A generalization of independence and multivariate Student's t-distributions

MATSUZOE Hiroshi
Nagoya Institute of Technology

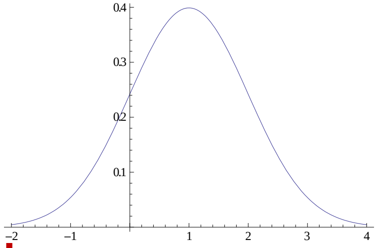
joint works with
SAKAMOTO Monta (Efrei, Paris)

Notions of expectations, independence are determined from the choice of statistical models.

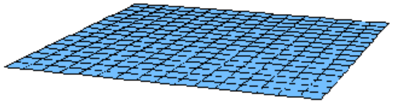
Introduction: Geometry and statistics

Probability density function $p(x;\theta)$

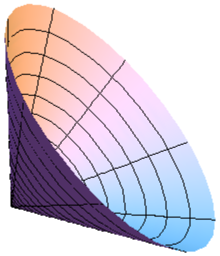
Random variable → x Parameter → θ



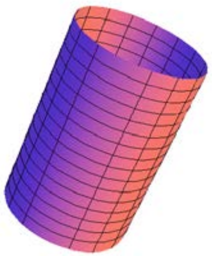
- Geometry for the sample space
- Kernel method
  - Reproducing kernel Hilbert space



Polynomial kernel



Gaussian kernel



- Geometry for the parameter space
- Wasserstein geometry
  - Optimal transport theory
  - A pdf is regarded as a distribution of mass
- Information geometry
  - Convexity of entropy and free energy
  - Duality of estimating function

1 Deformed exponential family (χ-exp. family)

$\chi : (0, \infty) \to (0, \infty)$: strictly increasing

χ-exponential, χ-logarithm

Definition 1.1

$$\log_\chi x := \int_1^x \frac{1}{\chi(t)} dt \qquad \chi\text{-logarithm}$$

$$\exp_\chi x := 1 + \int_0^x \lambda(t) dt \qquad \chi\text{-exponential}$$

$$\textit{where} \quad \lambda(\log_\chi t) = \chi(t)$$

In the case $\chi(t) = t$, χ-exponential and χ-logarithm recover the standard exponential and the standard logarithm.

Example 1.2 *In the case $\chi(t) = t^q$, we have*

$$\int_1^x \frac{1}{\chi(t)} dt = \int_1^x \frac{1}{t^q} dt = \frac{x^{1-q} - 1}{1-q} = \log_q x \qquad q\text{-logarithm}$$

$$\lambda(t) = (1 + (1-q)t)^{\frac{q}{1-q}}$$

$$1 + \int_0^x \lambda(t)\, dt = (1 + (1-q)x)^{\frac{1}{1-q}} \qquad q\text{-exponential}$$

$\chi : (0, \infty) \to (0, \infty)$: strictly increasing

Definition 1.1 (χ-exponential, χ-logarithm)

$$\log_\chi x := \int_1^x \frac{1}{\chi(t)} dt \qquad \chi\text{-logarithm}$$

$$\exp_\chi x := 1 + \int_0^x \lambda(t) dt \qquad \chi\text{-exponential}$$

$$\textit{where} \quad \lambda(\log_\chi t) = \chi(t)$$

$F_1(x), \ldots, F_n(x)$: functions on Ω

$\theta = \{\theta^1, \ldots, \theta^n\}$: parameters

$$S = \left\{ p(x, \theta) \;\middle|\; p(x; \theta) > 0, \int_\Omega p(x; \theta) dx = 1 \right\}: \text{ statistical model}$$

Definition 1.4

$S_\chi = \{p(x; \theta)\}$: χ-exponential family, deformed exponential family

$$\overset{\text{def}}{\Longleftrightarrow} \; S_\chi := \left\{ p(x, \theta) \middle| p(x; \theta) = \exp_\chi \left[\sum_{i=1}^n \theta^i F_i(x) - \psi(\theta) \right], p(x, \theta) \in S \right\}$$

Example 1.5 (Student t-distribution (q-normal distribution))
$\Omega = R, \; n = 2, \; \xi = (\mu, \sigma) \in R^2_+$ (the upper half plane), $1 < q < 3$.

$$p(x; \mu, \sigma) = \frac{1}{z_q}\left[1 - \frac{1-q}{3-q}\frac{(x-\mu)^2}{\sigma^2}\right]^{\frac{1}{1-q}}$$

Set

$$\theta^1 = \frac{2}{3-q} z_q^{q-1} \cdot \frac{\mu}{\sigma^2}, \qquad \theta^2 = -\frac{1}{3-q} z_q^{q-1} \cdot \frac{1}{\sigma^2}.$$

Then

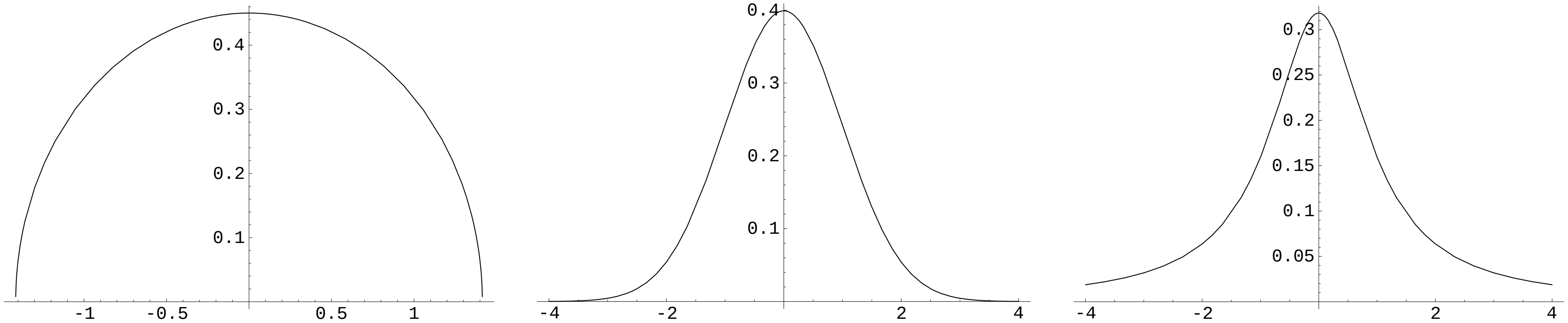
$$\begin{aligned}
\log_q p_q(x) &= \frac{1}{1-q}(p^{1-q} - 1) = \frac{1}{1-q}\left\{\frac{1}{z_q^{1-q}}\left(1 - \frac{1-q}{3-q}\frac{(x-\mu)^2}{\sigma^2}\right) - 1\right\} \\
&= \frac{2\mu z_q^{q-1}}{(3-q)\sigma^2}x - \frac{z_q^{q-1}}{(3-q)\sigma^2}x^2 - \frac{z_q^{q-1}}{3-q} \cdot \frac{\mu^2}{\sigma^2} + \frac{z_q^{q-1} - 1}{1-q} \\
&= \theta^1 x + \theta^2 x^2 - \psi(\theta)
\end{aligned}$$

$$\psi(\theta) = -\frac{(\theta^1)^2}{4\theta^2} - \frac{z_q^{q-1} - 1}{1-q}$$

The set of Student t-distributions is a q-exponential family.

$$p(x;\mu,\sigma) = \frac{1}{z_q}\left[1 - \frac{1-q}{3-q}\frac{(x-\mu)^2}{\sigma^2}\right]_+^{\frac{1}{1-q}}$$

q	distribution
$-\infty$	uniform on $[\mu-\sigma, \mu+\sigma]$
-1	semi-circle
1	normal
$1+\dfrac{1}{n+1}$	Student t
2	Cauchy distribution
3	uniform on $[-\infty, \infty]$



Example 1.6 (Multivariate Student t-distribution)

$$\Omega = R^d, \quad n = \frac{d(d+3)}{2}, \quad \xi = (\mu, \Sigma) \in R^d \times \mathrm{Syd}^+(d), \quad 1 < q < 1 + \frac{2}{d}.$$

By setting $\nu = -d - \dfrac{2}{1-q}, \quad z_q = \dfrac{(\pi\nu)^{\frac{d}{2}}\Gamma\left(\frac{\nu}{2}\right)\sqrt{\det(\Sigma)}}{\Gamma(\frac{1}{q-1})},$

$$p_q(x; \mu, \Sigma) := \frac{1}{z_q}\left[1 + \frac{1}{\nu}{}^t(x-\mu)\Sigma^{-1}(x-\mu)\right]^{\frac{1}{1-q}}.$$

Set

$$\tilde{R} = \frac{z_q^{q-1}}{(1-q)d+2}\Sigma^{-1}, \ \theta = 2\tilde{R}\mu, \ \psi(\theta) = \frac{1}{4}{}^t\theta\tilde{R}^{-1}\theta - \ln_q \frac{1}{z_q}.$$

$$\Longrightarrow \quad \log_q p_q(x; \mu, \Sigma) = \log_q \left(\frac{1}{z_q}\left[1 + \frac{1}{\nu}{}^t(x-\mu)\Sigma^{-1}(x-\mu)\right]^{\frac{1}{1-q}}\right)$$

$$= -{}^t(x-\mu)\tilde{R}(x-\mu) + \ln_q \frac{1}{z_q}$$

$$= \sum_{i=1}^d \theta^i x_i - \sum_{i=1}^d \tilde{R}_{ii}x_i^2 - 2\sum_{i<j}\tilde{R}_{ij}x_i x_j - \frac{1}{4}{}^t\theta\tilde{R}^{-1}\theta + \ln_q \frac{1}{z_q}.$$

The set of Student t-distributions is a q-exponential family.

Example 1.7 (discrete distributions)

$\Omega = \{x_0, x_1, \dots, x_n\}$

$$S_n = \left\{ p(x;\eta) \;\middle|\; \eta_i > 0,\ \sum_{i=0}^{n} \eta_i = 1,\ p(x;\eta) = \sum_{i=0}^{n} \eta_i \delta_i(x) \right\},$$

$$\eta_0 = 1 - \sum_{i=1}^{n} \eta_i$$

Set $\theta^i = \log_\chi p(x_i) - \log_\chi p(x_0) = \log_\chi \eta_i - \log_\chi \eta_0$

Then

$$\log_\chi p(x) = \log_\chi \left(\sum_{i=0}^{n} \eta_i \delta_i(x) \right) = \sum_{i=1}^{n} \left(\log_\chi \eta_i - \log_\chi \eta_0 \right) \delta_i(x) + \log_\chi(\eta_0)$$

$$\psi(\theta) = -\log_\chi \eta_0$$

The set of discrete distributions is a χ-exponential family for any χ.

2 Non-additive differentials and expectation functionals

Napier's constant

$$e := \lim_{n\to\infty}\left(1+\frac{1}{n}\right)^n$$

Napier's constant is defined by the limit of a monotone increasing series bounded above.

$$\begin{aligned}
a_1 &= \left(1+\frac{1}{1}\right)\\
a_2 &= \left(1+\frac{1}{2}\right)\left(1+\frac{1}{2}\right)\\
a_3 &= \left(1+\frac{1}{3}\right)\left(1+\frac{1}{3}\right)\left(1+\frac{1}{3}\right)\\
&\vdots\\
&\vdots
\end{aligned}$$

Another interpretation is that the generation speed of the copy of interval matches with the speed of the shrinkage of the interval.

2.2 Non-additive differential

$$\frac{d}{dx}f(x) = \{f(x)\}^q : \text{escort representation}$$

$$\Longrightarrow \quad f(x) = (1+(1-q)x)^{\frac{1}{1-q}} \qquad (= \exp_q x)$$

q-exponential, q-logarithm ($q > 0$)

$$\exp_q x := (1+(1-q)x)^{\frac{1}{1-q}} \quad (1+(1-q)x > 0) \qquad q\text{-exponential}$$

$$\log_q y := \frac{y^{1-q}-1}{1-q} \qquad (y > 0) \qquad q\text{-logarithm}$$

$q \to 1$, the standard exponential and logarithm are recovered.

q-sum, q-product

$$x_1 \tilde{\oplus}_q x_2 := x_1 + x_2 + (1-q)x_1 x_2 \qquad q\text{-sum}$$

$$y_1 \otimes_q y_2 := \left[y_1^{1-q} + y_2^{1-q} - 1\right]^{\frac{1}{1-q}} \qquad q\text{-product}$$

$$\exp_q(x_1 \tilde{\oplus}_q x_2) = \exp_q x_1 \cdot \exp_q x_2,$$
$$\log_q(y_1 \cdot y_2) = \log_q y_1 \tilde{\oplus}_q \log_q y_2,$$
$$\exp_q(x_1 + x_2) = \exp_q x_1 \otimes_q \exp_q x_2,$$
$$\log_q(y_1 \otimes_q y_2) = \log_q y_1 + \log_q y_2.$$

q-sum, q-product

$$x_1 \tilde{\oplus}_q x_2 := x_1 + x_2 + (1-q)x_1 x_2 \qquad q\text{-sum}$$

$$y_1 \otimes_q y_2 := \left[y_1^{1-q} + y_2^{1-q} - 1 \right]^{\frac{1}{1-q}} \qquad q\text{-product}$$

$$[-x] := \log_q \left(\frac{1}{\exp_q x} \right) = \frac{1}{1-q} \left(\frac{1}{1+(1-q)x} - 1 \right) : \text{the inverse of } x$$

$$x_1 \tilde{\ominus}_q x_2 := x_1 \tilde{\oplus}_q [-x_2] : \qquad q\text{-difference}$$

non-additive q-differential

$$\frac{d_q}{d_q x} f(x) := \lim_{x' \to x} \frac{f(x') - f(x)}{x' \tilde{\ominus}_q x} : \text{ non-additive } q\text{-differential}$$

$$\frac{d_q}{d_q x} f(x) = f(x) : \text{non-additive representation}$$

$$\Longrightarrow \quad f(x) = (1+(1-q)x)^{\frac{1}{1-q}} \qquad (= \exp_q x)$$

$$\frac{d}{dx} f(x) = \{f(x)\}^q : \text{escort representation}$$

$$\Longrightarrow \quad f(x) = (1+(1-q)x)^{\frac{1}{1-q}} \qquad (= \exp_q x)$$

q-sum, q-product

$$x_1 \tilde{\oplus}^q x_2 := x_1 + x_2 + (1-q)x_1 x_2 \qquad q\text{-sum}$$

$$y_1 \otimes_q y_2 := \left[y_1^{1-q} + y_2^{1-q} - 1 \right]^{\frac{1}{1-q}} \qquad q\text{-product}$$

Infinite product expression of q-exponential function

Proposition 2.1

For all $n \in N$, *suppose that* $n\left(1+\frac{x}{n}\right)^{1-q} - (n-1) > 0$

$$\Longrightarrow \qquad \exp_q x = \lim_{n\to\infty} \left(1+\frac{x}{n}\right)^{\otimes_q^n}$$

where $$\left(1+\frac{x}{n}\right)^{\otimes_q^n} := \underbrace{\left(1+\frac{x}{n}\right) \otimes_q \cdots \otimes_q \left(1+\frac{x}{n}\right)}_{n \text{ times}}$$

2.3 Expectation functionals

Example 2.2 (q-normal distributions)

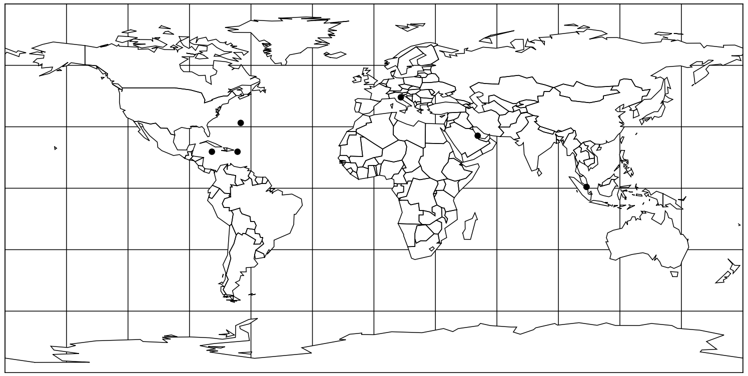
$$p(x;\mu,\sigma) = \frac{1}{Z_q}\left[1 - \frac{1-q}{3-q}\frac{(x-\mu)^2}{\sigma^2}\right]_+^{\frac{1}{1-q}}$$

q	distribution	mean	variance
-1	semi-circle	○	○
1	normal	○	○
$1+\frac{1}{n+1}$	student t	○	○
2	Cauchy distribution	×	×

$q \geq \frac{5}{3}$
$V[X]$ does not exist.
$q \geq 2$
$E[X]$ does not exist.

Expectations give nothing but a local coordinate system.

$\Longrightarrow$ It is better to choose a good local coordinate (expectations).



Local coordinates should be chosen to the purposes.

Formal definition of expectation functional

$(\Omega, \mathcal{F}, P)$: a probability space
$\mathcal{H}$ $(\subset L^1(\Omega))$: a set of measurable functions on Ω
$\Phi : \mathcal{H} \to C$: a bounded linear functional
$\Longrightarrow$ Φ is an expectation functional.

Definition 2.3

$E_p[f(x)]$: the simple (or standard) expectation of $f(x)$

$$\overset{\text{def}}{\Longleftrightarrow} \quad E_p[f(x)] = \int_\Omega f(x) p(x;\theta) dx$$

Definition 2.4 (Naudts (2004))

X : a random variable which follows to $p(x)$,
$r(x)$: a probability distribution on Ω. (an escort distribution of $p(x)$)
$F_r^{esc}[X]$: an escort expectation of X

$$\overset{\text{def}}{\Longleftrightarrow} \quad F_r^{esc}[X] := \int_\Omega x\, r(x) dx$$

Remark 2.5

(1) Tow probability distributions $p(x)$ and $r(x)$ have no relation.
(2) $r(x)$ is not necessary to be a probability distribution.

Definition 2.6

$P_q(x)$, $P_q^{esc}(x)$: an escort distribution and a normalized escort distribution of $p(x)$

$$\overset{\text{def}}{\Longleftrightarrow} \quad P_q(x;\theta) = \{p(x;\theta)\}^q,$$
$$P_q^{esc}(x;\theta) = \frac{1}{Z_q(\theta)}\{p(x;\theta)\}^q, \quad Z_q(\theta) = \int_\Omega \{p(x;\theta)\}^q dx$$

$E_{q,p}[f(x)]$: the q-canonical expectation of $f(x)$

$$\overset{\text{def}}{\Longleftrightarrow} \quad E_{q,p}[f(x)] = \int_\Omega f(x)P_q(x;\theta)dx = \int_\Omega f(x)\{p(x;\theta)\}^q dx$$

$E_{q,p}^{esc}[f(x)]$: the normalized q-escort expectation of $f(x)$

$$\overset{\text{def}}{\Longleftrightarrow} \quad E_{q,p}^{esc}[f(x)] = \int_\Omega f(x)P_q^{esc}(x;\theta)dx = \frac{1}{Z_q(\theta)}\int_\Omega f(x)\{p(x;\theta)\}^q dx$$

q-exponential, q-logarithm $(\chi(t) = t^q)$

$$\log_q x := \int_1^x \frac{1}{t^q}dt = \frac{x^{1-q}-1}{1-q} \qquad q\text{-logarithm}$$

$$\exp_\chi x := (1+(1-q)x)^{\frac{1}{1-q}} \qquad q\text{-exponential}$$

Theorem 2.7 (Wada, M, Scarfone)

S : a statistical model

Λ : a monotone increasing function on $(0, \infty)$ (Ran$\{S\{p(x;\theta)\}\}$)

$\tilde{s}(x;\theta)$: a generalized score functional

$$\tilde{s}^i(x;\theta) := \partial_i \Lambda(p(x;\theta)), \ \ p(x;\theta) \in S$$

Suppose that $P(x;\theta) \propto \left.\frac{d\Lambda^{-1}(t)}{dt}\right|_{t=\Lambda(p(x;\theta))}$

$\Longrightarrow$ $\tilde{s}(x;\theta)$ is unbiased with respect to $P(x;\theta)$, that is,

$$\int_\Omega \tilde{s}(x;\theta) P(x;\theta) dx \ = 0$$

Proof: The generalized score functional is unbiased if.

$$P(x;\theta)\partial_i \Lambda(p(x;\theta)) = P(x;\theta)\frac{\partial \Lambda(p)}{\partial p} \partial_i p(x;\theta) \propto \partial_i p(x;\theta). \tag{1}$$

This implies that

$$P(x;\theta) \propto \frac{1}{\partial \Lambda(p)/\partial p} = \left.\frac{\partial \Lambda^{-1}(t)}{\partial t}\right|_{t=\Lambda(p(x;\theta))}. \tag{2}$$

Remark 2.8 When the statistical model is a χ-exponential family, it is natural to choice $P(x;\theta) = \chi\{p(x;\theta)\}$.

3 Geometry of deformed exponential families

S_χ : a deformed exponential family

$$\overset{\text{def}}{\Longleftrightarrow} S_\chi := \left\{ p(x,\theta) \middle| p(x;\theta) = \exp_\chi \left[\sum_{i=1}^{n} \theta^i F_i(x) - \psi(\theta) \right], p(x,\theta) \in S \right\}$$

$P_\chi^{esc}(x)$: a normalized escort distribution of $p(x;\theta)$,

$$\overset{\text{def}}{\Longleftrightarrow} \quad P_\chi^{esc}(x;\theta) = \frac{1}{Z_\chi(\theta)} \chi(p(x;\theta)), \quad Z_\chi(\theta) = \int_\Omega \chi(p(x;\theta))dx$$

$E_{\chi,p}^{esc}[f(x)]$: the normalized χ-escort expectation of $p(x)$

$$\overset{\text{def}}{\Longleftrightarrow} E_{\chi,p}^{esc}[f(x)] = \int_\Omega f(x) P_\chi^{esc}(x;\theta)dx = \frac{1}{Z_\chi(\theta)} \int_\Omega f(x)\chi(p(x;\theta))dx$$

Definition 3.1 $S_\chi = \{p(x;\theta)\}$: a deformed exponential family

$$g_{ij}^\chi(\theta) = \partial_i \partial_j \psi(\theta) : \text{ the } \chi\text{-Fisher information metric}$$
$$C_{ijk}^\chi(\theta) = \partial_i \partial_j \partial_k \psi(\theta) : \text{ the } \chi\text{-cubic form}$$

Set

$$\Gamma_{ij,k}^{\chi(e)} := \Gamma_{ij,k}^{\chi(0)} - \frac{1}{2} C_{ijk}^\chi, \qquad \Gamma_{ij,k}^{\chi(m)} := \Gamma_{ij,k}^{\chi(0)} + \frac{1}{2} C_{ijk}^\chi,$$

Proposition 3.2 For S_χ, the following hold:

(1) $(S_\chi, g^\chi, \nabla^{\chi(e)}, \nabla^{\chi(m)})$ is a dually flat space.

(2) $\{\theta^i\}$ is a $\nabla^{\chi(e)}$-affine coordinate system on S_χ.

(3) ψ is the potential of g^χ with respect to $\{\theta^i\}$, that is,

$$g^\chi_{ij}(\theta) = \partial_i \partial_j \psi(\theta).$$

(4) Set the χ-expectation of $F_i(x)$ by $\eta_i = E^{esc}_{\chi,p}[F_i(x)]$.
$\Longrightarrow$ $\{\eta_i\}$ is the dual coordinate system of $\{\theta^i\}$ with respect to g^χ.

(5) Set $\phi(\eta) = E^{esc}_{\chi,p}[\log_\chi p(x;\theta)]$
$\Longrightarrow$ $\phi(\eta)$ is the potential of g^χ with respect to $\{\eta_i\}$.

The generalized relative entropy (or χ-relative entropy) of S_χ by

$$D^\chi(p(x;\theta), p(x;\theta')) := E^{esc}_{\chi,p}[\log_\chi p(x;\theta) - \log_\chi p(x;\theta')]$$
$$= \psi(\theta') + \phi(\theta) - \sum_{i=1}^{n} (\theta')^i \eta_i = D(p(x;\theta'), p(x;\theta))$$

The generalized relative entropy D^χ of S_χ coincides with the canonical divergence $D(p_{\theta'}, p_\theta)$ on $(S_\chi, \nabla^{\chi(e)}, g^\chi)$.

α-divergence ($\alpha = 1 - 2q$)

$$D^{(1-2q)}\left(p(x), r(x)\right) = \frac{1}{q}\int_{\Omega} p(x)^q \{\log_q p(x) - \log_q r(x)\} dx$$

$D^{(1-2q)}$ induces a non-flat invariant statistical manifold $(S_q, \nabla^{(1-2q)}, g^F)$.

normalized Tsallis relative entropy (χ-relative entropy)

$$D^q\left(p(x), r(x)\right) = E_{q,p}^{esc}\left[\log_q p(x) - \log_q r(x)\right]$$

$$= \int_{\Omega} \frac{p(x)^q}{Z_q(p)} \{\log_q p(x) - \log_q r(x)\} dx \quad \left(= \frac{q}{Z_q(p)} D^{(1-2q)}(p, r) \right)$$

D^q induces a Hessian manifold (flat statistical mfd.) $(S_q, \nabla^{q(m)}, g^q)$.

In general, if two contrast functions have the following relation:

$$D(p, r) = f(p) D(p, r),$$

then induced statistical manifolds are 1-conformally equivalent.

$\nu(x)$ pos. measure	$\not\longrightarrow$	$\dfrac{\nu(x)}{Z_q(\nu)}$ prob. measure	Normalization of a positive measure to a probability measure is NOT a trivial problem.

4 Generalization of independence

4.1 Deformed independences

$X \sim p_1(x), \; Y \sim p_2(y)$

X and Y are independent

$$\overset{\text{def}}{\iff} \quad p(x,y) = p_1(x)p_2(y).$$

$$\iff \quad p(x,y) = \exp\left[\log p_1(x) + \log p_2(y)\right] \quad (p_1(x) > 0, \; p_2(y) > 0)$$

Formal definition of χ-product

$y_1 \otimes_\chi y_2$: the χ-product of y_1 and y_2

$$\overset{\text{def}}{\iff} \quad y_1 \otimes_\chi y_2 = \exp_\chi\left[\log_\chi y_1 + \log_\chi y_2\right]$$

$$\exp_\chi x_1 \otimes_\chi \exp_\chi x_2 = \exp_\chi(x_1 + x_2),$$
$$\log_\chi(y_1 \otimes_\chi y_2) = \log_\chi y_1 + \log_\chi y_2.$$

X_1 and X_2 : χ-independent with m-normalization

$$\overset{\text{def}}{\iff} \quad p_\chi(x_1, x_2) = \frac{p_1(x_1) \otimes_\chi p_2(x_2)}{Z_{p_1,p_2}}$$

$$\text{where} \quad Z_{p_1,p_2} = \iint_{Supp\{p_\chi\} \subset \mathcal{X}_1\mathcal{X}_2} p_1(x_1) \otimes_\chi p_2(x_2) dx_1 dx_2$$

4 Generalization of independence

4.1 Deformed independences

$X \sim p_1(x),\ Y \sim p_2(y)$

X and Y are independent

$$\overset{\text{def}}{\iff} \quad p(x,y) = p_1(x)p_2(y).$$

$$\iff \quad p(x,y) = \exp\left[\log p_1(x) + \log p_2(y)\right] \quad (p_1(x) > 0,\ p_2(y) > 0)$$

Formal definition of χ-product

$y_1 \otimes_\chi y_2$: the χ-product of y_1 and y_2

$$\overset{\text{def}}{\iff} \quad y_1 \otimes_\chi y_2 = \exp_\chi\left[\log_\chi y_1 + \log_\chi y_2\right]$$

$$\exp_\chi x_1 \otimes_\chi \exp_\chi x_2 = \exp_\chi(x_1 + x_2),$$
$$\log_\chi(y_1 \otimes_\chi y_2) = \log_\chi y_1 + \log_\chi y_2.$$

X_1 and X_2 : χ-independent with e-normalization

$$\overset{\text{def}}{\iff} \quad p_\chi(x_1, x_2) = p_1(x_1) \otimes_\chi p_2(x_2) \otimes_\chi \exp_\chi(-c)$$

where c is determined by $\displaystyle\iint_{Supp\{p_\chi\}\subset \mathcal{X}_1\mathcal{X}_2} p_\chi(x_1, x_2)dx_1dx_2 = 1$

q-product

$x > 0, y > 0$ and $x^{1-q} + y^{1-q} - 1 > 0 \; (q > 0)$.

$x \otimes_q y$: the q-product of x and y

$$\overset{\text{def}}{\Longleftrightarrow} \quad x \otimes_q y := \left[x^{1-q} + y^{1-q} - 1\right]^{\frac{1}{1-q}} = \exp_q\left[\log_q x + \log_q y\right]$$

X and Y : q-independent $\overset{\text{def}}{\Longleftrightarrow} \quad p_q(x,y) = p_1(x) \otimes_q p_2(y)$

X and Y : q-independent with m-normalization (mixture normalization)

$$\overset{\text{def}}{\Longleftrightarrow} \quad p_q(x,y) = \frac{p_1(x) \otimes_q p_2(y)}{Z_{p_1,p_2}}$$

$$\text{where} \quad Z_{p_1,p_2} = \iint_{Supp\{p_q(x,y)\} \subset \mathcal{X}\mathcal{Y}} p_1(x) \otimes_q p_2(y) dx dy$$

X and Y : q-independent with e-normalization (exponential normalization)

$$\overset{\text{def}}{\Longleftrightarrow} \quad p_q(x,y) = p_1(x) \otimes_q p_2(y) \otimes_q \exp_q(-c)$$

$$\text{where } c \text{ is determined by} \quad \iint_{Supp\{p_q(x,y)\} \subset \mathcal{X}\mathcal{Y}} p_q(x,y) dx dy = 1$$

5 q-independence and Student's t-distributions

Theorem 1

$X_1, X_2 \sim p_1(x_1), p_2(x_2)$: *univariate Student's t-distributions with same parameter* q $(1 < q < 2)$

$\Longrightarrow$ *there exist a bivariate Student's t-distribution* $p(x_1, x_2)$ *such that*

$$p(x_1, x_2) = p(x_1; \mu_1, \sigma_1) \otimes_q p(x_2; \mu_2, \sigma_2) \otimes_q (-c)$$

(q-independent with e-normalization)

proof: We obtain a q-exponential representation of t-distribution:

$$p(x_1; \mu_1, \sigma_1) = \exp_q \left[\theta^1 x_1 - \theta^{11} {x_1}^2 - \frac{(\theta^1)^2}{4\theta^{11}} + \ln_q \frac{1}{z_q(\sigma_1)} \right],$$

where θ^1 and θ^{11} are natural parameters defined by

$$\theta^1 = \frac{2\mu_1 \{z_q(\sigma_1)\}^{q-1}}{(3-q)\sigma_1^2}, \quad \theta^{11} = \frac{\{z_q(\sigma_1)\}^{q-1}}{(3-q)\sigma_1^2}.$$

Since the normalization $z_q(\sigma_1)$ can be determined by the parameter θ^{11}, $p(x_1; \mu_1, \sigma_1)$ is uniquely determined from natural parameters θ^1 and θ^{11}. Set θ^2 and θ^{22} by changing parameters to μ_2 and σ_2.

Then we obtain a positive density by

$$\begin{aligned} & p(x_1;\mu_1,\sigma_1) \otimes_q p(x_2;\mu_2,\sigma_2) \\ = \ & \exp_q\left[\theta^1 x_1 + \theta^2 x_2 - \theta^{11}{x_1}^2 - \theta^{22}{x_2}^2 - \frac{(\theta^1)^2}{4\theta^{11}} - \frac{(\theta^2)^2}{4\theta^{22}} + A(\theta)\right], \\ & \quad \text{where} \quad A(\theta) = \ln_q \frac{1}{z_q(\sigma_1)} + \ln_q \frac{1}{z_q(\sigma_2)}. \end{aligned}$$

Recall that $p(x_1;\mu_1,\sigma_1) \otimes_q p(x_2;\mu_2,\sigma_2)$ is not a probability distribution. Setting the e-normalization function c by

$$c = A(\theta) - \ln_q \frac{1}{z_q} = \left(\ln_q \frac{1}{z_q(\sigma_1)} + \ln_q \frac{1}{z_q(\sigma_2)}\right) - \ln_q \frac{1}{z_q},$$

where z_q is the m-normalization of bivariate Student's t-distribution.

As a consequence, we have

$$\begin{aligned} p(x_1,x_2) &= p(x_1;\mu_1,\sigma_1) \otimes_q p(x_2;\mu_2,\sigma_2) \otimes_q (-c) \\ &= \exp_q\left[\theta^1 x_1 + \theta^2 x_2 - \theta^{11}{x_1}^2 - \theta^{22}{x_2}^2 - \frac{(\theta^1)^2}{4\theta^{11}} - \frac{(\theta^2)^2}{4\theta^{22}} + \ln_q \frac{1}{z_q}\right]. \end{aligned}$$

This implies that X_1 and X_2 are q-independent with e-normalization, and $p(x_1,x_2)$ is a bivariate Student's t-distribution.

Concluding Remarks

U-divergence $(S_\chi, g^M, \nabla^{M(e)}, \nabla^{M(m)})$

estimating function $u_\chi(x;\theta)$:

$$u_\chi^i(x;\theta) = \frac{\partial}{\partial\theta^i}\log_\chi p(x;\theta) - E_p\left[\frac{\partial}{\partial\theta^i}\log_\chi p(x;\theta)\right]$$

Riemannian metric g^M: $\quad g_{ij}^M(\theta) = \int_\Omega \partial_i p(x;\theta)\partial_j \log_\chi p(x;\theta)dx$

dual coordinates $\{\eta_i\}$: $\quad \eta_i = E_p[F_i(x)]$

χ-relative entropy $(S_\chi, g^\chi, \nabla^{\chi(e)}, \nabla^{\chi(m)})$

estimating function $(s^\chi)(x;\theta)$:

$$(s^\chi)^i(x;\theta) = \frac{\partial}{\partial\theta^i}\log_\chi p(x;\theta)$$ (unbiased under χ-expectation)

Riemannian metric g^χ: $\quad g_{ij}^\chi(\theta) = \frac{\partial^2}{\partial\theta^i\theta^j}\psi(\theta)$

dual coordinates $\{\eta_i\}$: $\quad \eta_i = E_{\chi,p}^{esc}[F_i(x)]$

The notion of expectations, independence are determined from a geometric structure of the statistical model.

Concluding Remarks

$P_\chi^{esc}(x)$: a normalized escort distribution of $p(x;\theta)$,

$$\overset{\text{def}}{\Longleftrightarrow} \quad P_\chi^{esc}(x;\theta) = \frac{1}{Z_\chi(\theta)}\chi(p(x;\theta)), \quad Z_\chi(\theta) = \int_\Omega \chi(p(x;\theta))dx$$

$E_{\chi,p}^{esc}[f(x)]$: the normalized χ-escort expectation of $p(x)$

$$\overset{\text{def}}{\Longleftrightarrow} \; E_{\chi,p}^{esc}[f(x)] = \int_\Omega f(x)P_\chi^{esc}(x;\theta)dx = \frac{1}{Z_\chi(\theta)}\int_\Omega f(x)\chi(p(x;\theta))dx$$

χ-relative entropy $\quad (S_\chi, g^\chi, \nabla^{\chi(e)}, \nabla^{\chi(m)})$

estimating function $(s^\chi)(x;\theta)$:

$$(s^\chi)^i(x;\theta) = \frac{\partial}{\partial\theta^i}\log_\chi p(x;\theta) \text{ (unbiased under } \chi\text{-expectation)}$$

Riemannian metric g^χ: $\quad g_{ij}^\chi(\theta) = \frac{\partial^2}{\partial\theta^i\theta^j}\psi(\theta)$

dual coordinates $\{\eta_i\}$: $\quad \eta_i = E_{\chi,p}^{esc}[F_i(x)]$

The notion of expectations, independence are determined from a geometric structure of the statistical model.

6 Geometry for χ-likelihood estimators

$S_\chi = \{p(x;\xi) | \xi \in \Xi\}$: a χ-exponential family
$\{x_1, \ldots, x_N\}$: N-observations from $p(x;\xi) \in S_\chi$.

$L_\chi(\xi)$: χ-likelihood function

$$\overset{\text{def}}{\iff} \quad L_\chi(\xi) = p(x_1;\xi) \otimes_\chi p(x_2;\xi) \otimes_\chi \cdots \otimes_\chi p(x_N;\xi)$$

$$\left(\iff \quad \log_\chi L_\chi(\xi) = \sum_{i=1}^{N} \log_\chi p(x_i;\xi) \right)$$

$\hat{\xi}$: the maximum χ-likelihood estimator

$$\overset{\text{def}}{\iff} \quad \hat{\xi} = \arg\max_{\xi \in \Xi} L_\chi(\xi) \quad \left(= \arg\max_{\xi \in \Xi} \log_\chi L_\chi(\xi) \right).$$

Theorem 6.1
the χ-likelihood is maximum
$\iff$ *the canonical divergence (χ-relative entropy) is minimum.*

$S_q = \{p(x;\xi) | \xi \in \Xi\}$: a q-exponential family
$\{x_1, \ldots, x_N\}$: N-observations from $p(x;\xi) \in S_q$.

$L_q(\xi)$: **q-likelihood function**

$$\stackrel{\text{def}}{\Longleftrightarrow} \qquad L_q(\xi) = p(x_1;\xi) \otimes_q p(x_2;\xi) \otimes_q \cdots \otimes_q p(x_N;\xi)$$

$$\left(\Longleftrightarrow \qquad \log_q L_q(\xi) = \sum_{i=1}^{N} \log_q p(x_i;\xi) \right)$$

In the case $q \to 1$, L_q is the standard likelihood function on Ξ.

$$\begin{aligned} & \exp_q(x_1 + x_2 + \cdots + x_N) \\ = & \exp_q x_1 \otimes_q \exp_q x_2 \otimes_q \cdots \otimes_q \exp_q x_N \\ = & \exp_q x_1 \cdot \exp_q \left(\frac{x_2}{1 + (1-q)x_1} \right) \cdots \exp_q \left(\frac{x_N}{1 + (1-q)\sum_{i=1}^{N-1} x_i} \right) \end{aligned}$$

Each measurement influences the others.

"**q-independent**", but random variables are strongly correlated.

Estimation preserving

Corollary 6.2 *Let $1 < q < 3$,*

$$N_q(\mu, \sigma^2) := \left\{ p(x;\mu,\sigma) \;\middle|\; p(x;\mu,\sigma) = \frac{1}{Z_q}\left[1 - \frac{1-q}{3-q}\frac{(x-\mu)^2}{\sigma^2}\right]^{\frac{1}{1-q}} \right\}$$

Student's t-distributions (q-normal distributions)

$\{x_1, \ldots, x_N\}$: *q-independent N-observations from* $p(x;\mu,\sigma) \in S_q$

$\Longrightarrow$ *q-maximum likelihood estimators in mixture coordinates are*

$$\hat{\eta}_1 = \frac{1}{N}\sum_{i=1}^{N} x_i, \qquad \hat{\eta}_2 = \frac{1}{N}\sum_{i=1}^{N} x_i^2$$

$$N(\mu, \sigma^2) \quad \Longrightarrow \quad \hat{\eta}_1 = \frac{1}{N}\sum_{i=1}^{N} x_i, \quad \hat{\eta}_2 = \frac{1}{N}\sum_{i=1}^{N} x_i^2$$

MLE

$$N_q(\mu, \sigma^2) \quad \Longrightarrow \quad \hat{\eta}_1 = \frac{1}{N}\sum_{i=1}^{N} x_i, \quad \hat{\eta}_2 = \frac{1}{N}\sum_{i=1}^{N} x_i^2$$

***q*-MLE**

Estimator preserving property

- MLE for normal distribution $\Longrightarrow \hat{\eta}_1 = \frac{1}{N}\sum_{i=1}^{N} x_i, \quad \hat{\eta}_2 = \frac{1}{N}\sum_{i=1}^{N} x_i^2$
- q-normal distribution (Student's t-distribution)
  - maximization of deformed entropy (Tsallis entropy)
  - infinite mixtures of normal distributions

$$p_q(x;\mu,\sigma) = \int_0^\infty N\left(\mu, \frac{1}{t}\right) \mathrm{Gamma}\left(t; \frac{3-q}{2(q-1)}, \frac{q-1}{3-q}\cdot\frac{2}{\sigma^2}\right) dt$$

Bayesian expression of q-normal distribution

- escort distributions, deformed algebras
  These are natural objects from the viewpoint of differential geometry.
- q-MLE for q-normal distributions $\Longrightarrow \hat{\eta}_1 = \frac{1}{N}\sum_{i=1}^{N} x_i, \ \hat{\eta}_2 = \frac{1}{N}\sum_{i=1}^{N} x_i^2$

A weight for parameter space and a weight for sample space are well balanced.

Concluding Remarks

U-divergence $\quad (S_\chi, g^M, \nabla^{M(e)}, \nabla^{M(m)})$

estimating function $u_\chi(x;\theta)$:

$$u_\chi^i(x;\theta) = \frac{\partial}{\partial\theta^i}\log_\chi p(x;\theta) - E_p\left[\frac{\partial}{\partial\theta^i}\log_\chi p(x;\theta)\right]$$

Riemannian metric g^M: $\quad g_{ij}^M(\theta) = \int_\Omega \partial_i p(x;\theta)\partial_j \log_\chi p(x;\theta)dx$

dual coordinates $\{\eta_i\}$: $\quad \eta_i = E_p[F_i(x)]$

χ-relative entropy $\quad (S_\chi, g^\chi, \nabla^{\chi(e)}, \nabla^{\chi(m)})$

estimating function $(s^\chi)(x;\theta)$:

$$(s^\chi)^i(x;\theta) = \frac{\partial}{\partial\theta^i}\log_\chi p(x;\theta) \text{ (unbiased under } \chi\text{-expectation)}$$

Riemannian metric g^χ: $\quad g_{ij}^\chi(\theta) = \frac{\partial^2}{\partial\theta^i\theta^j}\psi(\theta)$

dual coordinates $\{\eta_i\}$: $\quad \eta_i = E_{\chi,p}^{esc}[F_i(x)]$

The notion of expectations, independence are determined from a geometric structure of the statistical model.

Concluding Remarks

$P^{esc}_{\chi}(x)$: a normalized escort distribution of $p(x;\theta)$,

$$\overset{\text{def}}{\Longleftrightarrow} \quad P^{esc}_{\chi}(x;\theta) = \frac{1}{Z_{\chi}(\theta)}\chi(p(x;\theta)), \quad Z_{\chi}(\theta) = \int_{\Omega}\chi(p(x;\theta))dx$$

$E^{esc}_{\chi,p}[f(x)]$: the normalized χ-escort expectation of $p(x)$

$$\overset{\text{def}}{\Longleftrightarrow} E^{esc}_{\chi,p}[f(x)] = \int_{\Omega} f(x)P^{esc}_{\chi}(x;\theta)dx = \frac{1}{Z_{\chi}(\theta)}\int_{\Omega} f(x)\chi(p(x;\theta))dx$$

χ-relative entropy $\quad (S_{\chi}, g^{\chi}, \nabla^{\chi(e)}, \nabla^{\chi(m)})$

estimating function $(s^{\chi})(x;\theta)$:

$$(s^{\chi})^{i}(x;\theta) = \frac{\partial}{\partial\theta^{i}}\log_{\chi} p(x;\theta) \text{ (unbiased under } \chi\text{-expectation)}$$

Riemannian metric g^{χ}: $\quad g^{\chi}_{ij}(\theta) = \frac{\partial^2}{\partial\theta^i\theta^j}\psi(\theta)$

dual coordinates $\{\eta_i\}$: $\quad \eta_i = E^{esc}_{\chi,p}[F_i(x)]$

The notion of expectations, independence are determined from a geometric structure of the statistical model.

Projective transformation ((−1)-conf. transf.)

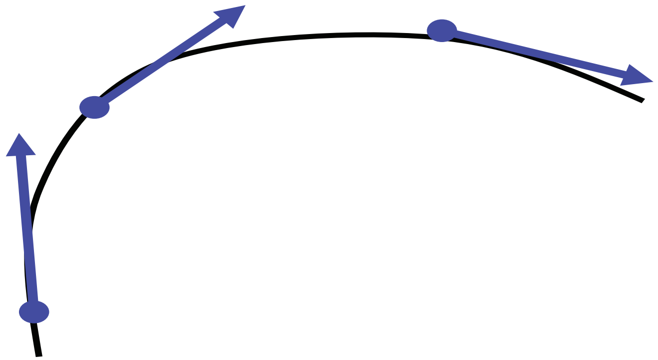
$c : I = (-\varepsilon, \varepsilon) \to M$:   a curve on M

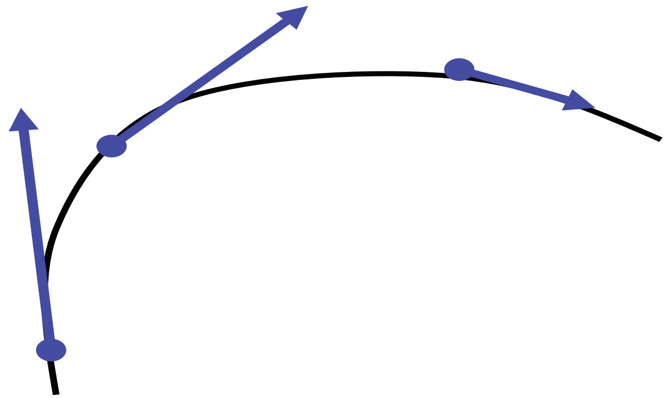
$$c \text{ is a geodesic} \iff \nabla_{\frac{d}{dt}} \dot{c} = 0$$

$$c \text{ is a pre-geodesic} \iff \nabla_{\frac{d}{dt}} \dot{c} = \gamma(t)\dot{c}$$

A projective transformation preserves pre-geodesics (unparametrized geodesics).

$$\nabla_{\frac{d}{dt}} \dot{c} = \beta(t)\dot{c} \iff \bar{\nabla}_{\frac{d}{dt}} \dot{c} = \bar{\gamma}(t)\dot{c}$$





Dual-projective transformation (1-conf. transf.)

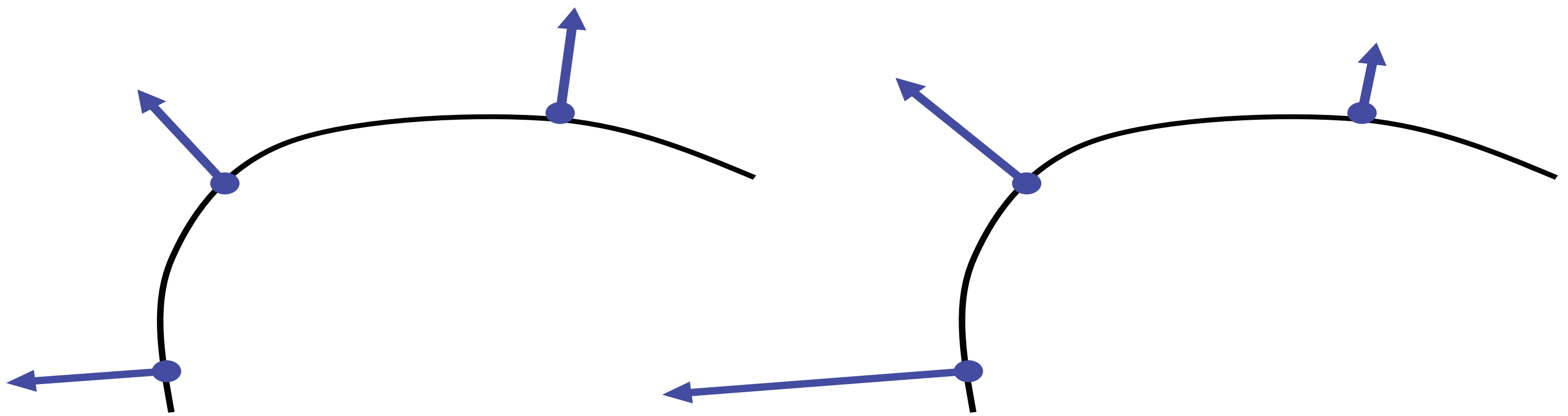
$c : I = (-\varepsilon, \varepsilon) \to M, \qquad \beta(X) = h(X, \dot{c}) :$ tangent 1-form

$$c \text{ is a dual geodesic} \iff \nabla_{\frac{d}{dt}}\beta = 0$$

$$c \text{ is a pre-dual geodesic} \iff \nabla_{\frac{d}{dt}}\beta = \gamma(t)\beta$$

A dual-projective transformation preserves pre-dual geodesics.

$$\nabla_{\frac{d}{dt}}\beta = \gamma(t)\beta \iff \bar{\nabla}_{\frac{d}{dt}}\beta = \bar{\gamma}(t)\beta$$



Definition 6.11

(M, ∇, h) and $(M, \bar{\nabla}, \bar{h})$ are 1-conformally equivalent

$\overset{\text{def}}{\Longleftrightarrow}$ There exists a function ψ such that

$$\bar{h}(X, Y) = e^{\psi} h(X, Y),$$
$$\bar{\nabla}_X Y = \nabla_X Y - h(X, Y)\mathrm{grad}_h \psi$$

Definition 6.12

(M, ∇, h) is 1-conformally flat

$\overset{\text{def}}{\Longleftrightarrow}$ (M, ∇, h) is locally 1-conformally equivalent to some flat statistical manifold.

Proposition 6.13

$\rho, \tilde{\rho}$: divergences (contrast functions) on M

$(M, \nabla, h), (M, \tilde{\nabla}, \tilde{h})$: induced statistical manifolds

(1) $\tilde{\rho}(p, q) = e^{\psi(q)} \rho(p, q) \quad \Longrightarrow$

(M, ∇, h) and $(M, \tilde{\nabla}, \tilde{h})$ are 1-conformally equivalent.

(2) $\tilde{\rho}(p, q) = e^{\psi(p)+\phi(q)} \rho(p, q) \quad \Longrightarrow$

(M, ∇, h) and $(M, \tilde{\nabla}, \tilde{h})$ are conformally-projectively equivalent.

Statistical inferences

Dually flat spaces

$(x_1, x_2, \ldots, x_N)$: N-independent observations

$L(\theta) = p(x_1;\theta)p(x_2;\theta)\cdots p(x_N;\theta)$

$\Longrightarrow$ Maximum likelihood estimator, Dually flat spaces

Generalized conformal geometry

$(x_1, x_2, \ldots, x_N)$: N-observations, but they are correlated.

$L_q(\theta) = p(x_1;\theta) \otimes_q p(x_2;\theta) \otimes_q \cdots \otimes_q p(x_N;\theta)$

$\Longrightarrow$ Tsallis statistics, deformed exponential family

$**$-conformally flat statistical manifolds

Non-integrable geometry

$(x_1, \ldots, x_N)$: independent, but we cannot observe.

Likelihood functions do not exist in the sense above.

$\Longrightarrow$ Non-conservative estimating function

Statistical manifolds admitting torsion

7 (Review) Geometry of statistical models

Definition 7.1

S is a statistical model or a parametric model on Ω

$\overset{\text{def}}{\iff}$ S is a set of probability densities with parameter $\xi \in \Xi$ s.t.

$$S = \left\{ p(x;\xi) \,\middle|\, \int_\Omega p(x;\xi)dx = 1, p(x;\xi) > 0, \xi \in \Xi \subset R^n \right\}.$$

We regard S as a manifold with a local coordinate system $\{\Xi; \xi^1, \ldots, \xi^n\}$

$g^F = (g^F_{ij})$ is the Fisher metric (Fisher information matrix) of S

$$\overset{\text{def}}{\iff} \quad g^F_{ij}(\xi) := \int_\Omega \frac{\partial}{\partial \xi^i} \log p(x;\xi) \frac{\partial}{\partial \xi^j} \log p(x;\xi) p(x;\xi) dx$$

$$= \int_\Omega \partial_i p_\xi \left(\frac{\partial}{\partial \xi^j} \log p_\xi \right) dx = E_\xi[\partial_i l_\xi \partial_j l_\xi]$$

$\partial_i p_\xi$ $\overset{\text{def}}{\iff}$ mixture representation,

$\partial_i l_\xi = \left(\dfrac{\partial_i p_\xi}{p_\xi} \right)$ $\overset{\text{def}}{\iff}$ exponential representation. (the score function)

For $\alpha \in R$, we define the α-connection $\nabla^{(\alpha)}$ by the following formula:

$$\Gamma_{ij,k}^{(\alpha)}(\xi) = E_\xi\left[\left(\partial_i\partial_j l_\xi + \frac{1-\alpha}{2}\partial_i l_\xi \partial_j l_\xi\right)(\partial_k l_\xi)\right]$$

$$g(\nabla_{\partial_i}^{(\alpha)}\partial_j, \partial_k) = \Gamma_{ij,k}^{(\alpha)}$$

We can check that $\nabla^{(\alpha)}$ ($^\forall\alpha \in R$) is torsion-free and $\nabla^{(0)}$ is the Levi-Civita connection of the Fisher metrics. On the other hand,

$\nabla^{(e)} := \nabla^{(1)}$: the exponential connection

$\nabla^{(m)} := \nabla^{(-1)}$: the mixture connection

(1) $\partial_i g(\partial_j, \partial_k) = g(\nabla_{\partial_i}^{(\alpha)}\partial_j, \partial_k) + g(\partial_j, \nabla_{\partial_i}^{(-\alpha)}\partial_k)$

($\nabla^{(\alpha)}$ and $\nabla^{(-\alpha)}$ are called dual (or conjugate) with respect to g)

(2) $g(\nabla_{\partial_i}^{(\alpha)}\partial_j, \partial_k) = g(\nabla_{\partial_i}^{(0)}\partial_j, \partial_k) - \dfrac{\alpha}{2}C(\partial_i, \partial_j, \partial_k)$

$C_\xi(\partial_i, \partial_j, \partial_k) := E_\xi[(\partial_i l_\xi)(\partial_j l_\xi)(\partial_k l_\xi)]$: the skewness or the cubic form

(3) $(\nabla_{\partial_i}^{(\alpha)} g)(\partial_j, \partial_k) = (\nabla_{\partial_j}^{(\alpha)} g)(\partial_i, \partial_k) = \alpha C(\partial_i, \partial_j, \partial_k)$

We say that $(S, \nabla^{(\alpha)}, g^F)$ is an invariant statistical manifold.

A statistical model S_e is an exponential family

$$\overset{\text{def}}{\Longleftrightarrow} \qquad S_e = \left\{ p(x;\theta) \;\middle|\; p(x;\theta) = \exp[Z(x) + \sum_{i=1}^{n} \theta^i F_i(x) - \psi(\theta)] \right\},$$

$Z, F_1, \cdots, F_n$: functions on Ω

ψ : a function on the parameter space Θ

The coordinate system $[\theta^i]$ is called the natural parameters.

Proposition 7.2 For an exponential family S_e,
(1) $\nabla^{(e)} := \nabla^{(1)}$ is flat
(2) $[\theta^i]$ is an affine coordinate, i.e., $\Gamma_{ij}^{(1)\,k} \equiv 0$

For simplicity, assume that $Z = 0$.

$$\begin{aligned} g_{ij}^F(\theta) &= E[(\partial_i \log p(x;\theta))(\partial_j \log p(x;\theta))] \\ &= \partial_i \partial_j \psi(\theta) \quad \text{:the Fisher metric} \\ C_{ijk}^F(\theta) &= E[(\partial_i \log p(x;\theta))(\partial_j \log p(x;\theta))(\partial_k \log p(x;\theta))] \\ &= \partial_i \partial_j \partial_k \psi(\theta) \quad \text{:the cubic form} \end{aligned}$$

The quadruplets $(S_e, g^F, \nabla^{(e)}, \nabla^{(m)})$ is a dually flat space.
The triplets $(S_e, \nabla^{(e)}, g^F)$ and $(S_e, \nabla^{(m)}, g^F)$ are Hessian manifolds,
$(S, \nabla^{(e)}, g^F)$ and $(S, \nabla^{(m)}, g^F)$ are flat statistical manifolds.

Normal distributions

$\Omega = R,\ n = 2,\ \xi = (\mu, \sigma) \in R^2_+$ (the upper half plane).

$$S = \left\{ p(x;\mu,\sigma) \ \middle|\ p(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-u)^2}{2\sigma^2}\right] \right\}$$

The Fisher metric is

$$(g_{ij}) = \frac{1}{\sigma^2}\begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix} \quad \left(S \text{ is a space of constant negative curvature } -\frac{1}{2} \right).$$

$\nabla^{(1)}$ and $\nabla^{(-1)}$ are flat affine connections. In addition,

$$\theta^1 = \frac{\mu}{\sigma^2},\ \theta^2 = -\frac{1}{2\sigma^2} \quad \psi(\theta) = -\frac{(\theta^1)^2}{4\theta^2} + \frac{1}{2}\log\left(-\frac{\pi}{\theta^2}\right)$$

$$\Longrightarrow p(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(x-u)^2}{2\sigma^2}\right] = \exp\left[x\theta^1 + x^2\theta^2 - \psi(\theta)\right].$$

$\{\theta^1, \theta^2\}$: natural parameters. ($\nabla^{(1)}$-geodesic coordinate system)

$\eta_1 = E[x] = \mu, \quad \eta_2 = E\left[x^2\right] = \sigma^2 + \mu^2.$

$\{\eta_1, \eta_2\}$: moment parameters. ($\nabla^{(-1)}$-geodesic coordinate system)

Finite sample space

$$\Omega = \{x_0, x_1, \cdots, x_n\}, \ \dim S_n = n$$

$$p(x_i;\eta) = \begin{cases} \eta_i & (1 \le i \le n) \\ 1 - \sum_{j=1}^n \eta_j & (i = 0) \end{cases}$$

$$\Xi = \left\{ \{\eta_1, \cdots, \eta_n\} \;\middle|\; \eta_i > 0 \ ({}^\forall i), \ \sum_{j=1}^n \eta_j < 1 \right\}$$

(an n-dimensional simplex)

The Fisher metric:

$$(g_{ij}) = \frac{1}{\eta_0} \begin{pmatrix} 1 + \frac{\eta_0}{\eta_1} & 1 & \cdots & 1 \\ 1 & 1 + \frac{\eta_0}{\eta_2} & & \vdots \\ \vdots & & \ddots & \vdots \\ 1 & \cdots & \cdots & 1 + \frac{\eta_0}{\eta_n} \end{pmatrix},$$

where $\eta_0 = 1 - \sum_{j=1}^n \eta_j$.

$\left(S_n \text{ is a space of constant positive curvature } \frac{1}{4} \right)$.

Finite sample space

$\Omega = \{x_0, x_1, \cdots, x_n\}$, $\dim S_n = n$

$$p(x_i; \eta) = \begin{cases} \eta_i & (1 \leq i \leq n) \\ 1 - \sum_{j=1}^{n} \eta_j & (i = 0) \end{cases}$$

$$\Xi = \left\{ \{\eta_1, \cdots, \eta_n\} \;\middle|\; \eta_i > 0 \; ({}^{\forall} i), \; \sum_{j=1}^{n} \eta_j < 1 \right\}$$

(an n-dimensional simplex)

$\{\theta^1, \cdots, \theta^n\}$: natural parameters. ($\nabla^{(1)}$-geodesic coordinate system)

where $\theta^i = \log p(x_i) - \log p(x_0) = \log \dfrac{\eta_i}{1 - \sum_{j=1}^{n} \eta_j}$

$$\psi(\theta) = \log \left(1 + \sum_{j=1}^{n} e^{\theta^j} \right)$$

$\{\eta_1, \cdots, \eta_n\}$: moment parameters. ($\nabla^{(-1)}$-geodesic coordinate system)

Proposition 7.3 For S_e, the following hold:

(1) $(S_e, g^F, \nabla^{(e)}, \nabla^{(m)})$ is a dually flat space.

(2) $\{\theta^i\}$ is a $\nabla^{(e)}$-affine coordinate system on S_e.

(3) $\psi(\theta)$ is the potential of g^F w.r.t. $\{\theta^i\}$:

$$g^F_{ij}(\theta) = \partial_i \partial_j \psi(\theta).$$

(4) Set the expectations of $F_i(x)$ by $\eta_i = E_\theta[F_i(x)]$
$\Longrightarrow$ $\{\eta_i\}$ is the dual coordinate system of $\{\theta^i\}$ with respect to g^M.

(5) Set $\phi(\eta) = E_\theta[\log p_\theta]$.
$\Longrightarrow$ $\phi(\eta)$ is the potential of g^F w.r.t. $\{\eta_i\}$.

Since $(S_e, g^F, \nabla^{(e)}, \nabla^{(m)})$ is a dually flat space, the Legendre transformation holds.

$$\frac{\partial \psi}{\partial \theta^i} = \eta_i, \quad \frac{\partial \phi}{\partial \eta_i} = \theta^i, \quad \psi(p) + \phi(p) - \sum_{i=1}^{m} \theta^i(p)\eta_i(p) = 0$$

$$g^F_{ij} = \frac{\partial^2 \psi}{\partial \theta^i \partial \theta^j}, \quad C^F_{ijk} = \frac{\partial^3 \psi}{\partial \theta^i \partial \theta^j \partial \theta^k}$$

Kullback-Leibler divergence (or relative entropy) on S

$$\overset{\text{def}}{\Longleftrightarrow} \quad D_{KL}(p,r) = \int_\Omega p(x)\log\frac{p(x)}{r(x)}dx$$
$$= E_p[\log p(x) - \log r(x)]$$
$$\left(= \psi(r) + \phi(p) - \sum_{i=1}^{n}\theta^i(r)\eta_i(p) \ = \ D(r,p) \ \right)$$

For S_e, D_{KL} coincides with the canonical divergence D on a dually flat space $(S_e, \nabla^{(m)}, g^F)$.

Construction of a divergence from an estimating function

$$s(x;\xi) = \begin{pmatrix} \partial/\partial\xi^1 \log p(x;\xi) \\ \vdots \\ \partial/\partial\xi^n \log p(x;\xi) \end{pmatrix}:$$ the score function of $p(x;\xi)$ (estimating function)

by integrating of the score function and by taking an expectation,

$$d_{KL}(p,r) := \int_\Omega p(x;\xi)\log r(x;\xi')dx \qquad \text{the cross entropy on } S$$

The KL-divergence is given by the difference of cross entropies.

$$D_{KL}(p,r) = d_{KL}(p,p) - d_{KL}(p,r)$$