On the Chi square and higher-order Chi distances for approximating f-divergences

Frank Nielsen[1] Richard Nock[2]
www.informationgeometry.org

[1]Sony Computer Science Laboratories, Inc.
[2]UAG-CEREGMIA

September 2013

Statistical divergences

Measures the separability between two distributions.

Examples: Pearson/Neymann χ^2, Kullback-Leibler divergence:

$$\begin{aligned}
\chi^2_P(X_1 : X_2) &= \int \frac{(x_2(x) - x_1(x))^2}{x_1(x)} \mathrm{d}\nu(x), \\
\chi^2_N(X_1 : X_2) &= \int \frac{(x_1(x) - x_2(x))^2}{x_2(x)} \mathrm{d}\nu(x), \\
\mathrm{KL}(X_1 : X_2) &= \int x_1(x) \log \frac{x_1(x)}{x_2(x)} \mathrm{d}\nu(x),
\end{aligned}$$

f-divergences: A generic definition

$$I_f(X_1 : X_2) = \int x_1(x) f\left(\frac{x_2(x)}{x_1(x)}\right) \mathrm{d}\nu(x) \geq 0,$$

where f is a convex function

$$f : (0, \infty) \subseteq \mathrm{dom}(f) \mapsto [0, \infty]$$

such that $f(1) = 0$.

Jensen inequality: $I_f(X_1 : X_2) \geq f(\int x_2(x) \mathrm{d}\nu(x)) = f(1) = 0$.

May consider $f'(1) = 0$ and fix the scale of divergence by setting $f''(1) = 1$.

Can always be symmetrized:

$$S_f(X_1 : X_2) = I_f(X_1 : X_2) + I_{f^*}(X_1 : X_2)$$

with $f^*(u) = uf(1/u)$, and $I_{f^*}(X_1 : X_2) = I_f(X_2 : X_1)$.

f-divergences: Some examples

Name of the f-divergence	Formula $I_f(P:Q)$	Generator $f(u)$ with $f(1)=0$
Total variation (metric)	$\frac{1}{2}\int \lvert p(x)-q(x)\rvert \mathrm{d}\nu(x)$	$\frac{1}{2}\lvert u-1\rvert$
Squared Hellinger	$\int(\sqrt{p(x)}-\sqrt{q(x)})^2\mathrm{d}\nu(x)$	$(\sqrt{u}-1)^2$
Pearson χ^2_P	$\int \frac{(q(x)-p(x))^2}{p(x)}\mathrm{d}\nu(x)$	$(u-1)^2$
Neyman χ^2_N	$\int \frac{(p(x)-q(x))^2}{q(x)}\mathrm{d}\nu(x)$	$\frac{(1-u)^2}{u}$
Pearson-Vajda χ^k_P	$\int \frac{(q(x)-\lambda p(x))^k}{p^{k-1}(x)}\mathrm{d}\nu(x)$	$(u-1)^k$
Pearson-Vajda $\lvert\chi\rvert^k_P$	$\int \frac{\lvert q(x)-\lambda p(x)\rvert^k}{p^{k-1}(x)}\mathrm{d}\nu(x)$	$\lvert u-1\rvert^k$
Kullback-Leibler	$\int p(x)\log\frac{p(x)}{q(x)}\mathrm{d}\nu(x)$	$-\log u$
reverse Kullback-Leibler	$\int q(x)\log\frac{q(x)}{p(x)}\mathrm{d}\nu(x)$	$u\log u$
α-divergence	$\frac{4}{1-\alpha^2}(1-\int p^{\frac{1-\alpha}{2}}(x)q^{1+\alpha}(x)\mathrm{d}\nu(x))$	$\frac{4}{1-\alpha^2}(1-u^{\frac{1+\alpha}{2}})$
Jensen-Shannon	$\frac{1}{2}\int(p(x)\log\frac{2p(x)}{p(x)+q(x)}+q(x)\log\frac{2q(x)}{p(x)+q(x)})\mathrm{d}\nu(x)$	$-(u+1)\log\frac{1+u}{2}+u\log u$

Stochastic approximations of f-divergences

$$\widehat{I_f^{(n)}}(X_1 : X_2) \sim \frac{1}{2n} \sum_{i=1}^{n} \left(f\left(\frac{x_2(s_i)}{x_1(s_i)}\right) + \frac{x_1(t_i)}{x_2(t_i)} f\left(\frac{x_2(t_i)}{x_1(t_i)}\right) \right),$$

with $s_1, ..., s_n$ and $t_1, ..., t_n$ IID. sampled from X_1 and X_2, respectively.

$$\lim_{n\rightarrow\infty} \widehat{I_f^{(n)}}(X_1 : X_2) \rightarrow I_f(X_1 : X_2)$$

- work for any generator f but...
- In practice, limited to small dimension support.

Exponential families

Canonical decomposition of the probability measure:

$$p_\theta(x) = \exp(\langle t(x), \theta \rangle - F(\theta) + k(x)),$$

Here, consider natural parameter space Θ affine.

$$\begin{aligned} \mathrm{Poi}(\lambda) &: \quad p(x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}, \lambda > 0, x \in \{0, 1, ...\} \\ \mathrm{Nor}_I(\mu) &: \quad p(x|\mu) = (2\pi)^{-\frac{d}{2}} e^{-\frac{1}{2}(x-\mu)^\top (x-\mu)}, \mu \in \mathbb{R}^d, x \in \mathbb{R}^d \end{aligned}$$

Family	θ	Θ	$F(\theta)$	$k(x)$	$t(x)$	ν
Poisson	$\log \lambda$	$\mathbb{R}$	e^θ	$-\log x!$	x	ν_c
Iso.Gaussian	μ	$\mathbb{R}^d$	$\frac{1}{2}\theta^\top\theta$	$\frac{d}{2}\log 2\pi - \frac{1}{2}x^\top x$	x	ν_L

χ^2 for affine exponential families

Bypass integral computation,

Closed-form formula

$$
\begin{aligned}
\chi^2_P(X_1 : X_2) &= e^{F(2\theta_2-\theta_1)-(2F(\theta_2)-F(\theta_1))} - 1, \\
\chi^2_N(X_1 : X_2) &= e^{F(2\theta_1-\theta_2)-(2F(\theta_1)-F(\theta_2))} - 1,
\end{aligned}
$$

Kullback-Leibler divergence amounts to a Bregman divergence [3]:

$$\mathrm{KL}(X_1 : X_2) = B_F(\theta_2 : \theta_1)$$

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

Higher-order Vajda χ^k divergences

$$
\begin{aligned}
\chi_P^k(X_1 : X_2) &= \int \frac{(x_2(x) - x_1(x))^k}{x_1(x)^{k-1}} \mathrm{d}\nu(x), \\
|\chi|_P^k(X_1 : X_2) &= \int \frac{|x_2(x) - x_1(x)|^k}{x_1(x)^{k-1}} \mathrm{d}\nu(x),
\end{aligned}
$$

are f-divergences for the generators $(u-1)^k$ and $|u-1|^k$.

- When $k=1$, $\chi_P^1(X_1 : X_2) = \int (x_1(x) - x_2(x)) \mathrm{d}\nu(x) = 0$ (never discriminative), and $|\chi_P^1|(X_1, X_2)$ is twice the total variation distance.
- χ_P^0 is the unit constant.
- χ_P^k is a signed distance

Higher-order Vajda χ^k divergences

Lemma

The (signed) χ^k_P distance between members $X_1 \sim \mathcal{E}_F(\theta_1)$ and $X_2 \sim \mathcal{E}_F(\theta_2)$ of the same affine exponential family is ($k \in \mathbb{N}$) always bounded and equal to:

$$\chi^k_P(X_1 : X_2) = \sum_{j=0}^{k} (-1)^{k-j} \binom{k}{j} \frac{e^{F((1-j)\theta_1 + j\theta_2)}}{e^{(1-j)F(\theta_1) + jF(\theta_2)}}.$$

Higher-order Vajda χ^k divergences:

For Poisson/Normal distributions, we get closed-form formula:

$$
\begin{aligned}
\chi_P^k(\lambda_1 : \lambda_2) &= \sum_{j=0}^{k} (-1)^{k-j} \binom{k}{j} e^{\lambda_1^{1-j}\lambda_2^j - ((1-j)\lambda_1 + j\lambda_2)}, \\
\chi_P^k(\mu_1 : \mu_2) &= \sum_{j=0}^{k} (-1)^{k-j} \binom{k}{j} e^{\frac{1}{2}j(j-1)(\mu_1-\mu_2)^\top(\mu_1-\mu_2)}.
\end{aligned}
$$

signed distances.

f-divergences from Taylor series

Lemma (extends Theorem 1 of [1])

When bounded, the f-divergence I_f can be expressed as the power series of higher order Chi-type distances:

$$
\begin{aligned}
I_f(X_1 : X_2) &= \int x_1(x) \sum_{i=0}^{\infty} \frac{1}{i!} f^{(i)}(\lambda) \left(\frac{x_2(x)}{x_1(x)} - \lambda \right)^i \mathrm{d}\nu(x), \\
&= \sum_{i=0}^{\infty} \frac{1}{i!} f^{(i)}(\lambda)\ \chi^i_{\lambda,P}(X_1 : X_2),
\end{aligned}
$$

$I_f < \infty$, and $\chi^i_{\lambda,P}(X_1 : X_2)$ is a generalization of the χ^i_P defined by:

$$
\chi^i_{\lambda,P}(X_1 : X_2) = \int \frac{(x_2(x) - \lambda x_1(x))^i}{x_1(x)^{i-1}} \mathrm{d}\nu(x).
$$

and $\chi^0_{\lambda,P}(X_1 : X_2) = 1$ by convention. Note that $\chi^i_{\lambda,P} \geq f(1) = (1-\lambda)^k$ is a f-divergence for $f(u) = (u-\lambda)^k - (1-\lambda)^k$

f-divergences: Analytic formula

- $\lambda = 1 \in \mathrm{int}(\mathrm{dom}(f^{(i)}))$, f-divergence (Theorem 1 of [1]):

$$|I_f(X_1 : X_2) - \sum_{k=0}^{s} \frac{f^{(k)}(1)}{k!} \chi_P^k(X_1 : X_2)| \leq \frac{1}{(s+1)!} \|f^{(s+1)}\|_\infty (M-m)^s,$$

where $\|f^{(s+1)}\|_\infty = \sup_{t\in[m,M]} |f^{(s+1)}(t)|$ and $m \leq \frac{p}{q} \leq M$.

- $\lambda = 0$ (whenever $0 \in \mathrm{int}(\mathrm{dom}(f^{(i)}))$) and affine exponential families, simpler expression:

$$\begin{aligned} I_f(X_1 : X_2) &= \sum_{i=0}^{\infty} \frac{f^{(i)}(0)}{i!} I_{1-i,i}(\theta_1 : \theta_2), \\ I_{1-i,i}(\theta_1 : \theta_2) &= \frac{e^{F(i\theta_2 + (1-i)\theta_1)}}{e^{iF(\theta_2)+(1-i)F(\theta_1)}}. \end{aligned}$$

Corollary: Approximating f-divergences by χ^2 divergences

Corollary

A second-order Taylor expansion yields

$$I_f(X_1 : X_2) \sim f(1) + f'(1)\chi^1_N(X_1 : X_2) + \frac{1}{2} f''(1)\chi^2_N(X_1 : X_2)$$

Since $f(1) = 0$ and $\chi^1_N(X_1 : X_2) = 0$, it follows that

$$I_f(X_1 : X_2) \sim \frac{f''(1)}{2}\chi^2_N(X_1 : X_2),$$

($f''(1) > 0$ follows from the strict convexity of the generator).
When $f(u) = u \log u$, this yields the well-known approximation [2]:

$$\chi^2_P(X_1 : X_2) \sim 2\,\mathrm{KL}(X_1 : X_2).$$

Kullback-Leibler divergence: Analytic expression

Kullback-Leibler divergence: $f(u) = -\log u$.

$$f^{(i)}(u) = (-1)^i (i-1)! u^{-i}$$

and hence $\frac{f^{(i)}(1)}{i!} = \frac{(-1)^i}{i}$, for $i \geq 1$ (with $f(1) = 0$).
Since $\chi^1_{1,P} = 0$, it follows that:

$$\mathrm{KL}(X_1 : X_2) = \sum_{j=2}^{\infty} \frac{(-1)^i}{i} \chi^j_P(X_1 : X_2).$$

$\rightarrow$ alternating sign sequence
Poisson distributions: $\lambda_1 = 0.6$ and $\lambda_2 = 0.3$, $\mathrm{KL} \sim 0.1158$ (exact using Bregman divergence), stochastic evaluation with $n = 10^6$ yields $\widehat{KL} \sim 0.1156$
KL divergence from Taylor truncation: $0.0809 (s = 2)$, $0.0910 (s = 3)$, $0.1017 (s = 4)$, $0.1135 (s = 10)$, $0.1150 (s = 15)$, etc.

Contributions

Statistical f-divergences between members of the same exponential family with **affine natural space**.

- Generic closed-form formula for the Pearson/Neyman χ^2 and Vajda χ^k-type distance
- Analytic expression of f-divergences using Pearson-Vajda-type distances.
- Second-order Taylor approximation for fast estimation of f-divergences.

Java™ package:
`www.informationgeometry.org/fDivergence/`

Thank you.

```
@article{fDivChi-arXiv1309.3029,
author="Frank Nielsen and Richard Nock",
title="On the {C}hi square and higher-order {C}hi distances for approximating $f$-divergences",
year="2013",
eprint="arXiv/1309.3029"
}
```

www.informationgeometry.org

Bibliographic references I

N.S. Barnett, P. Cerone, S.S. Dragomir, and A. Sofo.

Approximating Csiszár f-divergence by the use of Taylor's formula with integral remainder.

Mathematical Inequalities & Applications, 5(3):417–434, 2002.

Thomas M. Cover and Joy A. Thomas.

Elements of information theory.

Wiley-Interscience, New York, NY, USA, 1991.

Frank Nielsen and Sylvain Boltz.

The Burbea-Rao and Bhattacharyya centroids.

IEEE Transactions on Information Theory, 57(8):5455–5466, August 2011.