On the Chi Square and Higher-Order Chi Distances for Approximating f-Divergences

Frank Nielsen and Richard Nock (IEEE Signal Processing Letters, Jan., 2014)

Zoltán Szabó

Gatsby Unit, Tea Talk

December 20, 2013

Outline

- Motivation: uncertainty, 'distance' between distributions.
- Exponential family.
- Analytical expressions.

Random variables: uncertainty

- Keyword: entropy.
- Example: Shannon entropy

$$H(p) = -\int p(u) \log p(u) \mathrm{d}u. \tag{1}$$

Random variables: uncertainty

- Keyword: entropy.
- Some examples ($\alpha \neq 1$, $\beta \neq 1$):

$$H(p) = -\int p(u) \log p(u) \mathrm{d}u, \tag{1}$$

$$H_{R,\alpha}(p) = \frac{1}{1-\alpha} \log \int p^{\alpha}(u) \mathrm{d}u, \tag{2}$$

$$H_{T,\alpha}(p) = \frac{1}{\alpha - 1} \left(1 - \int p^{\alpha}(u) \mathrm{d}u\right), \tag{3}$$

$$H_{SM,\alpha,\beta}(p) = \frac{1}{1-\beta} \left[\left(\int p^{\alpha}(u) \mathrm{d}u\right)^{\frac{1-\beta}{1-\alpha}} - 1\right]. \tag{4}$$

Random variables: entropy

- Relations:

$$\lim_{\alpha\to 1} H_{\mathrm{R},\alpha} = H, \qquad \lim_{\alpha\to 1} H_{\mathrm{T},\alpha} = H, \tag{5}$$

$$\lim_{\beta\to 1} H_{\mathrm{SM},\alpha,\beta} = H_{\mathrm{R},\alpha}, \qquad \lim_{\beta\to \alpha} H_{\mathrm{SM},\alpha,\beta} = H_{\mathrm{T},\alpha}, \tag{6}$$

$$\lim_{(\alpha,\beta)\to(1,1)} H_{\mathrm{SM},\alpha,\beta} = H. \tag{7}$$

- Quantity of interest:

$$I_\alpha(p) = \int p^\alpha(u)\mathrm{d}u. \tag{8}$$

Random variables: 'distance' of distributions

- Keyword: divergence.
- Example: Kullback-Leibler divergence

$$D(p, q) = \int p(u) \log \left[\frac{p(u)}{q(u)} \right] \mathrm{d}u. \tag{9}$$

Random variables: 'distance' of distributions

- Keyword: divergence.
- Some examples ($\alpha \neq 1$):

$$D(p,q) = \int p(u) \log\left[\frac{p(u)}{q(u)}\right] \mathrm{d}u, \tag{9}$$

$$D_{R,\alpha}(p,q) = \frac{1}{\alpha - 1} \log \int p^{\alpha}(u) q^{1-\alpha}(u) \mathrm{d}u, \tag{10}$$

$$D_{T,\alpha}(p,q) = \frac{1}{\alpha - 1} \left(\int p^{\alpha}(u) q^{1-\alpha}(u) \mathrm{d}u - 1 \right). \tag{11}$$

Random variables: divergence

- Some examples continued ($0 < \alpha \neq 1; \beta \neq 1$):

$$D_{SM,\alpha,\beta} = \frac{1}{\beta - 1}\left[\left(\int p^{\alpha}(u)q^{1-\alpha}(u)\mathrm{d}u\right)^{\frac{1-\beta}{1-\alpha}} - 1\right], \quad (12)$$

$$D_{\chi^2}(p,q) = \int \frac{[q(u) - p(u)]^2}{p(u)}\mathrm{d}u = \int p^{-1}(u)q^2(u)\mathrm{d}u - 1.$$

- Quantity of interest:

$$I_{a,b}(p,q) = \int p^a(u)q^b(u)\mathrm{d}u. \quad (13)$$

Csiszár f-divergence

- Definition ($f \geq 0$, convex, $f(1) = 0$):

$$D_f(p,q) = \int p(u) f\left(\frac{q(u)}{p(u)}\right) \mathrm{d}u. \tag{14}$$

- Challenge:
  - in the general case: hard to estimate,
  - variational characterization,
  - convex programming.

Paper

- For the exponential family: D_{χ^2} – analytical formula.
- General (analytical) f:
  - series expansion of f,
  - each term is a D_{χ^2}-type quantity.

Exponential family

Definition:

$$p(u;\theta) = e^{\langle t(u),\theta\rangle - F(\theta) + k(u)} \quad (\theta \in \Theta), \tag{15}$$

where

- $t(u)$: sufficient statistic,
- $F(\theta) = -\log\left[\int e^{\langle t(u),\theta\rangle + k(u)} \mathrm{d}u\right]$:
  - log-normalizer (partition function, cumulant function),
  - characterizes the family,
  - in many cases: analytical formula!
- Θ: natural parameter space.

Exponential family - normal example

For the normal case $\left[N(m,\Sigma) \in \mathbb{R}^d\right]$:

$$\theta = (\theta_1, \theta_2) = \left(\Sigma^{-1} m, \frac{1}{2}\Sigma^{-1}\right), \tag{16}$$

$$F(\theta) = \frac{1}{4} tr\left(\theta_2^{-1}\theta_1\theta_1^T\right) - \frac{1}{2}\log\det(\theta_2) + \frac{d}{2}\log(\pi), \tag{17}$$

$$t(u) = \left(u, -uu^T\right), \tag{18}$$

$$k(u) = 0. \tag{19}$$

Exponential family – special cases

A few important special cases:

Gaussian or normal (generic, isotropic Gaussian, diagonal Gaussian, rectified Gaussian or Wald distributions, log-normal), Poisson, Bernoulli, binomial, multinomial (trinomial, Hardy-Weinberg distribution), Laplacian, Gamma (including the chi-squared), Beta, exponential, Wishart, Dirichlet, Rayleigh, negative binomial, Weibull, Fisher-von Mises, Pareto, skew logistic, hyperbolic secant, ...

D_{χ^2}: analytical formula

Proposition: if

- $p = \varepsilon_F(\theta_1)$, $q = \varepsilon_F(\theta_2)$,
- Θ: is affine, and
- $a + b = 1$

then

$$I_{a,b}(p,q) = \int p^a(u) q^b(u) \mathrm{d}u = e^{F(a\theta_1 + b\theta_2) - [aF(\theta_1) + bF(\theta_2)]}. \quad (20)$$

D_{χ^2}: analytical formula – proof

$$
\begin{aligned}
I_{a,b}(p,q) &= \int p^a(u) q^b(u) \mathrm{d}u \\
&= \int e^{a[\langle t(u),\theta_1\rangle - F(\theta_1) + k(u)]} e^{b[\langle t(u),\theta_2\rangle - F(\theta_2) + k(u)]} \mathrm{d}u \\
&= \int e^{\langle t(u), a\theta_1 + b\theta_2\rangle - [aF(\theta_1) + bF(\theta_2)] + k(u)} \mathrm{d}u \\
&= \int e^{\langle t(u), a\theta_1 + b\theta_2\rangle + k(u)} e^{-[aF(\theta_1) + bF(\theta_2)]} e^{F(a\theta_1 + b\theta_2)} e^{-F(a\theta_1 + b\theta_2)} \mathrm{d}u \\
&= e^{F(a\theta_1 + b\theta_2) - [aF(\theta_1) + bF(\theta_2)]} \int p_F(u; a\theta_1 + b\theta_2) \mathrm{d}u \\
&= e^{F(a\theta_1 + b\theta_2) - [aF(\theta_1) + bF(\theta_2)]} \times 1.
\end{aligned}
$$

Higher-order Pearson-Vajda $D_{\chi^2_k}$

- Definition:

$$D_{\chi^2_k}(p,q) = \int \frac{[q(u) - p(u)]^k}{p^{k-1}(u)} \mathrm{d}u. \tag{21}$$

- Specially:
  - $k = 0$: $D_{\chi^2_0}(p,q) = 1$,
  - $k = 1$: $D_{\chi^2_1}(p,q) = 0$,
  - $k = 2$: $D_{\chi^2_k}(p,q) = D_{\chi^2}(p,q)$.

f-divergence estimation

- $D_{\chi^2_k}$: by the binomial theorem reduction to D_{χ^2}.
- Let f be analytical:

$$f(u) = \sum_{k=0}^{\infty} \frac{f^{(k)}(\lambda)}{k!}(u-\lambda)^k, \qquad (22)$$

$$D_f(p,q) = \sum_{k=0}^{\infty} \frac{f^{(k)}(\lambda)}{k!} D_{\chi^2_k}(p,q;\lambda), \qquad (23)$$

where

$$D_{\chi^2_k}(p,q;\lambda) = \int \frac{[q(u)-\lambda p(u)]^k}{p^{k-1}(u)} \mathrm{d}u. \qquad (24)$$

f-divergence estimation

Important special case ($\lambda = 1$, $k = 2$):

$$D_f(p,q) \approx f(1) + f'(1) D_{\chi_1^2}(p,q) + \frac{f''(1)}{2} D_{\chi_2^2}(p,q) \tag{25}$$

$$= 0 + f'(1)0 + \frac{f''(1)}{2} D_{\chi^2}(p,q) \tag{26}$$

$$= \frac{f''(1)}{2} D_{\chi^2}(p,q). \tag{27}$$

Summary

- Information theoretical quantities:
  - simple functionals of the distributions.
- Exponential family (affine natural space):
  - analytical expressions for the
    - χ^2-divergence,
    - Pearson-Vajda divergence,
  - efficient approximations for the f-divergence.

Thank you for the attention!