Dissimilarities, divergences, and distances

Frank Nielsen
Sony Computer Science Laboratories Inc
Tokyo, Japan

13th August 2021

This is a working document which will be frequently updated with materials concerning the discrepancy between two distributions.

1 Statistical distances between densities with computationally intractable normalizers

Consider a density $p(x) = \frac{\tilde{p}(x)}{Z_p}$ where $\tilde{p}(x)$ is an unnormalized *computable* density and $Z_p = \int p(x)\mathrm{d}\mu(x)$ the *computationally intractable* normalizer (also called in statistical physics the partition function or free energy). A statistical distance $D[p_1 : p_2]$ between two densities $p_1(x) = \frac{\tilde{p}_1(x)}{Z_{p_1}}$ and $p_2(x) = \frac{\tilde{p}_2(x)}{Z_{p_2}}$ with computationally intractable normalizers Z_{p_1} and Z_{p_2} is said *projective* (or two-sided *homogeneous*) if and only if

$$\forall \lambda_1 > 0, \lambda_2 > 0, \quad D[p_1 : p_2] = D[\lambda_1 p_1 : \lambda_2 p_2].$$

In particular, letting $\lambda_1 = Z_{p_1}$ and $\lambda_2 = Z_{p_2}$, we have

$$D[p_1 : p_2] = D[\tilde{p}_1 : \tilde{p}_2].$$

Notice that the rhs. does not rely on the computationally intractable normalizers. These projective distances are useful in statistical inference based on minimum distance estimators [2] (see next Section).

Here are a few statistical projective distances:

- **γ-divergences** ($\gamma > 0$) [5, 3]:

$$D_\gamma[p : q] := \log\left(\int_{\mathbb{R}} q^{\alpha+1}\right) - \left(1 + \frac{1}{\alpha}\right)\log\left(\int_{\mathbb{R}} q^\alpha p\right) + \frac{1}{\alpha}\log\left(\int_{\mathbb{R}} p^{\alpha+1}\right), \quad \gamma \geq 0$$

  When $\gamma \to 0$, we have [3] $D_\gamma[p : q] = D_{\mathrm{KL}}[p : q]$, the Kullback-Leibler divergence (KLD). For example, we can estimate the KLD between two densities of an exponential-polynomial family by Monte Carlo stochastic integration of the γ-divergence for a small value of γ [9].

The γ-divergences (projective, Bregman-type) and the density power divergence [1] (non-projective, Bregman-type divergence):

$$D_\alpha^{\mathrm{dpd}}[p:q] := \int_{\mathbb{R}} q^{\alpha+1} - \left(1+\frac{1}{\alpha}\right)\int_{\mathbb{R}} q^\alpha p + \frac{1}{\alpha}\int_{\mathbb{R}} p^{\alpha+1}, \quad \alpha \geq 0,$$

can be encapsulated into the family of Φ-power divergences [12] (functional density power divergence class):

$$D_{\phi,\alpha}[p:q] := \phi\left(\int_{\mathbb{R}} q^{\alpha+1}\right) - \left(1+\frac{1}{\alpha}\right)\phi\left(\int_{\mathbb{R}} q^\alpha p\right) + \frac{1}{\alpha}\phi\left(\int_{\mathbb{R}} p^{\alpha+1}\right), \quad \alpha \geq 0,$$

where $\phi(e^x)$ convex and strictly increasing, ϕ continuous and twice continously differentiable with finite second order derivatives. We have $D_{\phi,0}[p:q] = \phi'(1)\int_{\mathbb{R}} p(x)\log\frac{p(x)}{q(x)}\mathrm{d}\mu(x) = \phi'(1)D_{\mathrm{KL}}[p:q]$.

- **Cauchy-Schwarz divergence** [4] (CSD, projective)

$$D_{\mathrm{CS}}[p:q] = -\log\left(\frac{\int p(x)q(x)\mathrm{d}\mu(x)}{\sqrt{\int p(x)^2\mathrm{d}\mu(x)\int q(x)^2\mathrm{d}\mu(x)}}\right) = D_{\mathrm{CS}}[\lambda_1 p : \lambda_2 q], \forall\lambda_1>0, \lambda_2>0,$$

and **Hölder divergences** [11] (HD, projective, which generalizes the CSD):

$$D_{\alpha,\gamma}^{\text{Hölder}}[p:q] = -\log\left(\frac{\int_{\mathcal{X}} p(x)^{\gamma/\alpha} q(x)^{\gamma/\beta}\mathrm{d}x}{\left(\int_{\mathcal{X}} p(x)^\gamma \mathrm{d}x\right)^{1/\alpha}\left(\int_{\mathcal{X}} q(x)^\gamma \mathrm{d}x\right)^{1/\beta}}\right), \quad \frac{1}{\alpha}+\frac{1}{\beta}=1.$$

We have

$$\forall\lambda_1>0, \lambda_2>0, D_{\alpha,\gamma}^{\text{Hölder}}[\lambda_1 p : \lambda_2 q] = D_{\alpha,\gamma}^{\text{Hölder}}[p:q],$$

and

$$D_{2,2}^{\text{Hölder}}[p:q] = D_{\mathrm{CS}}[p:q].$$

Hölder divergences between two densities p_{θ_p} and p_{θ_q} of an exponential family with cumulant function $F(\theta)$ is available in closed-form [11]:

$$D_{\alpha,\gamma}^{\text{Hölder}}[p:q] = \frac{1}{\alpha}F(\gamma\theta_p) + \frac{1}{\beta}F(\gamma\theta_q) - F\left(\frac{\gamma}{\alpha}\theta_p + \frac{\gamma}{\beta}\theta_q\right)$$

The CSD is available in closed-form between mixtures of an exponential family with a conic natural parameter [7]: This includes the case of Gaussian mixture models [6].

- **Hilbert distance** [10] (projective): Consider two probability mass functions $p=(p_1,\ldots,p_d)$ and $q=(q_1,\ldots,q_d)$ of the d-dimensional probability simplex. Then the Hilbert distance is

$$D^{\text{Hilbert}}[p:q] = \log\left(\frac{\max_{i\in\{1,\ldots,d\}}\frac{p_i}{q_i}}{\min_{j\in\{1,\ldots,d\}}\frac{p_j}{q_j}}\right).$$

We have

$$\forall\lambda_1>0, \lambda_2>0, D^{\text{Hilbert}}[\lambda_1 p : \lambda_2 q] = D^{\text{Hilbert}}[p:q].$$

2 Statistical distances between empirical distributions and densities with computationally intractable normalizers

When estimating the parameter $\hat{\theta}$ for a parametric family of distributions $\{p_\theta\}$ from i.i.d. observations $\mathcal{S} = \{x_1, \ldots, x_n\}$, we can define a minimum distance estimator:

$$\hat{\theta} = \arg\min_{\theta} D[p_{\mathcal{S}} : p_\theta],$$

where $p_{\mathcal{S}} = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ is the empirical distribution (normalized). Thus we need only a right-sided projective divergence to estimate models with computationally intractable normalizers.

- **Hyvärinen divergence** [] (also called **Fisher divergence**):

$$D^{\text{Hyvärinen}}[p : p_\theta] := \frac{1}{2}\int \|\nabla_x \log p(x) - \nabla_x \log p_\theta(x)\|^2 \; p(x)\mathrm{d}x.$$

The Hyvarinen divergence has been extended for order-α Hyvarinen divergences [8] (for $\alpha > 0$):

$$D_\alpha^{\text{Hyvärinen}}[p : q] := \frac{1}{2}\int p(x)^\alpha \left(\nabla_x \log p(x) - \nabla_x \log q(x)\right)^2 \mathrm{d}x, \quad \alpha > 0.$$

This column is also available in pdf: filename `Distance.pdf`

Initially created 13th August 2021 (last updated August 13, 2021).

References

[1] Ayanendranath Basu, Ian R Harris, Nils L Hjort, and MC Jones. Robust and efficient estimation by minimising a density power divergence. *Biometrika*, 85(3):549–559, 1998.

[2] Ayanendranath Basu, Hiroyuki Shioya, and Chanseok Park. *Statistical inference: the minimum distance approach*. Chapman and Hall/CRC, 2019.

[3] Hironori Fujisawa and Shinto Eguchi. Robust parameter estimation with a small bias against heavy contamination. *Journal of Multivariate Analysis*, 99(9):2053–2081, 2008.

[4] Robert Jenssen, Jose C Principe, Deniz Erdogmus, and Torbjørn Eltoft. The Cauchy–Schwarz divergence and Parzen windowing: Connections to graph theory and Mercer kernels. *Journal of the Franklin Institute*, 343(6):614–629, 2006.

[5] MC Jones, Nils Lid Hjort, Ian R Harris, and Ayanendranath Basu. A comparison of related density-based minimum divergence estimators. *Biometrika*, 88(3):865–873, 2001.

[6] Kittipat Kampa, Erion Hasanbelliu, and Jose C Principe. Closed-form Cauchy-Schwarz PDF divergence for mixture of Gaussians. In *The 2011 International Joint Conference on Neural Networks*, pages 2578–2585. IEEE, 2011.

[7] Frank Nielsen. Closed-form information-theoretic divergences for statistical mixtures. In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR)*, pages 1723–1726. IEEE, 2012.

[8] Frank Nielsen. Fast approximations of the Jeffreys divergence between univariate Gaussian mixture models via exponential polynomial densities. *arXiv preprint arXiv:2107.05901*, 2021.

[9] Frank Nielsen and Richard Nock. Patch matching with polynomial exponential families and projective divergences. In *International Conference on Similarity Search and Applications*, pages 109–116. Springer, 2016.

[10] Frank Nielsen and Ke Sun. Clustering in Hilbert's projective geometry: The case studies of the probability simplex and the elliptope of correlation matrices. In *Geometric Structures of Information*, pages 297–331. Springer, 2019.

[11] Frank Nielsen, Ke Sun, and Stéphane Marchand-Maillet. On hölder projective divergences. *Entropy*, 19(3):122, 2017.

[12] Souvik Ray, Subrata Pal, Sumit Kumar Kar, and Ayanendranath Basu. Characterizing the functional density power divergence class. *arXiv preprint arXiv:2105.06094*, 2021.