantke, E. Takimoto (Eds.), Proceedings of the 14th International Conference on Algorithmic Learning Theory (ALT'03), Sapporo, Japan, Lecture Notes in Artificial Intelligence, vol. 2842, Springer-Verlag, Berlin Heidelberg, October 17–19, 2003, pp. 190–204.
[276] L. Wu, R. Jin, S.C.-H. Hoi, J. Zhu, N. Yu, Learning Bregman distance functions and its application for semi-supervised clustering, in: Y. Bengio, D. Schuurmans, J. Lafferty, C.K.I. Williams, A. Culotta (Eds.), Advances in Neural Information Processing Systems 22, Vancouver, British Columbia, Canada, NIPS Foundation, December 7–10, 2009, pp. 2089–2097.
[277] Y. Wu, Y. Jin, M. Chen, Model selection in loglinear models using ϕ-divergence measures and $M\phi E$ s, Sankhyā: The Indian Journal of Statistics—Series A 71 (2) (2009) 260–283.
[278] R.W. Yeung, A First Course in Information Theory. Information Technology: Transmission, Processing and Storage, Springer-Verlag, 2002.
[279] R.W. Yeung, Information Theory and Network Coding. Information Technology: Transmission, Processing and Storage, Springer-Verlag, 2008.
[280] W. Yin, S. Osher, D. Goldfarb, J. Darbon, Bregman iterative algorithms for ℓ_1-minimization with applications to compressed sensing, SIAM Journal on Imaging Sciences 1 (1) (2008) 143–168.
[281] S. Yu, P.G. Mehta, The Kullback–Leibler rate pseudo-metric for comparing dynamical systems, IEEE Transactions on Automatic Control 55 (July (7)) (2010) 1585–1598.
[282] R.G. Zaripov, New Measures and Methods in Information Theory. A. N. Tupolev State Technical University Press, Kazan, Tatarstan, ⟨www.imm.knc.ru/zaripov-measures.html⟩, 2005 (in Russian).
[283] J. Zhang, Divergence function, duality, and convex analysis, Neural Computation 16 (January (1)) (2004) 159–195.
[284] J. Ziv, M. Zakaï, On functionals satisfying a data-processing theorem, IEEE Transactions on Information Theory 19 (May (3)) (1973) 275–283.