THE **CHORD GAP DIVERGENCE** AND A GENERALIZATION OF THE BHATTACHARYYA DISTANCE

— THE CHORD JENSEN DIVERGENCE —

Frank Nielsen

Sony Computer Science Laboratories Inc, Japan
École Polytechnique, France

@FrnkNlsn

18th April 2018
— ICASSP —

Outline of the talk

- Background on divergences:
  Statistical divergences versus *parameter* divergences
- Definition of the chord gap divergence and review of its properties
- Chord gap divergence yields a generalization of the renown Burbea-Rao divergence/Jensen divergences [4].
  Used as a distance in matrix signal processing [7, 10, 5, 12]
  (as known as Jensen-Bregman LogDet, JBLD)
- Center-based k-means(++) clustering with respect to the chord gap divergence
- Concluding remarks and perspectives

Background on statistical and parameter divergences

- In statistics, divergence = *distortion measure* between *probability measures*. E.g., Kullback-Leibler (KL) divergence/deviance (= relative entropy in IT):

$$\mathrm{KL}[p:q] := \int_{\mathcal{X}} p(x) \log \frac{p(x)}{q(x)} \mathrm{d}\mu(x)$$

- In information geometry [1], divergence = smooth dissimilarity measure between *parameters*: $D(\theta : \theta') \geq 0$ with equality iff $\theta = \theta'$. Non-metric measure when it violates the triangle inequality. E.g., Bregman divergence for a strictly convex and smooth generator F:

$$B_F(\theta : \theta') := F(\theta) - F(\theta') - (\theta - \theta')^\top \nabla F(\theta')$$

- Potential confusion: BD for $F(\theta) = \sum_i \theta_i \log \theta_i$ yields *discrete* $\mathrm{KL}[p:q] = \sum_i p_i \log \frac{p_i}{q_i} + q_i - p_i = B_F(p:q)$ extended to discrete positive measures. On the probability simplex, $\mathrm{KL}[p:q] = \mathrm{KL}(p:q)$.

Principled parametric statistical divergences

- Statistical divergences on parametric models $\mathcal{F} = \{p_\theta\}$ amount to an equivalent parameter divergence:

$$D_{\mathcal{F}}(\theta : \theta') := D[p_\theta : p_{\theta'}]$$

- Principled statistical divergences: Invariant *f-divergences* (including KL for $f(u) = -\log u$) in information geometry

$$I_f[p : q] := \int_{\mathcal{X}} p(x) f\left(\frac{q(x)}{p(x)}\right) \mathrm{d}\mu(x)$$

  Invariance by Markov kernel on sample space and *information monotonicity* when $Y = T(X)$ [1]:

$$I_f[p_Y : q_Y] \leq I_f[p_X : q_X]$$

- Parametric families of divergences useful in practice for fine tuning performance in applications (increase DOFs).

Parameter divergence families from convex generators

- Skew Jensen divergences [4, 13, 6] (Burbea-Rao divergences [4]) for a **strictly convex** function F:

$$\boxed{J_F^\alpha(\theta : \theta') {:=} (F(\theta) F(\theta'))_\alpha - F((\theta\theta')_\alpha)}$$

where $(\theta\theta')_\lambda {:=} (1-\lambda)\theta + \lambda\theta' = \theta + \lambda(\theta' - \theta)$.

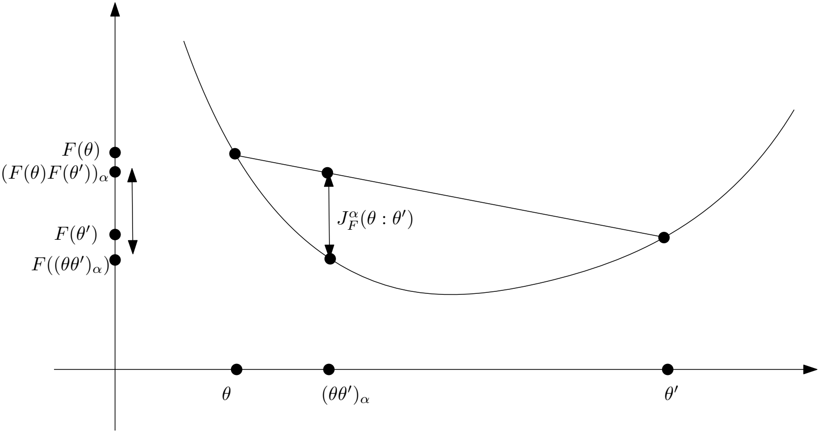
- Related *asymptotically* to Bregman divergences [3, 2]:

$$\lim_{\alpha \to 0+} \frac{1}{\alpha(1-\alpha)} J_F^\alpha(\theta : \theta') = B_F(\theta' : \theta)$$

$$\lim_{\alpha \to 1-} \frac{1}{\alpha(1-\alpha)} J_F^\alpha(\theta : \theta') = B_F(\theta : \theta')$$

Geometric interpretation: Skew Jensen inequality gap

$$J_F^\alpha(\theta : \theta') := (F(\theta)F(\theta'))_\alpha - F((\theta\theta')_\alpha)$$



Can be generalized to (M, N)-convexity [11]: $(\theta\theta')_\alpha = M_{1-\alpha}(\theta : \theta')$ and $(F(\theta)F(\theta'))_\alpha = N_{1-\alpha}(F(\theta) : F(\theta'))$. Usual skew Jensen divergence is for M=N=A, the weighted Arithmetic mean.

Statistical distances on parametric families

- $\mathcal{F} = \{p(x;\theta)\}$ **exponential family** [9] with density $p_\theta(x) := \exp(\theta^\top x - F(\theta))$
(include Gaussian, Gamma/Beta, Poisson, etc.)
- Statistical skew Bhattacharrya divergences [7]:

$$\begin{aligned} \mathrm{Bhat}_\alpha[p:q] &:= -\log \int p^{1-\alpha}(x) q^\alpha(x) \mathrm{d}\mu(x) \\ \mathrm{Bhat}_\alpha[p_{\theta_1}:p_{\theta_2}] &= J_F^\alpha(\theta_1:\theta_2) = J_F^{1-\alpha}(\theta_2:\theta_1). \end{aligned}$$

- Asymptotic cases (general/exponential families):

$$\begin{aligned} \lim_{\alpha\rightarrow 0^+} \frac{1}{\alpha(1-\alpha)} \mathrm{Bhat}_\alpha[p:q] &= \mathrm{KL}[p:q] \\ \lim_{\alpha\rightarrow 1^-} \frac{1}{\alpha(1-\alpha)} \mathrm{Bhat}_\alpha[p:q] &= \mathrm{KL}[q:p] \end{aligned}$$

$$\begin{aligned} \lim_{\alpha\rightarrow 0^+} \frac{1}{\alpha(1-\alpha)} \mathrm{Bhat}_\alpha(p_\theta:p_{\theta'}) &= B_F(\theta':\theta) \\ \lim_{\alpha\rightarrow 1^-} \frac{1}{\alpha(1-\alpha)} \mathrm{Bhat}_\alpha(p_\theta:p_{\theta'}) &= B_F(\theta:\theta') \end{aligned}$$

Relationships between statistical/parameter divergences

Relationships between statistical distances and parameter divergences when the distributions belong to the *same* exponential family.

Statistical divergences

Parameter divergences

$\mathrm{Bhat}_\alpha[p:q] = -\log \int p(x)^{1-\alpha} q(x)^\alpha \mathrm{d}\mu(x)$

$J_F^\alpha(\theta_p : \theta_q) = (F(\theta_p)F(\theta_q))_\alpha - F((\theta_p\theta_q)_\alpha)$

$\lim_{\alpha\rightarrow 0+} \frac{1}{\alpha}\mathrm{Bhat}_\alpha[p:q]$

$p(x) = p(x;\theta_p)$

$q(x) = p(x;\theta_q)$

$\lim_{\alpha\rightarrow 0+} \frac{1}{\alpha}\mathrm{J}_F^\alpha(\theta:\theta')$

$\mathrm{KL}[p:q] = \int p(x)\log\frac{p(x)}{q(x)}\mathrm{d}\mu(x)$

$B_F(\theta_q:\theta_p) = F(\theta_q) - F(\theta_p) - (\theta_q-\theta_p)^\top \nabla F(\theta_q)$

Generic distributions

Exponential families

$p(x;\theta) = \exp(\theta^\top x - F(\theta))$

Skew Jensen-Bregman divergence

Skew Jensen divergence rewritten as a skew Jensen-Bregman divergence

Skew Jensen-Bregman (JB) divergence [6] (inspired by statistical Jensen-Shannon divergence):

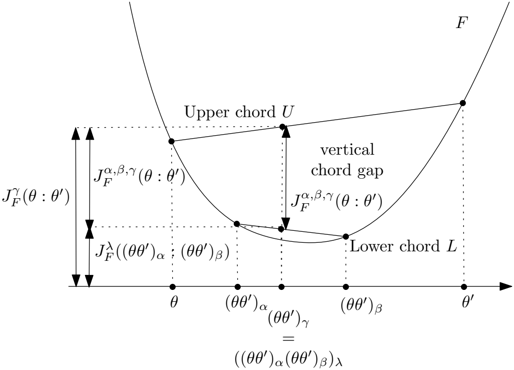
$$\boxed{\mathrm{JB}_F^\alpha(\theta:\theta'):=(1-\alpha)B_F(\theta:(\theta\theta')_\alpha)+\alpha B_F(\theta':(\theta\theta')_\alpha)}$$

$$\mathrm{JB}_F^\alpha(\theta:\theta')=J_F^\alpha(\theta:\theta')$$

$\Rightarrow$ since $\theta-(\theta\theta')_\alpha=\alpha(\theta-\theta')$ and $\theta'-(\theta\theta')_\alpha=(1-\alpha)(\theta'-\theta)$, the gradient terms $\nabla F((\theta\theta')_\alpha)$ in the Bregman divergences canceled out!

The novel triparametric chord gap divergence

Vertical distances between an *outer upper chord U* and *inner lower chord L* is always non-negative:

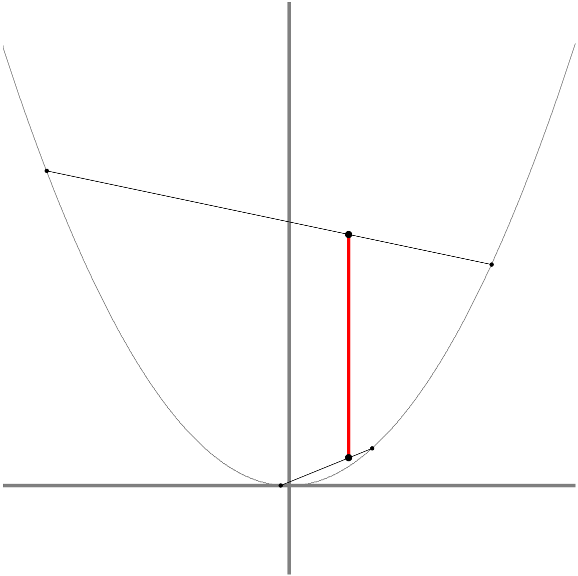


The chord gap divergence induced by a strictly convex function F is defined for $\alpha, \beta \in [0,1]$ and $\gamma \in (\alpha, \beta)$ as

$$J_F^{\alpha,\beta,\gamma}(\theta:\theta') = (F(\theta)F(\theta'))_\gamma - (F((\theta\theta')_\alpha)F((\theta\theta')_\beta))_{\frac{\gamma-\alpha}{\beta-\alpha}}$$

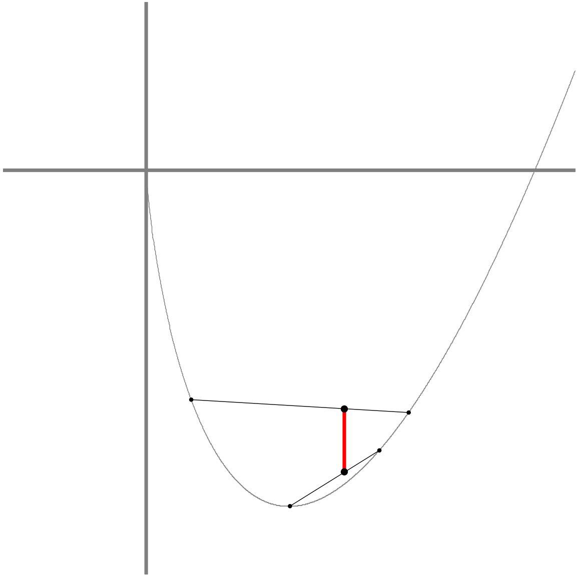
Chord gap divergence: Quadratic generator

$F(\theta) = \frac{1}{2}\sum_i \theta_i^2$ (BD = half squared Euclidean distance)



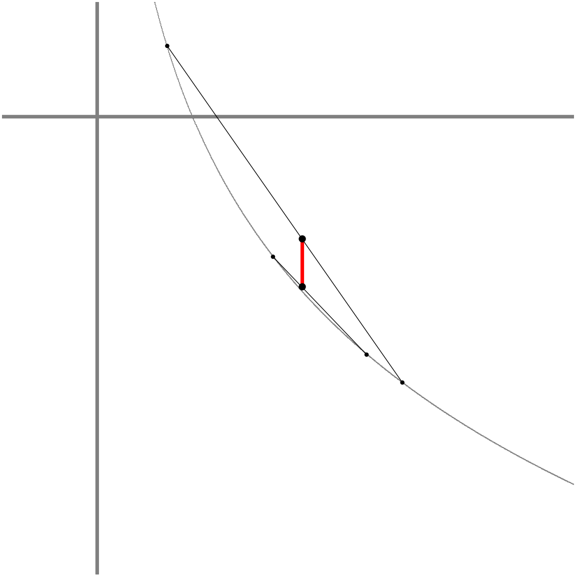
Chord gap divergence: Shannon information

$F(\theta) = \sum_i \theta_i \log \theta_i$ (BD = extended KL, F = negentropy)



Chord gap divergence: Burg information generator

$F(\theta) = -\sum_i \log \theta_i$ (BD = Itakura-Saito divergence)



Some basic properties of the chord gap parameter divergence

- Generalization of skew Jensen divergence:

$$J_F^{\alpha,\alpha,\alpha}(\theta:\theta') = J_F^{\alpha}(\theta:\theta')$$

(visually speaking, lower chord collapses to a point) For $\alpha = 0$, $\beta = 1$, we have $\lambda = \gamma$, and we also recover the skew γ-Jensen divergence.

- Reference duality ($\theta \leftrightarrow \theta'$):

$$J_F^{\alpha,\beta,\gamma}(\theta':\theta) = J_F^{1-\alpha,1-\beta,1-\gamma}(\theta:\theta')$$

In particular $J_F^{1-\alpha,1-\alpha,1-\alpha}(\theta:\theta') = J_F^{\alpha}(\theta':\theta)$

- Interpreted as the difference of two skew Jensen divergences:

$$\boxed{J_F^{\alpha,\beta,\gamma}(\theta:\theta') = J_F^{\gamma}(\theta:\theta') - J_F^{\lambda}((\theta\theta')_\alpha:(\theta\theta')_\beta)}$$

with $\lambda = \frac{\gamma-\alpha}{\beta-\alpha}$ (i.e., $\gamma = \lambda(\beta-\alpha)+\alpha$).
$\Rightarrow$ **Chord Jensen Divergence**

A biparametric subfamily of chord gap divergences

Consider $\alpha = 0$ so that $(\theta\theta')_\alpha = \theta$.
Then upper & lower chords coincide at extremity $(\theta, F(\theta))$.

$$
\begin{aligned}
J_F^{\beta,\gamma}(\theta:\theta') &= (F(\theta)F(\theta'))_\gamma - (F(\theta)F(\theta\theta'_\beta))_{\frac{\gamma}{\beta}},\\
&= \left(\frac{\gamma}{\beta} - \gamma\right) F(\theta) + \gamma F(\theta') - \frac{\gamma}{\beta} F((\theta\theta')_\beta),\\
&= \gamma \left(\left(\frac{1}{\beta} - 1\right) F(\theta) + F(\theta') - \frac{1}{\beta} F((\theta\theta')_\beta) \right)
\end{aligned}
$$

In particular, when $\beta = \frac{1}{2}$:

$$
J_F^\gamma(\theta:\theta') = 2\gamma \left(\frac{F(\theta)+F(\theta')}{2} - F\left(\frac{\theta+\theta'}{2}\right)\right)
$$

$=$ ordinary (scaled) Jensen divergence.

When $\beta \to 0$, $\lim_{\beta\to 0} \frac{1}{\gamma} J_F^{\beta,\gamma}(\theta:\theta') = B_F(\theta':\theta)$ (with $\gamma \in (0,\beta)$)

Generalization of the statistical Bhattacharyya divergence

- First, let us use the equivalence of chord gap divergence (difference of two skew Jensen divergences) with the statistical Bhattacharrya divergences between distributions of a same exponential family:

$$\mathrm{Bhat}^{\alpha,\beta,\gamma}[p_\theta : p_{\theta'}] = -\log \frac{\int p^{1-\gamma}(x;\theta)p^{\gamma}(x;\theta')\mathrm{d}\mu(x)}{\int p^{1-\lambda}(x;(\theta\theta')_\alpha)p^{\lambda}(x;(\theta\theta')_\beta)\mathrm{d}\mu(x)}$$

- Then relax/extrapolate the definition to arbitrary densities: (need to normalize distributions on Bhattacharyya arcs)

$$\mathrm{Bhat}^{\alpha,\beta,\gamma}[p:q] :=$$

$$-\log\left(\frac{\int p(x)^{1-\gamma}q(x)^{\gamma}\mathrm{d}\mu(x)}{\int\left(\frac{p(x)^{1-\alpha}q(x)^{\alpha}}{\int p(x)^{1-\alpha}q(x)^{\alpha}\mathrm{d}\mu(x)}\right)^{1-\lambda}\left(\frac{p(x)^{1-\beta}q(x)^{\beta}}{\int p(x)^{1-\beta}q(x)^{\beta}\mathrm{d}\mu(x)}\right)^{\lambda}\mathrm{d}\mu(x)}\right)$$

Clustering: Centroid wrt. to the chord gap divergence

- The centroid of n parameter $\{\theta_1, \ldots, \theta_n\}$ is defined as the minimizer of

$$\min_{\theta} \sum_{i=1}^{n} J_F^{\alpha,\beta,\gamma}(\theta_i : \theta)$$

- Express the function using a difference of convex functions
- Iteratively optimize using the Concave-Convex Procedure (CCCP): $\theta^{(t+1)} = \nabla F^{-1}\left(\frac{1}{\gamma}\sum_i w_i((1-\lambda)\alpha\nabla F((\theta_i\theta^{(t)})_\alpha) + \lambda\beta\nabla F((\theta_i\theta^{(t)})_\beta))\right)$
- Guaranteed to converge [6] to a (local) minimum.

But no need to compute centroids with k-means++ initialization!

Guaranteed probabilistic initialization of k-means++

By pass the centroid computations in k-means that minimizes loss function

$$\sum_{i=1}^{n} \min_{j\in[k]} D(\theta_i : C_j)$$

For a general divergence D, to get an expected competitive ratio of $2U^2(1+V)(2+\log k)$, we need to bound [8]:

- U such that the divergence $D = J_F^{\alpha,\beta\gamma}$ satisfies the U-triangular inequality:

$$D(x:z) \leq U(D(x:y) + D(y:z))$$

For any squared Mahalanobis distance $D_Q(\theta,\theta'){:=}(\theta'-\theta)^\top Q(\theta'-\theta)$ (with $Q \succ 0$), we have $\boxed{U=2}$.

- V such that the divergence satisfies the symmetric inequality:

$$D(y:x) \leq V\, D(x:y)$$

Bounding U and V for the chord gap divergence

Using Jensen-Bregman divergence and the Lagrange remainder of first-order Taylor expansion of Bregman divergences

$$J_F^\alpha(\theta : \theta') = (1-\alpha) B_F(\theta : (\theta\theta')_\alpha) + \alpha B_F(\theta' : (\theta\theta')_\alpha)$$

We get

$$\boxed{J_F^\alpha(\theta : \theta') = (\theta' - \theta)^\top H_\alpha(\theta : \theta')(\theta' - \theta)}$$

with

$$H_\alpha(\theta : \theta') = \frac{\alpha(1-\alpha)}{2}(\alpha \nabla^2 F(\xi_1) + (1-\alpha)\nabla^2 F(\xi_2)) \succ 0,$$

$\xi_1 \in [\theta(\theta\theta')_\alpha]$ and $\xi_2 \in [(\theta\theta')_\alpha \theta']$

Chord Jensen Divergence as a squared Mahalanobis distance

Since we have $(\theta\theta')_\alpha - (\theta\theta')_\beta = (\alpha-\beta)(\theta'-\theta)$, it follows that $J_F^\lambda((\theta\theta')_\alpha : (\theta\theta')_\beta) = (\alpha-\beta)^2(\theta'-\theta)^\top H_\lambda(\theta',\theta)$

Finally, from the difference of two skew Jensen divergences, it follows that the squared Mahalanobis expression ($U=2$)

$$\boxed{J_F^{\alpha,\beta,\gamma}(\theta:\theta') = \frac{1}{2}(\theta'-\theta)^\top H_F^{\alpha,\beta,\gamma}(\theta:\theta')(\theta'-\theta)}$$

$$\begin{aligned} H_F^{\alpha,\beta,\gamma}(\theta:\theta') &= \frac{1}{2}\gamma(1-\gamma)\nabla^2 F(\xi') - \frac{1}{2}\lambda(1-\lambda)(\alpha-\beta)^2\nabla^2 F(\xi'') \\ &= \frac{1}{2}\left(\gamma(1-\gamma)\nabla^2 F(\xi') - (\gamma-\alpha)(\gamma-\beta)\nabla^2 F(\xi'')\right) \end{aligned}$$

Therefore, we bound $V \leq \rho$ for $\mathcal{P} = \{\theta_i\}$ (co: convex hull) with

$$\rho = \frac{\sup_{\xi',\xi'',\theta,\theta'\in\mathrm{co}(\mathcal{P})} \|(\nabla^2 F(\xi'))^{\frac{1}{2}}(\theta'-\theta)\|}{\inf_{\xi',\xi'',\theta,\theta'\in\mathrm{co}(\mathcal{P})} \|(\nabla^2 F(\xi''))^{\frac{1}{2}}(\theta'-\theta)\|} < \infty$$

and the chord gap divergence k-means++ yields a guaranteed probabilistic initialization

Summary and perspectives

- Statistical divergences $D[p_\theta : p_{\theta'}]$ on families of parametric probabilities $\mathcal{F} = \{p_\theta\}$ amount to equivalent parametric divergences $D_\mathcal{F}(\theta : \theta')$
- For exponential families, link between skew Jensen parameter divergences and skew Bhattacharrya statistical divergences (and Bregman divergence with Kullback-Leibler divergence asymptotically)
- Parameter divergences can be geometrically constructed from a convex function by taking vertical gaps in the function graph
- The chord gap divergence is an extension of the skew Jensen/Burbea-Rao divergence by taking the **vertical gap between an upper chord and a lower chord**. Can be expressed as the difference of two skew Jensen gap divergences
- Perspective: Demonstrate its usefulness in applications like clustering or statistical inference.

More in the paper and in arXiv:1709.10498

References I

S.-i. Amari.
Information geometry and its applications.
Springer, 2016.

A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh.
Clustering with Bregman divergences.
Journal of machine learning research, 6(Oct):1705–1749, 2005.

Lev 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.

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, 1982.

Anoop Cherian, Suvrit Sra, Arindam Banerjee, and Nikolaos Papanikolopoulos.
Jensen-Bregman logdet divergence with application to efficient similarity search for covariance matrices.
IEEE transactions on pattern analysis and machine intelligence, 35(9):2161–2174, 2013.

F. Nielsen and S. Boltz.
The Burbea-Rao and Bhattacharyya centroids.
IEEE Transactions on Information Theory, 57(8):5455–5466, 2011.

F. Nielsen and R. Nock.
Skew Jensen-Bregman Voronoi diagrams.
Trans. Computational Science, 14:102–128, 2011.

References II

F. Nielsen and R. Nock.
Total Jensen divergences: definition, properties and clustering.
In *IEEE ICASSP*, pages 2016–2020, 2015.

Frank Nielsen and Vincent Garcia.
Statistical exponential families: A digest with flash cards.
CoRR, abs/0911.4863, 2009.

Frank Nielsen, Meizhu Liu, Xiaojing Ye, and Baba C Vemuri.
Jensen divergence based SPD matrix means and applications.
In *Pattern Recognition (ICPR), 2012 21st International Conference on*, pages 2841–2844. IEEE, 2012.

Frank Nielsen and Richard Nock.
Generalizing skew Jensen divergences and Bregman divergences with comparative convexity.
IEEE signal processing letters, 24(8):1123–1127, 2017.

Hui Song, Wen Yang, Xin Xu, and Mingsheng Liao.
Unsupervised PolSAR imagery classification based on Jensen-Bregman logdet divergence.
In *EUSAR 2014; 10th European Conference on Synthetic Aperture Radar; Proceedings of*, pages 1–4. VDE, 2014.

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