Divergence measures for statistical data processing

Michèle Basseville

Publications Internes de l'IRISA
ISSN : 2102-6327
PI 1961 – Novembre 2010

Divergence measures for statistical data processing

Michèle Basseville *
michele.basseville@irisa.fr

Abstract: This note provides a bibliography of investigations based on or related to divergence measures for theoretical and applied inference problems.

Key-words: Distance measures, f-divergences, Bregman divergences, α-divergences, barycenters, divergence-based statistical inference, spectral divergence measures.

Mesures de distance pour le traitement statistique de données

Résumé : *Cette note contient une bibliographie de travaux concernant l'utilisation de divergences dans des problèmes relatifs à l'inférence statistique et ses applications.*

Mots clés : *Mesures de distance, f-divergences, divergences de Bregman, α-divergences, barycentres, inférence statistique et divergences, divergences spectrales.*

---

* CNRS-IRISA

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, recognition, indexation, diagnosis, model selection ...

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, 37, 59, 100, 139, 147, 203, 235], a number of books have been published during the last decade or so [14, 19, 31, 57, 60, 74, 101, 126, 153, 181, 224, 243, 244, 247].

In a report on divergence measures and their tight connections with the notions of entropy, information and mean values, an attempt has been made to describe various procedures for building divergences measures from entropy functions or from generalized mean values, and conversely for defining entropies from divergence measures [25]. Another goal was to clarify the properties of and relationships between two main classes of divergence measures, namely the f-divergences and the Bregman divergences. This was summarized in the conference paper [27].

The purpose of this note is to provide a bibliography for a wide variety of investigations based on or related to divergence measures for theoretical and applied inference problems. The note 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 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 [6, 61, 62, 170] and then again in the seventies [5, 249]. Kullback-Leibler divergence, Hellinger distance, χ^2-divergence, Csiszár α-divergence, and Kolmogorov total variation distance are some well known instances of f-divergences. Other instances may be found in [134, 135, 149].

f-divergences enjoy some **invariance** properties investigated in [147, 148] (see also [25]) and a universal **monotonicity** property known under the name of generalized data processing theorem [165, 249]. Their topological properties are investigated in [63]. An extension of the family of f-divergences to squared metric distances is introduced in [237].

Non-asymptotic **variational formulations** of f-divergences have been recently investigated in [41, 174, 204, 205]. Early results on that issue obtained for Kullback-Leibler divergence trace back to [79, 100]. 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 [175].

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 [174], as surrogate functions for independence and ICA in [161], and the α-divergence in (9) is used in [160] 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 [35, 104], respectively.

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 [226, 227].

f-divergences also turn out useful for defining robust **projection pursuit** indices [172]. 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 [231].

Nonnegative matrix factorization (**NMF**), of widespread use in multivariate analysis and linear algebra [78], 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 [86], Itakura-Saito divergence in [90], f-divergences in [55], or α-divergences in [56, 137]; see also [57].

The maximizers of the Kullback-Leibler divergence from an exponential family and from any hierarchical log-linear model are derived in [201] and [163], respectively.

Other investigations include comparison of experiments [230]; minimizing f-divergences on sets of signed measures [42]; minimizing multivariate entropy functionals with application to minimizing f-divergences in both variables [69]; determining the joint range of f-divergences [106, 108]; 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 [216].

Recent **applications** involving the use of f-divergences concern feature selection in fuzzy approximation spaces [155], the selection of discriminative genes from micro-array data [154], medical image registration [192], or speech recognition [193], to mention but a few examples. A modification of f-divergences for finite measures turns out to be useful for right-censored observations [221].

3 Bregman divergences

Bregman divergences, introduced in [39], are defined for vectors, matrices, functions and probability distributions.

The Bregman divergence between vectors is defined as:

$$D_\varphi(x,y) = \varphi(x) - \varphi(y) - (x-y)^T \ \nabla\varphi(y) \tag{2}$$

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

$$\bar{D}_\varphi(x,y) = (\nabla\varphi(x) - \nabla\varphi(y))^T \ \ (x-y) \tag{3}$$

The Bregman matrix divergence is defined as:

$$D_\phi(X,Y) = \phi(X) - \phi(Y) - \mathrm{Tr}\left((\nabla\phi(Y))^T \ (X-Y)\right) \tag{4}$$

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 [77].

For $\phi(X) = \ln|X|$, the divergence (4) 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 [138], that distance is likely to trace back to [116]. For example, it has been used for estimating structured covariance matrices [48] and for designing residual generation criteria for monitoring [26].

The divergence (4), in the general case for ϕ, has been recently proposed for designing and investigating a new family of self-scaling quasi-Newton methods [122, 123].

The Bregman divergence between probability densities is defined as [66, 119]:

$$D_\varphi(p,q) = \int \left(\varphi(p) - \varphi(q) - (p-q) \ \varphi'(q)\right) \ dx \tag{5}$$

A Bregman divergence can also be seen as the limit of a **Jensen difference** [27, 176], namely:

$$D_\varphi(p,q) = \lim_{\beta \to 1} \ \frac{1}{1-\beta} \ J_\varphi^{(\beta)}(p,q) \tag{6}$$

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

$$J_\varphi^{(\beta)}(p,q) \stackrel{\Delta}{=} \beta\ \varphi(p) + (1-\beta)\ \varphi(q) - \varphi\left(\beta p + (1-\beta)q\right) \tag{7}$$

For $\beta = 1/2$, Jensen difference is the **Burbea-Rao divergence** [47]; see also [82, 149]. The particular case where D and J are identical is of interest [200].

The Bregman divergence measures enjoy a number of properties useful for learning, clustering and many other inference [23, 66, 87] and quantization [21] problems. In the discrete case, an asymptotic equivalence with the f-divergence and the Burbea-Rao divergence is investigated in [183].

One important instance of the use of Bregman divergences for **learning** is the case of inverse problems [119, 142][1], where convex duality is extensively used. Convex duality is also used for minimizing a class of Bregman divergences subject to linear constraints in [71], whereas a simpler proof using convex analysis is provided in [32], and the results are used in [58].

Bregman divergences have been used for a generalization of the LMS adaptive filtering algorithm [133], for Bayesian estimation of distributions [88] using functional divergences introduced for quadratic discriminant analysis [217], and for l_1-regularized logistic regression posed as Bregman distance minimization and solved with non-linear constrained optimization techniques in [70]. Iterative l_1-minimization with application to compressed sensing is investigated in [245]. Mixture models are estimated using Bregman divergences within a modified EM algorithm in [89]. Learning continuous latent variable models with Bregman divergences and alternating minimization is addressed in [241]. 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 [180].

Learning structured models, such as Markov networks or combinatorial models, is performed in [225] 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 [202].

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 [226, 227]. Clustering with Bregman divergences unifying k-means, LBG and other information theoretic clustering approaches is investigated in [23], together with a connection with rate distortion theory[2]. Scaled Bregman distances are used in [220] and compared with f-divergences for key properties: optimality of the k-means algorithm [23] and invariance w.r.t. statistically sufficient transformations. Quantization and clustering with Bregman divergences are investigated in [87] together with convergence rates. k-means and hierarchical classification algorithms w.r.t. Burbea-Rao divergences (expressed as Jensen-Bregman divergences) are studied in [176].

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 [58, 132, 141], and [140] for an earlier study. Some controversy does exist however, see for example [143] 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 geometry is investigated in [171], 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

[1] See also [214] for another investigation of regularization.

[2] Note that minimum cross-entropy classification was addressed as an extension of coding by vector quantization in [213].

gradient updates and Bregman projections is investigated in [232]. Learning low-rank positive semidefinite (kernel) matrices for machine learning applications is also addressed with Bregman divergences in [138]. Matrix rank minimization is achieved with a Bregman iterative algorithm in [152]. Nonnegative matrix approximation with low rank matrices is discussed in [76], 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 [22]. Nonnegative matrix factorization (NMF) with Bregman divergences is addressed in [56]; see also [57]. The particular case of using the density power divergence and a surrogate function for NMF is investigated in [91].

Applications involving the use of Bregman divergences concern nearest neighbor retrieval [51], color image segmentation [176], 3D image segmentation and word alignment [225], cost-sensitive classification for medical diagnosis (UCI datasets) [209], magnetic resonance image analysis [238], semi-supervised clustering of high dimensional text benchmark datasets and low dimensional UCI datasets[3] [242], content-based multimedia retrieval with efficient neighbor queries [178], efficient range search from a query in a large database [52].

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 [223], and the reader is referred to [2, 65] for critical surveys. However, a number of useful α-divergence measures have been proposed, tracing back to Chernoff [53] and Rao [195]. A recent synthesis can be found in [54]; see also [57].

The **Csiszár I-divergences of order α**, also called **Havrda-Charvát's α-divergences**, have been introduced in [62, 110] as *f-divergences* (1) associated with:

$$f_\alpha(u) = \begin{cases} \frac{4}{1-\alpha^2}\ \left(1 - u^{(1+\alpha)/2}\right) - \frac{2}{1-\alpha}\ (u-1) & \alpha \neq \pm 1 \\ u \ln u - (u-1) & \alpha = +1 \\ -\ln u + (u-1) & \alpha = -1 \end{cases} \tag{8}$$

namely:

$$D_{f_\alpha}(p,q) = \begin{cases} \frac{4}{1-\alpha^2}\ \left(1 - \int p^{(1-\alpha)/2}\ q^{(1+\alpha)/2}\ dx\right) & \alpha \neq \pm 1 \\ K(p,q) & \alpha = +1 \\ K(q,p) & \alpha = -1 \end{cases} \tag{9}$$

See also [3, 147, 234, 235].

They are also known under the name of **Amari's α-divergences** [8, 14] - see also [10], and identical to **Tsallis divergences** [233]. The **Read-Cressie's power divergence** [147, 150, 203, 235] turns out to be identical to the Havrda-Charvát's α-divergence, up to a transformation of α. For $\alpha = +1$ this divergence is nothing but the Kullback-Leibler information, and for $\alpha = 0$ the Hellinger distance. The α-divergence (9) has been used earlier by Chernoff for investigating the asymptotic efficiency of statistical tests [53].

Some applications of those divergences are described in particular for model integration in [10], in EM algorithms in [158, 159, 160], and message-passing algorithms for complex Bayesian networks approximation in [167].

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 [66] that Kullback-Leibler divergence is the only Bregman divergence which is an f-divergence.

f-divergences based on Arimoto's entropies [17] and introduced in [218] define α-divergences different from the above ones.

[3] For which a non-parametric approach to learning φ in (3) 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.

A different class of α-divergences, known under the name of **density power divergences**, have been introduced in [28] by Basu *et al.* as:

$$D_\alpha(p,q) = \begin{cases} 1/\alpha \int \left(p^{\alpha+1} - (\alpha+1)\ p\ q^\alpha + \alpha\ q^{\alpha+1}\right)\ dx & \alpha > 0 \\ K(p,q) & \alpha = 0 \end{cases} \tag{10}$$

They can be seen as *Bregman divergences* D_{φ_α} (5) associated with:

$$\varphi_\alpha(u) = \begin{cases} 1/\alpha\ \left(u^{\alpha+1} - u\right) & \alpha > 0 \\ u \ln u & \alpha = 0 \end{cases} \tag{11}$$

They have been used for robust blind source separation [166], analyzing mixtures ICA models [169], model selection [124, 162] and estimation of tail index of heavy tailed distributions [131]. They have recently been handled in [81] 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 [206] as:

$$D_\alpha(p,q) = \begin{cases} \frac{1}{\alpha(\alpha-1)}\ \ln \int p^{1-\alpha}\ q^\alpha\ dx & \alpha \neq 1 \\ -\ K(p,q) & \alpha = 1 \end{cases} \tag{12}$$

although they may have been proposed earlier [34, 210]. Those divergences exhibit direct links with the Chernoff distance and with the moment generating function of the likelihood ratio [25, 112]. Their use for channel capacity[4] and their link with cutoff rates are addressed in [67]. Their involvment in estimation and coding problems is also address in [15, 16, 173]. Scale and concentration invariance properties have been investigated in [136].

A surrogate function for the Rényi's α-divergence is the α-Jensen difference [112]. Entropy functionals derived from Rényi's divergences have been recently studied in [33], whereas characterizations of maximum Rényi's entropy distributions are provided in [109, 239].

In recent years, the Rényi's α-divergences have been used for robust image registration [111], differentiating brain activity [20], for feature classification, indexing, and retrieval in image and other databases [112], and detecting distributed denial-of-service attacks [146]. They have been shown to be possibly irrelevant for blind source separation [191, 240].

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 [248], 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 [164] and taxonomy [196, 197], 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 [105].

Generalized Jensen divergences are defined in [149] as:

$$J_\varphi^{(\beta)}(p_1,\ldots,p_n) \triangleq \sum_{i=1}^n \beta_i\ \varphi(p_i) - \varphi\left(\sum_{i=1}^n \beta_i\ p_i\right) \tag{13}$$

In the case of the Shannon entropy, namely $\varphi(x) = -x \ln x$, 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.

[4] In [17], another α-information is used for channel capacity.

Jensen divergence can also be written as the arithmetic mean of Bregman divergences:

$$\mathbf{J}_H^{(\beta)}(P_1, P_2) = \beta\, \mathbf{D}_H(P_1, \beta P_1 + (1-\beta)P_2) + (1-\beta)\, \mathbf{D}_H(P_2, \beta P_1 + (1-\beta)P_2) \tag{14}$$

This has been applied to word clustering for text classification [75].

Actually the interplay between divergences, entropies and mean values is quite tight [2, 25, 27, 206], and some other means than the arithmetic mean may be used in the definition of such generalized divergences. The **information radius** introduced in [215] is the generalized mean of the Rényi's divergences between each of the p_i's and the generalized mean of all the p_i's, which boils down to [25]:

$$\mathcal{S}_\alpha^{(\beta)}(p_1, \ldots, p_n) = \frac{\alpha}{\alpha - 1}\, \ln \int \left(\sum_{i=1}^{n}\, \beta_i\, p_i^\alpha(x) \right)^{1/\alpha} d\lambda(x) \tag{15}$$

See also [67].

Mean values, barycenters, centroids have been widely investigated. The barycenter of a set of probability measures is studied in [188]. Barycenters in a dually flat space are introduced as minimizers of averaged Amari's divergences in [186]. Left-sided, right-sided and symmetrized centroids are defined as minimizers of averaged Bregman divergences in [177], whereas Burbea-Rao centroids are the topic of [176] with application to color image segmentation.

Geometric means of symmetric positive definite matrices are investigated in [18]. A number of Fréchet means are discussed in [80] with application to diffusion tensor imaging. Quasi-arithmetic means of multiple positive matrices by symmetrization from the mean of two matrices are investigated in [190]. Riemannian metrics on space of matrices are addressed in [189].

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 [207]. 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 [31, 60, 147, 153, 181, 235], 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 (11) and written as M-estimates, are investigated in [28] in terms of consistency, influence function, equivariance, and robustness; see also [29, 150, 184]. Iteratively reweighted estimating equations for robust minimum distance estimation are proposed in [30] whereas a boostrap root search is discussed in [156]. Recent investigations of the power divergence estimates include robustness to outliers and a local learning property [81] and Bahadur efficiency [107]. An application to robust blind source separation is described in [166].

The role of duality when investigating divergence minimization for statistical inference is addressed in [7, 41]; see also [103, 248]. A further investigation of the minimum divergence estimates introduced in [41] can be found in [228], which addresses the issues of influence function, asymptotic relative efficiency, and empirical performances. See also [187] for another investigation.

A comparison of density-based minimum divergence estimates is presented in [120]. A recent comparative study of four types of minimum divergence estimates is reported in [44], in terms of consistency and influence curves: this includes the power divergence estimates [28], the so-called power superdivergence estimates [40, 148, 236], the power subdivergence estimates [41, 236], and the Rényi pseudo-distance estimates [144, 236].

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 [85].

Robust LS estimates with a Kullback-Leibler divergence constraint are introduced and investigated in [145] where a connection with risk-sensitive filtering is established.

Hypothesis testing may also be addressed within such a framework [36]. The asymptotic distribution of a generalized entropy functional and the application to hypothesis testing and design of confidence intervals are studied in [83]. The asymptotic distribution of tests statistics built on divergence measures based on entropy functions is derived in [181, 182]. This includes extensions of Burbea-Rao's J-divergences [46, 47] and of the Sibson's information radius [215].

Tests statistics based on entropy or divergence of hypothetical distributions with ML estimated values of the parameter have been recently investigated in [41, 43, 44]. Robustness properties of these tests are proven in [228]. 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 [106, 108].

Maximum entropy, minimum divergence and Bayesian decision theory are investigated in [103] 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.

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 [113] approximately follows the gradient of the difference of two divergences:

$$CD_n = KL(p_0, p^\star) - KL(p_n, p^\star) \tag{16}$$

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 [50, 168]. The convergence properties of contrastive divergence learning are analysed in [222].

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 $\psi_\theta(\xi) \stackrel{\Delta}{=} \nabla_\xi \ln p_\theta(\xi)$ and $\psi_x(\cdot) \stackrel{\Delta}{=} \nabla_\xi \ln p_x(\cdot)$. This objective function turns out to be very useful for estimating non-normalized statistical models [114, 115].

7 Spectral divergence measures

Spectral distance measures for scalar signal processing have been thoroughly investigated in [99, 102]. Spectral distances between vector Gaussian processes have been studied in [127, 128, 129, 211, 212]; see also [130][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 [219]. A distance based on mutual information for Gaussian processes is investigated in [38].

Kullback-Leibler and/or Hellinger distances have been used for spectral interpolation [49, 92, 125], spectral approximation [84, 95, 185], spectral estimation [194], and ARMA modeling [96].

Differential-geometric structures for prediction and smoothing problems for spectral density functions are introduced in [93]. This work has been recently pursued in [246] 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 [94].

8 Miscellanea

Information geometry, which investigates information for statistical inference based on differential geometry, has been studied by Csiszár [64], Rao [45, 199], Amari [8, 9, 13, 14] and Kass [126]. A recent book [19] explores neighbourhoods 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 [97]. The role of information geometry in asymptotic statistical theory is discussed in [24]. The information geometric structure of the generalized empirical likelihood method based on minimum divergence estimates is investigated in [179]. 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 [72, 73, 117, 121, 151]. Inequalities involving f-divergences are studied in [98, 106, 108, 229].

The **axiomatic characterization** of information and divergence measures is addressed in [3, 66, 118, 157, 198, 206] [235, chap10]; see also [1]. Additional references are provided in [25]. More recent treatments may be found in [68, 208] and [94].

References

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

[2] J. Aczél. Measuring information beyond communication theory - Why some generalized information measures may be useful, others not. *Aequationes mathematicae*, 27:1–19, 1984.

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

[4] A. Agarwal and H. Daume III. A geometric view of conjugate priors, May 2010. arXiv.

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

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

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

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

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

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

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

[12] S.-I. Amari. Information geometry and its applications: Convex function and dually flat manifold. In *Emerging Trends in Visual Computing - LIX Colloquium, Nov. 2008*, volume 5416 of *Lecture Notes in Computer Science*, pages 75–102. Springer-Verlag, 2009.

[13] S.-I. Amari. Information geometry derived from divergence functions. In *Proceedings of the 3rd International Symposium on Information Geometry and its Applications, Leipzig, FRG, August 2-6, 2010*, 2010.

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

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

[16] S. Arimoto. Information-theoretical considerations on estimation problems. *Information and Control*, 19(3):181–194, Oct. 1971.

[17] S. Arimoto. Information measures and capacity of order α for discrete memoryless channels. In *Topics in Information Theory - 2nd Colloquium, Keszthely, HU, 1975*, volume 16 of *Colloquia Mathematica Societatis János Bolyai*, pages 41–52. North Holland, Amsterdam, NL, 1977.

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

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

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

[21] A. Banerjee, I. Dhillon, J. Ghosh, and S. Merugu. An information theoretic analysis of maximum likelihood mixture estimation for exponential families. In C. E. Brodley, editor, *Proceedings of the 21st International Conference on Machine Learning (ICML'04), Banff, Alberta, Canada, July 4-8, 2004*, ACM International Conference Proceeding Series, New York, NY, USA, 2004.

[22] A. Banerjee, I. Dhillon, J. Ghosh, S. Merugu, and D. S. Modha. A generalized maximum entropy approach to Bregman co-clustering and matrix approximation. *Journal of Machine Learning Research*, 8:1919–1986, Aug. 2007.

[23] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh. Clustering with Bregman divergences. *Journal of Machine Learning Research*, 6:1705–1749, Oct. 2005.

[24] O. E. Barndorff-Nielsen, D. R. Cox, and N. Reid. The role of differential geometry in statistical theory. *International Statistical Review*, 54(1):83–96, Apr. 1986.

[25] M. Basseville. Information: entropies, divergences et moyennes. Research Report 1020, IRISA, May 1996. In French - Online.

[26] M. Basseville. Information criteria for residual generation and fault detection and isolation. *Automatica*, 33(5):783–803, May 1997.

[27] M. Basseville and 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*, page 330, Sept. 1995.

[28] A. Basu, I. R. Harris, N. Hjort, and M. C. Jones. Robust and efficient estimation by minimising a density power divergence. *Biometrika*, 85(3):549–559, Sept. 1998.

[29] A. Basu and B. G. Lindsay. Minimum disparity estimation for continuous models: efficiency, distributions and robustness. *Annals of the Institute of Statistical Mathematics*, 46(4):683–705, Dec. 1994.

[30] A. Basu and B. G. Lindsay. The iteratively reweighted estimating equation in minimum distance problems. *Computational Statistics and Data Analysis*, 45(2):105–124, Mar. 2004.

[31] A. Basu, H. Shioya, and C. Park. *Statistical Inference: The Minimum Distance Approach*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. CRC Press, Boca Raton, FL, Jan. 2011. To appear.

[32] H. H. Bauschke. Duality for Bregman projections onto translated cones and affine subspaces. *Journal of Approximation Theory*, 121(1):1–12, Mar. 1983.

[33] J.-F. Bercher. On some entropy functionals derived from Rényi information divergence. *Information Sciences*, 178(12):2489–2506, June 2008.

[34] A. Bhattacharyya. On a measure of divergence between two statistical populations defined by their probability distributions. *Bulletin of the Calcutta Mathematical Society*, 35:99–109, 1943.

[35] L. Birgé. A new lower bound for multiple hypothesis testing. *IEEE Transactions on Information Theory*, 51(4):1611–1615, Apr. 2005.

[36] R. E. Blahut. Hypothesis testing and information theory. *IEEE Transactions on Information Theory*, 20(4):405–417, July 1974.

[37] R. E. Blahut. *Principles and Practice of Information Theory*. Series in Electrical and Computer Engineering. Addison Wesley Publishing Co, 1987.

[38] J. Boets, K. De Cock, and B. De Moor. A mutual information based distance for multivariate Gaussian processes. In A. Chiuso, A. Ferrante, and S. Pinzoni, editors, *Modeling, Estimation and Control, Festschrift in Honor of Giorgio Picci on the Occasion of his Sixty-Fifth Birthday*, volume 364 of *Lecture Notes in Control and Information Sciences*, pages 15–33. Springer-Verlag, Berlin, FRG, Oct. 2007.

[39] 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):200–217, 1967.

[40] M. Broniatowski and A. Keziou. Minimization of φ-divergences on sets of signed measures. *Studia Scientiarum Mathematicarum Hungarica*, 43(4):403–442, Dec. 2006.

[41] M. Broniatowski and A. Keziou. Parametric estimation and tests through divergences and the duality technique. *Journal of Multivariate Analysis*, 100(1):16–36, Jan. 2009.

[42] M. Broniatowski and A. Keziou. Minimization of divergences on sets of signed measures, Mar. 2010. arXiv.

[43] M. Broniatowski and A. Keziou. On generalized empirical likelihood methods, Feb. 2010. arXiv.

[44] M. Broniatowski and I. Vajda. Several applications of divergence criteria in continuous families, Nov. 2009. arXiv.

[45] J. Burbea and C. R. Rao. Entropy differential metric, distance and divergence measures in probability spaces: a unified approach. *Journal of Multivariate Analysis*, 12(4):575–596, Dec. 1982.

[46] J. Burbea and C. R. Rao. On the convexity of higher order Jensen differences based on entropy functions. *IEEE Transactions on Information Theory*, 28(6):961–963, Nov. 1982.

[47] J. Burbea and C. R. Rao. On the convexity of some divergence measures based on entropy functions. *IEEE Transactions on Information Theory*, 28(3):489–495, May 1982.

[48] J. Burg, D. Luenberger, and D. Wenger. Estimation of structured covariance matrices. *Proceedings of the IEEE*, 70(9):963–974, Sept. 1982.

[49] C. I. Byrnes, T. T. Georgiou, and A. Lindquist. A generalized entropy criterion for Nevanlinna-Pick interpolation with degree constraint. *IEEE Transactions on Automatic Control*, 46(6):822–839, June 2001.

[50] M. A. Carreira-Perpiñán and G. E. Hinton. On contrastive divergence learning. In R. Cowell and Z. Ghahramani, editors, *Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics (AISTATS'05), Barbados, UK, January 6-8, 2005*, pages 59–66, 2005.

[51] L. Cayton. Fast nearest neighbor retrieval for Bregman divergences. In W. W. Cohen, A. McCallum, and S. T. Roweis, editors, *Proceedings of the 25th International Conference on Machine Learning (ICML'08)*, pages 112–119, Helsinki, Finland, June 2008.

[52] L. Cayton. Efficient Bregman range search. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22, Vancouver, British Columbia, Canada, December 7-10, 2009*, pages 243–251. NIPS Foundation, 2009.

[53] H. Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Annals of Mathematical Statistics*, 23(4):493–507, Dec. 1952.

[54] A. Cichocki and S.-I. Amari. Families of alpha- beta- and gamma- divergences: Flexible and robust measures of similarities. *Entropy*, 12(6):1532–1568, June 2010.

[55] A. Cichocki, R. Zdunek, and S.-I. Amari. Csiszár's divergences for non-negative matrix factorization: Family of new multiplicative algorithm. In J. P. Rosca, D. Erdogmus, J. C. Príncipe, and S. Haykin, editors, *Proceedings of the 6th International Conference on Independent Component Analysis and Blind Source Separation (ICA'06), Charleston, South Carolina, USA, March 5-8, 2006*, volume 3889 of *Lecture Notes in Computer Science*, pages 32–39. Springer-Verlag, Berlin Heidelberg, FRG, 2006.

[56] A. Cichocki, R. Zdunek, and S.-I. Amari. Nonnegative matrix and tensor factorization. *IEEE Signal Processing Magazine*, 25(1):142–145, Jan. 2008.

[57] A. Cichocki, R. Zdunek, A. Phan, and S.-I. Amari. *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation.* John Wiley & Sons Ltd, 2009.

[58] M. Collins, R. E. Schapire, and Y. Singer. Logistic regression, AdaBoost and Bregman distances. *Machine Learning*, 48(1-3):253–285, July 2002.

[59] T. M. Cover and J. A. Thomas. *Elements of Information Theory.* Wiley Series in Telecommunications. John Wiley & Sons Ltd, 1991.

[60] T. M. Cover and J. A. Thomas. *Elements of Information Theory, Second Edition.* John Wiley & Sons Ltd, 2006.

[61] I. Csiszár. Eine informationstheoretische ungleichung und ihre anwendung auf den beweis der ergodizitat von markoffschen ketten. *Magyar. Tud. Akad. Mat. Kutato Int. Kozl*, 8:85–108, 1963.

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

[63] I. Csiszár. On topological properties of f-divergence. *Studia Scientiarum Mathematicarum Hungarica*, 2:329–339, 1967.

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

[65] I. Csiszár. Information measures: a critical survey. In J. Kozesnik, editor, *Transactions of the 7th Conference on Information Theory, Statistical Decision Functions, Random Processes, Prague, August 18-23, 1974*, volume B, pages 73–86. Academia, Prague, 1978.

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

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

[68] I. Csiszár. Axiomatic characterizations of information measures. *Entropy*, 10(3):261–273, Sept. 2008.

[69] I. Csiszár and F. Matus. On minimization of multivariate entropy functionals. In V. Anantharam and I. Kontoyiannis, editors, *Proceedings of the IEEE Information Theory Workshop on Networking and Information Theory (ITW'09), Volos, Greece, June 10-12, 2009*, pages 96–100, 2009.

[70] M. Das Gupta and T. S. Huang. Bregman distance to l_1 regularized logistic regression, Apr. 2010. arXiv.

[71] S. Della Pietra, V. Della Pietra, and J. Lafferty. Duality and auxiliary functions for Bregman distances. Technical Report Collection CMU-CS-01-109R, School of Computer Science, Carnegie Mellon University, Feb. 2002.

[72] A. Dembo. Information inequalities and concentration of measure. *Annals of Probability*, 25(2):927–939, Apr. 1997.

[73] A. Dembo, T. M. Cover, and J. A. Thomas. Information theoretic inequalities. *IEEE Transactions on Information Theory*, 37(6):1501–1518, Nov. 1991.

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

[75] I. S. Dhillon, S. Mallela, and R. Kumar. A divisive information-theoretic feature clustering algorithm for text classification. *Journal of Machine Learning Research*, 3:1265–1287, Mar. 2003.

[76] I. S. Dhillon and S. Sra. Generalized nonnegative matrix approximations with Bregman divergences. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems 18, Vancouver, British Columbia, Canada, December 5-8, 2005*, pages 283–290. MIT Press, Cambridge, MA, 2006.

[77] I. S. Dhillon and J. A. Tropp. Matrix nearness problems with Bregman divergences. *SIAM Journal on Matrix Analysis and Applications*, 29(4):1120–1146, 2007.

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

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

[80] I. L. Dryden, A. Kolydenko, D. Zhou, and B. Li. Non-Euclidean statistical analysis of covariance matrices and diffusion tensors, Oct. 2010. arXiv.

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

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

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

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

[85] D. Ferrari and Y. Yang. Maximum L_q-likelihood estimation. *Annals of Statistics*, 38(2):753–783, Apr. 2010.

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

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

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

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

[90] C. Févotte, N. Bertin, and J.-L. Durrieu. Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis. *Neural Computation*, 21(3):793–830, Mar. 2009.

[91] C. Févotte and J. Idier. Algorithms for nonnegative matrix factorization with the beta-divergence, Oct. 2010. arXiv.

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

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

[94] T. T. Georgiou, J. Karlsson, and M. S. Takyar. Metrics for power spectra: An axiomatic approach. *IEEE Transactions on Signal Processing*, 57(3):859–867, Mar. 2009.

[95] T. T. Georgiou and A. Lindquist. Kullback-Leibler approximation of spectral density functions. *IEEE Transactions on Information Theory*, 49(11):2910–2917, Nov. 2003.

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

[97] P. Gibilisco, E. Riccomagno, M. P. Rogantin, and H. P. Wynn, editors. *Algebraic and Geometric Methods in Statistics*. Cambridge University Press, Cambridge, UK, 2010.

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

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

[100] R. M. Gray. *Entropy and Information Theory.* Springer-Verlag, New York, NY, USA, 1990. Corrected version, 2009, Online.

[101] R. M. Gray. *Entropy and Information Theory, Second Edition.* Springer-Verlag, New York, NY, USA, 2010.

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

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

[104] A. Guntuboyina. Lower bounds for the minimax risk using f-divergences and applications, Feb. 2010. arXiv.

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

[106] P. Harremoës and I. Vajda. Joint range of f-divergences. In M. Gastpar, R. Heath, and K. Narayanan, editors, *Proceedings of the IEEE International Symposium on Information Theory (ISIT'10), Austin, TX, USA, June 13-18, 2010*, pages 1345–1349, 2010.

[107] P. Harremoës and I. Vajda. On Bahadur efficiency of power divergence statistics, Feb. 2010. arXiv.

[108] P. Harremoës and I. Vajda. On pairs of f-divergences and their joint range, July 2010. arXiv.

[109] P. Harremoës and C. Vignat. Rényi entropies of projections. In A. Barg and R. W. Yeung, editors, *Proceedings of the IEEE International Symposium on Information Theory (ISIT'06), Seattle, WA, USA, July 9-14, 2006*, pages 1827–1830, 2006.

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

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

[112] A. O. Hero, B. Ma, O. Michel, and J. Gorman. Alpha-divergence for classification, indexing and retrieval. Research Report CSPL-328, University of Michigan, Communications and Signal Processing Laboratory, May 2001.

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

[114] A. Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6:695–709, Apr. 2005.

[115] A. Hyvärinen. Some extensions of score matching. *Computational Statistics and Data Analysis*, 51(5):2499–2512, Feb. 2007.

[116] W. James and C. Stein. Estimation with quadratic loss. In J. Neyman, editor, *Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 361–379. University of California Press, Berkeley, CA, USA, 1961.

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

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

[119] L. K. Jones and 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(1):23–30, Jan. 1990.

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

[121] A. Kagan and T. Yu. Some inequalities related to the Stam inequality. *Applications of Mathematics*, 53(2):195–205, 2008.

[122] T. Kanamori and A. Ohara. A Bregman extension of quasi-Newton updates I: An information geometrical framework, Oct. 2010. arXiv.

[123] T. Kanamori and A. Ohara. A Bregman extension of quasi-Newton updates II: Convergence and robustness properties, Oct. 2010. arXiv.

[124] A. Karagrigoriou and T. Papaioannou. On measures of information and divergence and model selection criteria. In F. Vonta, M. Nikulin, N. Limnios, and C. Huber-Carol, editors, *Statistical Models and Methods for Biomedical and Technical Systems*, Statistics for Industry and Technology, pages 503–518. Birkhäuser, Boston, MA, USA, 2008.

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

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

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

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

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

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

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

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

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

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

[135] L. Knockaert. Statistical thermodynamics and natural f-divergences, 1994. Unpublished paper, `http://users.ugent.be/ lknockae/`.

[136] L. Knockaert. On scale and concentration invariance in entropies. *Information Sciences*, 152:139–144, June 2003.

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

[138] B. Kulis, M. A. Sustik, and I. S. Dhillon. Low-rank kernel learning with Bregman matrix divergences. *Journal of Machine Learning Research*, 10:341–376, Feb. 2009.

[139] S. Kullback, J. C. Keegel, and J. H. Kullback. *Topics in Statistical Information Theory*, volume 42 of *Lecture Notes in Statistics*. Springer-Verlag, New York, NY, USA, 1987.

[140] J. D. Lafferty. Statistical learning algorithms based on Bregman distances. In *Proceedings of the 1997 Canadian Workshop on Information Theory, Toronto, Canada, June 3-6, 1997*, pages 77–80, 1997.

[141] 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, July 7-9, 1999*, pages 125–133. ACM, 1999.

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

[143] G. Lebanon and J. Lafferty. Boosting and maximum likelihood for exponential models. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14, Vancouver, British Columbia, Canada, December 3-8, 2001*, Cambridge, MA, 2002. MIT Press.

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

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

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

[147] F. Liese and I. Vajda. *Convex Statistical Distances*, volume 95 of *Texte zur Mathematick*. Teubner, Leipzig, 1987.

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

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

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

[151] E. Lutwak, D. Yang, and G. Zhang. Cramér-Rao and moment-entropy inequalities for Rényi entropy and generalized Fisher information. *IEEE Transactions on Information Theory*, 51(2):473–478, Feb. 2005.

[152] S. Ma, D. Goldfarb, and L. Chen. Fixed point and Bregman iterative methods for matrix rank minimization. *Mathematical Programming, Series A*, 2010 - To appear.

[153] D. MacKay. *Information Theory, Inference & Learning Algorithms*. Cambridge University Press, Cambridge, UK, 2003.

[154] P. Maji. f-information measures for efficient selection of discriminative genes from microarray data. *IEEE Transactions on Biomedical Engineering*, 56(4):1063–1069, Apr. 2009.

[155] P. Maji and S. K. Pal. Feature selection using f-information measures in fuzzy approximation spaces. *IEEE Transactions on Knowledge and Data Engineering*, 22(6):854–867, June 2010.

[156] M. Markatou, A. Basu, and B. G. Lindsay. Weighted likelihood equations with bootstrap root search. *Journal of the American Statistical Association*, 93(442):740–750, June 1998.

[157] A. Mathai and P. Rathie. *Basic Concepts in Information Theory and Statistics*. Wiley Eastern Ltd, New Delhi, India, 1975.

[158] 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*, page 385, 1998.

[159] Y. Matsuyama. The α-EM algorithm and its applications. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'00), Istanbul, Turkey, June 5-9, 2000*, volume 1, pages 592–595, 2000.

[160] Y. Matsuyama. The α-EM algorithm: surrogate likelihood maximization using α-logarithmic information measures. *IEEE Transactions on Information Theory*, 49(3):692–706, Mar. 2003.

[161] Y. Matsuyama, N. Katsumata, and S. Imahara. Convex divergence as a surrogate function for independence: The f-divergence. In T.-W. Lee, T.-P. Jung, S. Makeig, and T. J. Sejnowski, editors, *Proceedings of the 3rd International Conference on Independent Component Analysis and Blind Signal Separation*, pages 31–36, San Diego, California, USA, Dec. 2001.

[162] K. Mattheou, S. Lee, and A. Karagrigoriou. A model selection criterion based on the BHHJ measure of divergence. *Journal of Statistical Planning and Inference*, 139(2):228–235, 2009.

[163] F. Matus. Divergence from factorizable distributions and matroid representations by partitions. *IEEE Transactions on Information Theory*, 55(12):5375–5381, Dec. 2009.

[164] K. Matusita. Discrimination and the affinity of distributions. In T. Cacoullos, editor, *Discriminant Analysis and Applications*, pages 213–223. Academic Press, New York, USA, 1973.

[165] N. Merhav. Data processing theorems and the second law of thermodynamics, July 2010. arXiv.

[166] M. Minami and S. Eguchi. Robust blind source separation by beta divergence. *Neural Computation*, 14(8):1859–1886, Aug. 2002.

[167] T. Minka. Divergence measures and message passing. TechReport MSR-TR-2005-173, Microsoft Research Ltd, 2005.

[168] A. Mnih and G. Hinton. Learning nonlinear constraints with contrastive backpropagation. In D. V. Prokhorov, editor, *Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN'05), Montréal, Québec, Canada, July 31-August 4, 2005*, volume 2, pages 1302–1307, 2005.

[169] M. N. H. Mollah, M. Minami, and S. Eguchi. Exploring latent structure of mixture ICA models by the minimum β-divergence method. *Neural Computation*, 18(1):166–190, Jan. 2006.

[170] T. Morimoto. Markov processes and the H-theorem. *Journal of the Physical Society of Japan*, 18(3):328–331, Mar. 1963.

[171] N. Murata, T. Takenouchi, T. Kanamori, and S. Eguchi. Information geometry of U-Boost and Bregman divergence. *Neural Computation*, 16(7):1437–1481, July 2004.

[172] G. Nason. Robust projection indices. *Journal of the Royal Statistical Society - Series B Methodological*, 63(3):551–567, 2001.

[173] P. Nath. On a coding theorem connected with Rényi's entropy. *Information and Control*, 29(3):234–242, Nov. 1975.

[174] X. Nguyen, M. J. Wainwright, and M. I. Jordan. On surrogate loss functions and f-divergences. *Annals of Statistics*, 37(2):876–904, Apr. 2009.

[175] X. Nguyen, M. J. Wainwright, and M. I. Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, Nov. 2010.

[176] F. Nielsen and S. Boltz. The Burbea-Rao and Bhattacharyya centroids, Apr. 2010. arXiv.

[177] F. Nielsen and R. Nock. Sided and symmetrized Bregman centroids. *IEEE Transactions on Information Theory*, 55(6):2882–2904, June 2009.

[178] F. Nielsen, P. Piro, and M. Barlaud. Bregman vantage point trees for efficient nearest neighbor queries. In Q. Sun and Y. Rui, editors, *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME'09), New York, NY, USA, June 28 - July 3, 2009*, pages 878–881, 2009.

[179] T. Nishimura and F. Komaki. The information geometric structure of generalized empirical likelihood estimators. *Communications in Statistics - Theory and Methods*, 37(12):1867–1879, Jan. 2008.

[180] R. Nock and F. Nielsen. Bregman divergences and surrogates for learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(11):2048–2059, Nov. 2009.

[181] 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.

[182] L. Pardo, M. Salicrú, M. L. Menéndez, and D. Morales. Divergence measures based on entropy functions and statistical inference. *Sankhyā: The Indian Journal of Statistics - Series B*, 57(3):315–337, 1995.

[183] M. Pardo and I. Vajda. On asymptotic properties of information-theoretic divergences. *IEEE Transactions on Information Theory*, 49(7):1860–1867, July 2003.

[184] R. K. Patra, A. Mandal, and A. Basu. Minimum Hellinger distance estimation with inlier modification. *Sankhyā: The Indian Journal of Statistics - Series B*, 70(2):310–323, Nov. 2008.

[185] M. Pavon and A. Ferrante. On the Georgiou-Lindquist approach to constrained Kullback-Leibler approximation of spectral densities. *IEEE Transactions on Automatic Control*, 51(4):639–644, Apr. 2006.

[186] B. Pelletier. Informative barycentres in statistics. *Annals of the Institute of Statistical Mathematics*, 57(4):767–780, Dec. 2005.

[187] B. Pelletier. Inference in ϕ-families of distributions. *Statistics - A Journal of Theoretical and Applied Statistics*, 2010. To appear.

[188] A. Perez. Barycenter of a set of probability measures and its application in statistical decision. In T. Havranek, Z. Sidak, and M. Novak, editors, *Proceedings of the 6th Prague Symposium on Computational Statistics, COMPSTAT'84*, pages 154–159. Physica-Verlag, Wien, AUS, 1984.

[189] D. Petz. Monotone metrics on matrix spaces. *Linear Algebra and its Applications*, 244:81–96, Sept. 1996.

[190] D. Petz and R. Temesi. Means of positive numbers and matrices. *SIAM Journal on Matrix Analysis and Applications*, 27(3):712–720, 2005.

[191] D.-T. Pham, F. Vrins, and M. Verleysen. On the risk of using Rényi's entropy for blind source separation. *IEEE Transactions on Signal Processing*, 56(10):4611–4620, Oct. 2008.

[192] J. P. W. Pluim, J. B. A. Maintz, and M. A. Viergever. f-information measures in medical image registration. *IEEE Transactions on Medical Imaging*, 23(12):1508–1516, Dec. 2004.

[193] Y. Qiao and N. Minematsu. A study on invariance of f-divergence and its application to speech recognition. *IEEE Transactions on Signal Processing*, 58(7):3884–3890, July 2010.

[194] F. Ramponi, A. Ferrante, and M. Pavon. A globally convergent matricial algorithm for multivariate spectral estimation. *IEEE Transactions on Automatic Control*, 54(10):2376–2388, Oct. 2009.

[195] C. R. Rao. Information and accuracy attainable in the estimation of statistical parameters. *Bulletin of the Calcutta Mathematical Society*, 37:81–91, 1945.

[196] C. R. Rao. Diversity and dissimilarity coefficients: A unified approach. *Theoretical Population Biology*, 21(1):24–43, Feb. 1982.

[197] C. R. Rao. Diversity: its measurement, decomposition, apportionment and analysis. *Sankhyā: The Indian Journal of Statistics - Series A*, 44(1):1–22, 1982.

[198] C. R. Rao. Rao's axiomatization of diversity measures. In S. Kotz and N. L. Johnson, editors, *Encyclopedia of Statistical Sciences*, volume 7, pages 614–617. John Wiley & Sons Ltd, 1986.

[199] C. R. Rao. Differential metrics in probability spaces. In S.-I. Amari, O. E. Barndorff-Nielsen, R. E. Kass, S. L. Lauritzen, and C. R. Rao, editors, *Differential Geometry in Statistical Inference*, volume 10 of *Lecture Notes - Monograph Series*. Institute of Mathematical Statistics, Hayward, CA, USA, 1987.

[200] C. R. Rao and T. Nayak. Cross entropy, dissimilarity measures, and characterizations of quadratic entropy. *IEEE Transactions on Information Theory*, 31(5):589–593, Sept. 1985.

[201] J. Rauh. Finding the maximizers of the information divergence from an exponential family, Dec. 2009. arXiv.

[202] P. Ravikumar, A. Agarwal, and M. J. Wainwright. Message-passing for graph-structured linear programs: Proximal methods and rounding schemes. *Journal of Machine Learning Research*, 11:1043–1080, Mar. 2010.

[203] T. Read and N. Cressie. *Goodness-of-Fit Statistics for Discrete Multivariate Data*. Statistics. Springer, NY, 1988.

[204] M. D. Reid and R. C. Williamson. Information, divergence and risk for binary experiments, Jan. 2009. arXiv.

[205] M. D. Reid and R. C. Williamson. Composite binary losses. *Journal of Machine Learning Research*, pages 2387–2422, Sept. 2010.

[206] A. Rényi. On measures of information and entropy. In J. Neyman, editor, *Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 547–561. University of California Press, Berkeley, CA, USA, 1961.

[207] A. Rényi. On some basic problems of statistics from the point of view of information theory. In L. M. Le Cam and J. Neyman, editors, *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 531–543. University of California Press, Berkeley, CA, USA, 1967.

[208] W. Sander. Measures of information. In E. Pap, editor, *Handbook of Measure Theory*, volume 2, pages 1523–1565. North-Holland, Amsterdam, NL, 2002.

[209] R. Santos-Rodriguez, D. Garcia-Garcia, and J. Cid-Sueiro. Cost-sensitive classification based on Bregman divergences for medical diagnosis. In M. A. Wani, editor, *Proceedings of the 8th International Conference on Machine Learning and Applications (ICMLA'09), Miami Beach, Fl., USA, December 13-15, 2009*, pages 551–556, 2009.

[210] M. P. Schützenberger. *Contribution aux Applications Statistiques de la Théorie de l'Information*. Thèse d'État, Inst. Stat. Univ. Paris, 1953. In French.

[211] F. C. Schweppe. On the Bhattacharyya distance and the divergence between Gaussian processes. *Information and Control*, 11(4):373–395, Oct. 1967.

[212] F. C. Schweppe. State space evaluation of the Bhattacharyya distance between two Gaussian processes. *Information and Control*, 11(3):352–372, Sept. 1967.

[213] J. E. Shore and R. M. Gray. Minimum cross-entropy pattern classification and cluster analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 4(1):11–17, Jan. 1982.

[214] S. Si, D. Tao, and B. Geng. Bregman divergence-based regularization for transfer subspace learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(7):929–942, July 2010.

[215] R. Sibson. Information radius. *Probability Theory and Related Fields*, 14(2):149–160, June 1969.

[216] B. K. Sriperumbudur, A. Gretton, K. Fukumizu, G. R. G. Lanckriet, and B. Schölkopf. On integral probability metrics, ϕ-divergences and binary classification, Jan. 2009. arXiv.

[217] S. Srivastava, M. R. Gupta, and B. A. Frigyik. Bayesian quadratic discriminant analysis. *Journal of Machine Learning Research*, 8:1277–1305, June 2007.

[218] F. Österreicher and I. Vajda. A new class of metric divergences on probability spaces and its applicability in statistics. *Annals of the Institute of Statistical Mathematics*, 55(3):639–653, Sept. 2003.

[219] A. A. Stoorvogel and J. H. van Schuppen. Approximation problems with the divergence criterion for Gaussian variables and Gaussian processes. *Systems and Control Letters*, 35(4):207–218, Nov. 1998.

[220] W. Stummer and I. Vajda. On Bregman distances and divergences of probability measures, Nov. 2009. arXiv.

[221] W. Stummer and I. Vajda. On divergences of finite measures and their applicability in statistics and information theory. *Statistics - A Journal of Theoretical and Applied Statistics*, 44(2):169–187, Apr. 2010.

[222] I. Sutskever and T. Tieleman. On the convergence properties of contrastive divergence. In Y. W. Teh and M. Titterington, editors, *Proceedings of the 13th International Workshop on Artificial Intelligence and Statistics (AISTATS'10), Chia Laguna, Sardinia, Italy, May 13-15, 2010*, pages 78–795, 2010.

[223] I. J. Taneja. On generalized information measures and their applications. *Advances in Electronics and Electron Physics*, 76:327–413, 1989.

[224] I. J. Taneja. *Generalized Information Measures and Their Applications*. 2001. Online.

[225] B. Taskar, S. Lacoste-Julien, and M. I. Jordan. Structured prediction, dual extragradient and Bregman projections. *Journal of Machine Learning Research*, 7:1627–1653, July 2006.

[226] M. Teboulle. A unified continuous optimization framework for center-based clustering methods. *Journal of Machine Learning Research*, 8:65–102, Jan. 2007.

[227] M. Teboulle, P. Berkhin, I. S. Dhillon, Y. Guan, and J. Kogan. Clustering with entropy-like k-means algorithms. In J. Kogan, C. Nicholas, and M. Teboulle, editors, *Grouping Multidimensional Data - Recent Advances in Clustering*, pages 127–160. Springer-Verlag, Berlin Heidelberg, 2006.

[228] A. Toma and M. Broniatowski. Dual divergence estimators and tests: robustness results, Dec. 2009. arXiv.

[229] F. Topsoe. Some inequalities for information divergence and related measures of discrimination. *IEEE Transactions on Information Theory*, 46(4):1602–1609, July 2000.

[230] E. Torgersen. *Comparison of Statistical Experiments*, volume 36 of *Encyclopedia of Mathematics and Its Applications*. Cambridge University Press, Cambridge, UK, 1991.

[231] J. Touboul. Projection pursuit through ϕ-divergence minimisation. *Entropy*, 12(6):1581–1611, June 2010.

[232] K. Tsuda, G. Rätsch, and M. K. Warmuth. Matrix exponentiated gradient updates for on-line learning and Bregman projection. *Journal of Machine Learning Research*, 6:995–1018, June 2005.

[233] M. Tsukada and 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*, pages 149–153, 2009.

[234] I. Vajda. χ^α-divergence and generalized Fisher's information. In J. Kozesnik, editor, *Transactions of the 6th Prague Conference on Information Theory, Statistical Decision Functions, Random Processes, Prague, September 19-25, 1971*, pages 873–886. Academia, Prague, 1973.

[235] I. Vajda. *Theory of Statistical Inference and Information*, volume 11 of *Series B: Mathematical and Statistical Methods*. Kluwer Academic Publishers, Dordrecht, 1989.

[236] 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, Nov. 2008.

[237] I. Vajda. On metric divergences of probability measures. *Kybernetika*, 45(6):885–900, 2009.

[238] B. Vemuri, M. Liu, S. Amari, and F. Nielsen. Total Bregman divergence and its applications to DTI analysis. *IEEE Transactions on Medical Imaging*, 2010. To appear.

[239] C. Vignat, A. O. Hero, and 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*, pages 1822–1826, July 2006.

[240] F. Vrins, D.-T. Pham, and 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, and M. D. Plumbley, editors, *Proceedings of the 7th International Conference on Independent Component Analysis and Blind Source Separation (ICA'07), London, UK, September 9-12, 2007*, volume 4666 of *Lecture Notes in Computer Science*, pages 129–136. Springer-Verlag, Berlin Heidelberg, FRG, 2007.

[241] S. Wang and D. Schuurmans. Learning continuous latent variable models with Bregman divergences. In R. Gavaldà, K. P. Jantke, and E. Takimoto, editors, *Proceedings of the 14th International Conference on Algorithmic Learning Theory (ALT'03), Sapporo, Japan, October 17-19, 2003*, volume 2842 of *Lecture Notes in Artificial Intelligence*, pages 190–204. Springer-Verlag, Berlin Heidelberg, 2003.

[242] L. Wu, R. Jin, S. C.-H. Hoi, J. Zhu, and N. Yu. Learning Bregman distance functions and its application for semi-supervised clustering. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22, Vancouver, British Columbia, Canada, December 7-10, 2009*, pages 2089–2097. NIPS Foundation, 2009.

[243] R. W. Yeung. *A First Course in Information Theory*. Information Technology: Transmission, Processing and Storage. Springer-Verlag, 2002.

[244] R. W. Yeung. *Information Theory and Network Coding*. Information Technology: Transmission, Processing and Storage. Springer-Verlag, 2008.

[245] W. Yin, S. Osher, D. Goldfarb, and J. Darbon. Bregman iterative algorithms for ℓ_1-minimization with applications to compressed sensing. *SIAM Journal on Imaging Sciences*, 1(1):143–168, 2008.

[246] S. Yu and P. G. Mehta. The Kullback-Leibler rate pseudo-metric for comparing dynamical systems. *IEEE Transactions on Automatic Control*, 55(7):1585–1598, July 2010.

[247] R. G. Zaripov. *New Measures and Methods in Information Theory*. A. N. Tupolev State Technical University Press, Kazan, Tatarstan, 2005. In Russian - Online.

[248] J. Zhang. Divergence function, duality, and convex analysis. *Neural Computation*, 16(1):159–195, Jan. 2004.

[249] J. Ziv and M. Zakai. On functionals satisfying a data-processing theorem. *IEEE Transactions on Information Theory*, 19(3):275–283, May 1973.