The Infinitesimal Invariance Criterion for Statistical Transformation Models

Linyu Peng (彭　林玉)

http://www.peng.mech.keio.ac.jp

Department of Mechanical Engineering
Keio University (慶應義塾大学)

CMCAA, Beijing, September 13th, 2020

慶應義塾
Keio University
Tokyo, Japan

Outline

Statistical manifolds $(\mathcal{S}^n, g)$

- A statistical model:

$$\mathcal{S} = \Big\{ p(x;\theta) \mid x \in \Omega \subseteq \mathbb{R}^m, \quad \theta \in \Theta \subseteq \mathbb{R}^n \Big\}$$

Here, $p(x;\theta)$ are probability density functions (pdfs).

Statistical manifolds $(\mathcal{S}^n, g)$

- A statistical model:

$$\mathcal{S} = \left\{ p(x;\theta) \mid x \in \Omega \subseteq \mathbb{R}^m, \quad \theta \in \Theta \subseteq \mathbb{R}^n \right\}$$

  Here, $p(x;\theta)$ are probability density functions (pdfs).

- The Kullback–Leibler (KL) divergence on $\mathcal{S}$ (Kullback–Leibler, 1951):

$$D_{\mathrm{KL}} : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$$

$$(p_1, p_2) \mapsto D_{\mathrm{KL}}(p_1, p_2) := \int_\Omega p(x;\theta_1) \ln \frac{p(x;\theta_1)}{p(x;\theta_2)} \, \mathrm{d}x$$

Statistical manifolds $(\mathcal{S}^n, g)$

- A statistical model:

$$\mathcal{S} = \left\{ p(x;\theta) \mid x \in \Omega \subseteq \mathbb{R}^m, \quad \theta \in \Theta \subseteq \mathbb{R}^n \right\}$$

  Here, $p(x;\theta)$ are probability density functions (pdfs).

- The Kullback–Leibler (KL) divergence on $\mathcal{S}$ (Kullback–Leibler, 1951):

$$\begin{aligned} D_{\mathrm{KL}} &: \mathcal{S} \times \mathcal{S} \to \mathbb{R} \\ &(p_1, p_2) \mapsto D_{\mathrm{KL}}(p_1, p_2) := \int_\Omega p(x;\theta_1) \ln \frac{p(x;\theta_1)}{p(x;\theta_2)} \, \mathrm{d}x \end{aligned}$$

- The Fisher information matrix g (Fisher, 1922) can be derived from

$$D_{\mathrm{KL}}\left(p(x;\theta), p(x;\theta + \mathrm{d}\theta)\right) = \frac{1}{2} g_{ij}(\theta) \, \mathrm{d}\theta^i \, \mathrm{d}\theta^j + O\left((\mathrm{d}\theta)^3\right)$$

- Entries of the matrix:

$$\begin{aligned} g_{ij}(\theta) &= E[\partial_i \ln p \quad \partial_j \ln p] \\ &= \int_\Omega \partial_i \ln p(x;\theta) \partial_j \ln p(x;\theta) p(x;\theta)\, \mathrm{d}x \end{aligned}$$

Note $\partial_i = \frac{\partial}{\partial \theta^i}$ and $i, j = 1, 2, \ldots, n$.

- Entries of the matrix:

$$\begin{aligned} g_{ij}(\theta) &= E[\partial_i \ln p \quad \partial_j \ln p] \\ &= \int_\Omega \partial_i \ln p(x;\theta) \partial_j \ln p(x;\theta) p(x;\theta) \, \mathrm{d}x \end{aligned}$$

  Note $\partial_i = \frac{\partial}{\partial \theta^i}$ and $i, j = 1, 2, \ldots, n$.

- The corresponding Riemannian metric (Rao, 1945):

$$g(\partial_i, \partial_j) := g_{ij}(\theta)$$

- Entries of the matrix:

$$g_{ij}(\theta) = E[\partial_i \ln p \quad \partial_j \ln p]$$
$$= \int_\Omega \partial_i \ln p(x;\theta) \partial_j \ln p(x;\theta) p(x;\theta) \, \mathrm{d}x$$

Note $\partial_i = \frac{\partial}{\partial \theta^i}$ and $i, j = 1, 2, \ldots, n$.

- The corresponding Riemannian metric (Rao, 1945):

$$g(\partial_i, \partial_j) := g_{ij}(\theta)$$

Definition. The n-dimensional Riemannian manifold $(\mathcal{S}^n, g)$ is called a *statistical manifold*.

Levi-Civita connection

The unique Levi-Civita connection $\nabla^{(0)}$ satisfies

- Torsion free:

$$\nabla^{(0)}_X Y - \nabla^{(0)}_Y X = [X, Y], \quad \forall X, Y \in \mathfrak{X}(\mathcal{S})$$

- Compatibility with the metric g: $\nabla^{(0)} g = 0$, i.e.,

$$Zg(X, Y) = g(\nabla^{(0)}_Z X, Y) + g(X, \nabla^{(0)}_Z Y), \quad \forall X, Y, Z \in \mathfrak{X}(\mathcal{S})$$

Levi-Civita connection

The unique Levi-Civita connection $\nabla^{(0)}$ satisfies

- Torsion free:

$$\nabla^{(0)}_X Y - \nabla^{(0)}_Y X = [X, Y], \quad \forall X, Y \in \mathfrak{X}(\mathcal{S})$$

- Compatibility with the metric g: $\nabla^{(0)} g = 0$, i.e.,

$$Zg(X, Y) = g(\nabla^{(0)}_Z X, Y) + g(X, \nabla^{(0)}_Z Y), \quad \forall X, Y, Z \in \mathfrak{X}(\mathcal{S})$$

Locally,

$$g\left(\nabla^{(0)}_{\partial_i} \partial_j, \partial_k\right) = \Gamma^{(0)}_{ij,k},$$

where

$$\Gamma^{(0)}_{ij,k} = \frac{1}{2}\left(\partial_i g_{jk} + \partial_j g_{ki} - \partial_k g_{ij}\right)$$

Dual affine connections

Some history of dual connections for statistical models:

- Chentsov, 1972 and before: Introduced a family of dual connections but only used the Riemannian structure (Originally in Russian, English translation published in 1982)
- Efron, 1975: Defined a curvature (independently from Chentsov) but did not realise it corresponds to the exponential connection
- Dawid, 1975: Showed the relation between Efron's curvature and the exponential connection, also suggested to define the mixture connection
- Amari, 1980, 1982: Defined a one-parameter family of affine connections, i.e., α-connections, that are *equivalent* to Chentsov's ones

Dual affine connections

A pair of affine connections ∇ and ∇^* are dual to each other if they satisfy

- Torsion free
- Duality condition:

$$Zg(X,Y) = g(\nabla_Z X, Y) + g(X, \nabla^*_Z Y), \quad \forall X, Y, Z \in \mathfrak{X}(\mathcal{S})$$

Dual affine connections

A pair of affine connections ∇ and ∇^* are dual to each other if they satisfy

- Torsion free
- Duality condition:

$$Zg(X, Y) = g(\nabla_Z X, Y) + g(X, \nabla_Z^* Y), \quad \forall X, Y, Z \in \mathfrak{X}(\mathcal{S})$$

Remark. 1. The Levi-Civita connection is

$$\nabla^{(0)} = \frac{\nabla + \nabla^*}{2}.$$

2. For any statistical manifold $\mathcal{S}$, there exists a one-parameter family of connections $\nabla^{(\alpha)}$ ($\alpha \in \mathbb{R}$) such that $\nabla^{(\alpha)}$ and $\nabla^{(-\alpha)}$ are dual.

Example: Gaussian distributions

- pdfs:

$$p(x;\theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, \quad x \in \mathbb{R}, \theta = (\mu, \sigma) \in \mathbb{R} \times \mathbb{R}^+$$

- Fisher information matrix:

$$g(\theta) = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{\sigma^2} \end{pmatrix}$$

- Constant curvature:

$$-\frac{1}{2}$$

Example: Weibull distributions

- pdfs:

$$p(x;\theta) = \frac{\beta}{\alpha}\left(\frac{x}{\alpha}\right)^{\beta-1}\exp\left\{-\left(\frac{x}{\alpha}\right)^{\beta}\right\}, \quad x \in \mathbb{R}^+, \theta = (\alpha,\beta) \in \mathbb{R}^+ \times \mathbb{R}^+$$

Example: Weibull distributions

- pdfs:

$$p(x;\theta) = \frac{\beta}{\alpha}\left(\frac{x}{\alpha}\right)^{\beta-1} \exp\left\{-\left(\frac{x}{\alpha}\right)^{\beta}\right\}, \quad x \in \mathbb{R}^+, \theta = (\alpha, \beta) \in \mathbb{R}^+ \times \mathbb{R}^+$$

- Fisher information matrix:

$$g(\theta) = \begin{pmatrix} \frac{\beta^2}{\alpha^2} & \frac{\gamma-1}{\alpha} \\ \frac{\gamma-1}{\alpha} & \frac{(\gamma-1)^2}{\beta^2} + \frac{\pi^2}{6\beta^2} \end{pmatrix}$$

The number γ is the Euler–Mascheroni constant, equaling

$$\gamma = -\int_0^{+\infty} e^{-x} \ln x \, \mathrm{d}x$$

Example: Weibull distributions

- pdfs:

$$p(x;\theta) = \frac{\beta}{\alpha}\left(\frac{x}{\alpha}\right)^{\beta-1}\exp\left\{-\left(\frac{x}{\alpha}\right)^{\beta}\right\}, \quad x \in \mathbb{R}^+, \theta = (\alpha,\beta) \in \mathbb{R}^+ \times \mathbb{R}^+$$

- Fisher information matrix:

$$g(\theta) = \begin{pmatrix} \frac{\beta^2}{\alpha^2} & \frac{\gamma-1}{\alpha} \\ \frac{\gamma-1}{\alpha} & \frac{(\gamma-1)^2}{\beta^2} + \frac{\pi^2}{6\beta^2} \end{pmatrix}$$

The number γ is the Euler–Mascheroni constant, equaling

$$\gamma = -\int_0^{+\infty} e^{-x}\ln x \,\mathrm{d}x$$

- Constant curvature (Cao–Sun–Wang, 2008):

$$-\frac{6}{\pi^2}$$

Natural gradient descent

Definition. Consider extrema of a function $J(\theta)$ defined on ta statistical manifold $(\mathcal{S}, g)$. The steepest descent direction is given by the natural gradient (Amari, 1997, 1998)

$$-\operatorname{grad}_N J(\theta) := -(g_{ij}(\theta))^{-1} \operatorname{grad} J(\theta).$$

Natural gradient descent

Definition. Consider extrema of a function $J(\theta)$ defined on ta statistical manifold $(\mathcal{S}, g)$. The steepest descent direction is given by the natural gradient (Amari, 1997, 1998)

$$-\operatorname{grad}_N J(\theta) := -(g_{ij}(\theta))^{-1} \operatorname{grad} J(\theta).$$

A natural gradient descent method can then be defined as a generalisation of Newton's gradient descent method on statistical manifolds:

$$\theta_{k+1} = \theta_k - h \operatorname{grad}_N J(\theta_k).$$

Natural gradient descent

Definition. Consider extrema of a function $J(\theta)$ defined on ta statistical manifold $(\mathcal{S}, g)$. The steepest descent direction is given by the natural gradient (Amari, 1997, 1998)

$$-\operatorname{grad}_N J(\theta) := -(g_{ij}(\theta))^{-1} \operatorname{grad} J(\theta).$$

A natural gradient descent method can then be defined as a generalisation of Newton's gradient descent method on statistical manifolds:

$$\theta_{k+1} = \theta_k - h \operatorname{grad}_N J(\theta_k).$$

The *difficulty* lies in the computation of matrix inversion $(g_{ij}(\theta_k))^{-1}$ for each k, especially when $\dim \mathcal{S}$ is big.

Group actions

A **group of transformations** (or a (left) **group action**) acting on a smooth manifold $\mathcal{M}$ is given by a (local) Lie group G, and a smooth map $\mathcal{T} : G \times \mathcal{M} \to \mathcal{M}$ satisfying:

- $\mathcal{T}(\rho_1, \mathcal{T}(\rho_2, z)) = \mathcal{T}((\rho_1 \cdot \rho_2), z)$ and $\mathcal{T}(e, z) = z$.

Group actions

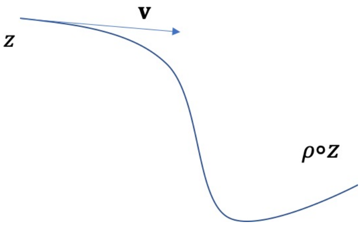
A **group of transformations** (or a (left) **group action**) acting on a smooth manifold $\mathcal{M}$ is given by a (local) Lie group G, and a smooth map $\mathcal{T} : G \times \mathcal{M} \to \mathcal{M}$ satisfying:

- $\mathcal{T}(\rho_1, \mathcal{T}(\rho_2, z)) = \mathcal{T}((\rho_1 \cdot \rho_2), z)$ and $\mathcal{T}(e, z) = z$.

Remark. For any $\rho \in G$, we denote $\mathcal{T}_\rho : \mathcal{M} \to \mathcal{M}$ by

$$\mathcal{T}_\rho(z) = \mathcal{T}(\rho, z) = \rho \circ z = \widetilde{z}.$$

Infinitesimal generators



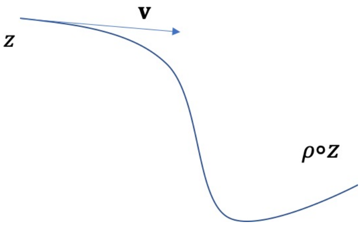
Locally, in a small neighbourhood of e, the group G can be parameterised by $\rho = (\rho^1, \rho^2, \ldots, \rho^r)$, where $r = \dim G$. The **infinitesimal generators** are defined as

$$\mathbf{v}_i = \xi_i^j(z)\partial_{z^j},$$

where

$$\xi_i^j(z) = \frac{\partial \widetilde{z}^j}{\partial \rho^i}\bigg|_{\rho=e}.$$

Infinitesimal generators



Locally, in a small neighbourhood of e, the group G can be parameterised by $\rho = (\rho^1, \rho^2, \dots, \rho^r)$, where $r = \dim G$. The **infinitesimal generators** are defined as

$$\mathbf{v}_i = \xi_i^j(z)\partial_{z^j},$$

where

$$\xi_i^j(z) = \frac{\partial \widetilde{z}^j}{\partial \rho^i}\bigg|_{\rho=e}.$$

Remark. Group actions and infinitesimal generators are connected by a system of linear PDEs:

$$\frac{\partial \widetilde{z}^j}{\partial \rho^i} = \xi_i^j(\widetilde{z})$$

subject to initial conditions

$$\widetilde{z}\Big|_{\rho=e} = z.$$

Example

Consider the special orthogonal group $G = SO(2)$ acting on the plane $\mathbb{R}^2$ (i.e., rotations):

$$\begin{pmatrix} x \\ y \end{pmatrix} \mapsto \begin{pmatrix} \widetilde{x} \\ \widetilde{y} \end{pmatrix} = \begin{pmatrix} \cos\varepsilon & -\sin\varepsilon \\ \sin\varepsilon & \cos\varepsilon \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}.$$

Example

Consider the special orthogonal group $G = SO(2)$ acting on the plane $\mathbb{R}^2$ (i.e., rotations):

$$\begin{pmatrix} x \\ y \end{pmatrix} \mapsto \begin{pmatrix} \widetilde{x} \\ \widetilde{y} \end{pmatrix} = \begin{pmatrix} \cos\varepsilon & -\sin\varepsilon \\ \sin\varepsilon & \cos\varepsilon \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}.$$

The infinitesimal generator is

$$\begin{aligned} \mathbf{v} &= \frac{\mathrm{d}\widetilde{x}}{\mathrm{d}\varepsilon}\Big|_{\varepsilon=0} \partial_x + \frac{\mathrm{d}\widetilde{y}}{\mathrm{d}\varepsilon}\Big|_{\varepsilon=0} \partial_y \\ &= -y\partial_x + x\partial_y, \end{aligned}$$

Invariance of functions

Definition. A smooth function $f(z)$ ($z \in \mathcal{M}$) is called invariant w.r.t. a group G acting on $\mathcal{M}$ if we have

$$f(z) = f(\rho \circ z), \quad \forall \rho \in G.$$

For instance, $f(x, y) = x^2 + y^2$ is invariant w.r.t. rotations in $\mathbb{R}^2$.

Invariance of functions

Definition. A smooth function $f(z)$ $(z \in \mathcal{M})$ is called invariant w.r.t. a group G acting on $\mathcal{M}$ if we have

$$f(z) = f(\rho \circ z), \quad \forall \rho \in G.$$

For instance, $f(x, y) = x^2 + y^2$ is invariant w.r.t. rotations in $\mathbb{R}^2$.

Theorem. A smooth function $f(z)$ $(z \in \mathcal{M})$ is invariant w.r.t. a group G acting on $\mathcal{M}$ *if and only if* for each infinitesimal generator $\mathbf{v}$, the following vanishment holds

$$\mathbf{v}(f) \equiv 0.$$

Invariance of integrals

Definition. Let $f(z)$ be a smooth function in $\mathcal{M}$. An integral $\int_\Omega f(z)\,\mathrm{d}z$, defined in an open, connected subspace $\Omega \subseteq \mathcal{M}$ with smooth boundary, is called invariant w.r.t. a group G acting on Ω if we have

$$\int_{\Omega_0} f(z)\,\mathrm{d}z = \int_{\rho\circ\Omega_0} f(\rho\circ z)\,\mathrm{d}(\rho\circ z), \quad \forall \rho \in G$$

for any subdomain Ω_0 such that $\overline{\Omega}_0 \subseteq \Omega$, or alternatively,

$$f(z)\,\mathrm{d}z = f(\rho\circ z)\,\mathrm{d}(\rho\circ z), \quad \forall \rho \in G.$$

Invariance of integrals

Definition. Let $f(z)$ be a smooth function in $\mathcal{M}$. An integral $\int_\Omega f(z)\,\mathrm{d}z$, defined in an open, connected subspace $\Omega \subseteq \mathcal{M}$ with smooth boundary, is called invariant w.r.t. a group G acting on Ω if we have

$$\int_{\Omega_0} f(z)\,\mathrm{d}z = \int_{\rho\circ\Omega_0} f(\rho\circ z)\,\mathrm{d}(\rho\circ z), \quad \forall \rho \in G$$

for any subdomain Ω_0 such that $\overline{\Omega}_0 \subseteq \Omega$, or alternatively,

$$f(z)\,\mathrm{d}z = f(\rho\circ z)\,\mathrm{d}(\rho\circ z), \quad \forall \rho \in G.$$

Theorem. Under the same assumptions of the definition above, an integral $\int_\Omega f(z)\,\mathrm{d}z$ is invariant *if and only if* the following identity holds for each infinitesimal generator $\mathbf{v} = \xi^i(z)\partial_{z^i}$:

$$\mathbf{v}(f) + f\,\mathrm{Div}\,\xi \equiv 0, \text{ where } \mathrm{Div}\,\xi := D_{z^i}\xi^i.$$

Group actions on measurable/Borel spaces

- Let $(\mathcal{X}, \mathcal{B})$ be a measurable space.

Group actions on measurable/Borel spaces

- Let $(\mathcal{X}, \mathcal{B})$ be a measurable space.
- Let ν be an arbitrary measure on $(\mathcal{X}, \mathcal{B})$. For a function $f \in L^1(\nu)$, we have

$$\nu(f) = \int_{\mathcal{X}} f(x)\nu(\mathrm{d}x).$$

Group actions on measurable/Borel spaces

- Let $(\mathcal{X}, \mathcal{B})$ be a measurable space.
- Let ν be an arbitrary measure on $(\mathcal{X}, \mathcal{B})$. For a function $f \in L^1(\nu)$, we have
$$\nu(f) = \int_{\mathcal{X}} f(x)\nu(\mathrm{d}x).$$
- Consider a group action
$$\begin{aligned} \mathcal{T} : G \times \mathcal{X} &\to \mathcal{X} \\ (\rho, x) &\mapsto \widetilde{x} = \rho \circ x, \end{aligned}$$
which induces transformations on a measure ν:
$$\rho \circ \nu(f) := \nu(f \circ \rho), \quad f \in L^1(\nu).$$

Group actions on measurable/Borel spaces

- Let $(\mathcal{X}, \mathcal{B})$ be a measurable space.
- Let ν be an arbitrary measure on $(\mathcal{X}, \mathcal{B})$. For a function $f \in L^1(\nu)$, we have

$$\nu(f) = \int_{\mathcal{X}} f(x)\nu(\mathrm{d}x).$$

- Consider a group action

$$\mathcal{T} : G \times \mathcal{X} \to \mathcal{X}$$
$$(\rho, x) \mapsto \widetilde{x} = \rho \circ x,$$

which induces transformations on a measure ν:

$$\rho \circ \nu(f) := \nu(f \circ \rho), \quad f \in L^1(\nu).$$

Definition. A measure ν is said to be invariant w.r.t. the group action $\mathcal{T}$ if

$$\rho \circ \nu = \nu, \quad \forall \rho \in G.$$

Probability measure

- Let X be a random variable in the measurable space $(\mathcal{X}, \mathcal{B})$ corresponding to a probability measure P on $(\mathcal{X}, \mathcal{B})$.

Probability measure

- Let X be a random variable in the measurable space $(\mathcal{X}, \mathcal{B})$ corresponding to a probability measure P on $(\mathcal{X}, \mathcal{B})$.
- The density of X w.r.t. a reference measure μ on $(\mathcal{X}, \mathcal{B})$ is derived using the Radon–Nikodym derivative:

$$p = \frac{\mathrm{d}P}{\mathrm{d}\mu}, \text{ or equivalently, } \mathrm{d}P = p \, \mathrm{d}\mu.$$

Probability measure

- Let X be a random variable in the measurable space $(\mathcal{X}, \mathcal{B})$ corresponding to a probability measure P on $(\mathcal{X}, \mathcal{B})$.
- The density of X w.r.t. a reference measure μ on $(\mathcal{X}, \mathcal{B})$ is derived using the Radon–Nikodym derivative:

$$p = \frac{\mathrm{d}P}{\mathrm{d}\mu}, \text{ or equivalently, } \mathrm{d}P = p\,\mathrm{d}\mu.$$

- The probability measure P is invariant w.r.t. a group action $\mathcal{T}$ if $\rho \circ P = P$, that, locally, is written as

$$P(\mathrm{d}x) = P\left(\mathrm{d}\widetilde{x}\right), \text{ i.e., } p(x)\mu(\mathrm{d}x) = p(\widetilde{x})\mu(\mathrm{d}\widetilde{x}).$$

Probability measure

- Let X be a random variable in the measurable space $(\mathcal{X}, \mathcal{B})$ corresponding to a probability measure P on $(\mathcal{X}, \mathcal{B})$.
- The density of X w.r.t. a reference measure μ on $(\mathcal{X}, \mathcal{B})$ is derived using the Radon–Nikodym derivative:

$$p = \frac{\mathrm{d}P}{\mathrm{d}\mu}, \text{ or equivalently, } \mathrm{d}P = p\,\mathrm{d}\mu.$$

- The probability measure P is invariant w.r.t. a group action $\mathcal{T}$ if $\rho \circ P = P$, that, locally, is written as

$$P(\mathrm{d}x) = P\left(\mathrm{d}\widetilde{x}\right), \text{ i.e., } p(x)\mu(\mathrm{d}x) = p(\widetilde{x})\mu(\mathrm{d}\widetilde{x}).$$

- Further assume μ is the Lebesgue measure, then the invariance becomes

$$p(x)\,\mathrm{d}x = p(\widetilde{x})\,\mathrm{d}\widetilde{x}$$

Statistical transformation models

Definition. Let $p(x;\theta)$ be the pdfs where $x \in \Omega \subseteq \mathbb{R}^m$ and $\theta \in \Theta$ with Θ an n-dimensional Lie group. The statistical model $\mathcal{S} = \{p(x;\theta\}$ is called a **transformation model** if there exists a group action $\mathcal{T} : \Theta \times \Omega \to \Omega$ such that the probability measure is invariant in the sense that

$$p(x;\theta)\,\mathrm{d}x = p(\widetilde{x};\rho\cdot\theta)\,\mathrm{d}\widetilde{x}, \quad \forall\rho\in\Theta,$$

where $\widetilde{x} = \rho \circ x$.

Statistical transformation models

Definition. Let $p(x;\theta)$ be the pdfs where $x \in \Omega \subseteq \mathbb{R}^m$ and $\theta \in \Theta$ with Θ an n-dimensional Lie group. The statistical model $\mathcal{S} = \{p(x;\theta\}$ is called a **transformation model** if there exists a group action $\mathcal{T} : \Theta \times \Omega \to \Omega$ such that the probability measure is invariant in the sense that

$$\boxed{p(x;\theta)\,\mathrm{d}x = p(\widetilde{x};\rho\cdot\theta)\,\mathrm{d}\widetilde{x}, \quad \forall \rho \in \Theta,}$$

where $\widetilde{x} = \rho \circ x$.

Remark. This is in fact a special transformation model according to Barndorff-Nielsen–Blæsild–Eriksen, 1989.

Example. The Gaussian distributions form a transformation model.

Example. The Gaussian distributions form a transformation model.

- Lie group structure of $\Theta = \{\rho = (\mu, \sigma) \mid \mu \in \mathbb{R}, \sigma \in \mathbb{R}^+\}$ (non-Abelian):

$$(\mu_1, \sigma_1) \cdot (\mu_2, \sigma_2) = (\mu_1 + \mu_2\sigma_1, \sigma_1\sigma_2) .$$

  - Identity:

$$e = (0, 1)$$

  - Inversion:

$$\rho^{-1} = \left(-\frac{\mu}{\sigma}, \frac{1}{\sigma}\right)$$

Example. The Gaussian distributions form a transformation model.

- Lie group structure of $\Theta = \{\rho = (\mu, \sigma) \mid \mu \in \mathbb{R}, \sigma \in \mathbb{R}^+\}$ (non-Abelian):

$$(\mu_1, \sigma_1) \cdot (\mu_2, \sigma_2) = (\mu_1 + \mu_2\sigma_1, \sigma_1\sigma_2).$$

  - Identity:

$$e = (0, 1)$$

  - Inversion:

$$\rho^{-1} = \left(-\frac{\mu}{\sigma}, \frac{1}{\sigma}\right)$$

- The group action:

$$\rho \circ x = \mu + \sigma x.$$

Theorem. (Amari–Nagaoka, 1993) Components of the Fisher information matrix g satisfy

$$g_{ij}(\theta) = B_i^l(\theta) g_{lm}(e) B_j^m(\theta),$$

where

$$B_i^l(\theta) := \frac{\partial}{\partial \rho^i}\Big|_{\rho=\theta} \left(\theta^{-1} \cdot \rho\right)^l .$$

In matrix form, it reads

$$g(\theta) = B(\theta) g(e) B^T(\theta),$$

where $B = (B_i^l)$ with i the row index and l the column index.
[A detailed proof is available in Sun et al., 2016. Examples available in Barndorff-Nielsen–Blæsild–Eriksen, 1989; Amari–Nagaoka, 2000; Sun et al., 2016.]

Theorem. (Amari–Nagaoka, 1993) Components of the Fisher information matrix g satisfy

$$g_{ij}(\theta) = B_i^l(\theta) g_{lm}(e) B_j^m(\theta),$$

where

$$B_i^l(\theta) := \frac{\partial}{\partial \rho^i}\Big|_{\rho=\theta} \left(\theta^{-1} \cdot \rho\right)^l .$$

In matrix form, it reads

$$g(\theta) = B(\theta) g(e) B^T(\theta),$$

where $B = (B_i^l)$ with i the row index and l the column index.
[A detailed proof is available in Sun et al., 2016. Examples available in Barndorff-Nielsen–Blæsild–Eriksen, 1989; Amari–Nagaoka, 2000; Sun et al., 2016.]

Corollary. Every 2-dimensional statistical transformation model has constant curvature.
[Some references on statistical manifolds of constant curvature: Cao–Sun–Wang, 2008; Rylov, 2016; Peng–Zhang, 2019.]

A modified natural gradient

If the transformation structure for a statistical model is known, then inversion of the Fisher information matrix becomes

$$g^{-1}(\theta) = B^{-T}(\theta) g^{-1}(e) B^{-1}(\theta)$$

and the natural gradient becomes

$$-\operatorname{grad}_N J(\theta) = -B^{-T}(\theta) g^{-1}(e) B^{-1}(\theta) \operatorname{grad} J(\theta).$$

Consequently, in the natural gradient descent method

$$\theta_{k+1} = \theta_k - h \operatorname{grad}_N J(\theta_k),$$

what left is to compute inversion of $g(e)$ and inversions of matrices $B(\theta_k)$ that are totally determined by the Lie group structure.

A modified natural gradient

If the transformation structure for a statistical model is known, then inversion of the Fisher information matrix becomes

$$g^{-1}(\theta) = B^{-T}(\theta) g^{-1}(e) B^{-1}(\theta)$$

and the natural gradient becomes

$$-\operatorname{grad}_N J(\theta) = -B^{-T}(\theta) g^{-1}(e) B^{-1}(\theta) \operatorname{grad} J(\theta).$$

Consequently, in the natural gradient descent method

$$\theta_{k+1} = \theta_k - h \operatorname{grad}_N J(\theta_k),$$

what left is to compute inversion of $g(e)$ and inversions of matrices $B(\theta_k)$ that are totally determined by the Lie group structure.

The Problem. Historically, people have mainly been focused on the existence of measures for a given Lie group action. In practice, it would be more important to determine the transformation structure for a given distribution.

Theorem. Assume $p(x;\theta)$ are pdfs for a statistical model $\mathcal{S} = \{p(x;\theta)\}$ with $x \in \Omega \subset \mathbb{R}^m$. The parameters θ are elements of an n-dimensional Lie group Θ, that are supposed to act on Ω, i.e., $\mathcal{T} : \Theta \times \Omega \to \Omega$. Then, $\mathcal{S}$ is a transformation model, namely, invariance of the probability measure, if and only if the **infinitesimal invariance criterion** is satisfied, namely.

$$\boxed{\mathbf{v}_i(p(x;\theta)) + p(x;\theta)\,\mathrm{Div}_x\,\xi_i \equiv 0}$$

holds for each infinitesimal generator

$$\mathbf{v}_i = \xi_i^j(x)\frac{\partial}{\partial x^j} + \eta_i^k(\theta)\frac{\partial}{\partial \theta^k}, \quad i = 1, 2, \ldots, n,$$

where ($\rho \in \Theta$, $j = 1, 2, \ldots, m$, $k = 1, 2, \ldots, n$)

$$\xi_i^j(x) = \frac{\partial}{\partial \rho^i}\Big|_{\rho=e}(\rho \circ x)^j, \quad \eta_i^k(\theta) = \frac{\partial}{\partial \rho^i}\Big|_{\rho=e}(\rho \cdot \theta)^k.$$

Theorem. Assume $p(x;\theta)$ are pdfs for a statistical model $\mathcal{S} = \{p(x;\theta)\}$ with $x \in \Omega \subset \mathbb{R}^m$. The parameters θ are elements of an n-dimensional Lie group Θ, that are supposed to act on Ω, i.e., $\mathcal{T} : \Theta \times \Omega \to \Omega$. Then, $\mathcal{S}$ is a transformation model, namely, invariance of the probability measure, if and only if the **infinitesimal invariance criterion** is satisfied, namely.

$$\boxed{\mathbf{v}_i(p(x;\theta)) + p(x;\theta)\,\mathrm{Div}_x\,\xi_i \equiv 0}$$

holds for each infinitesimal generator

$$\mathbf{v}_i = \xi_i^j(x)\frac{\partial}{\partial x^j} + \eta_i^k(\theta)\frac{\partial}{\partial \theta^k}, \quad i = 1, 2, \ldots, n,$$

where ($\rho \in \Theta$, $j = 1, 2, \ldots, m$, $k = 1, 2, \ldots, n$)

$$\xi_i^j(x) = \frac{\partial}{\partial \rho^i}\bigg|_{\rho=e}(\rho \circ x)^j, \quad \eta_i^k(\theta) = \frac{\partial}{\partial \rho^i}\bigg|_{\rho=e}(\rho \cdot \theta)^k.$$

LP [2020], Infinitesimal invariance criterion for statistical transformation models, draft.

Example. (Weibull distributions.)

$$p(x;\theta) = \frac{\beta}{\alpha}\left(\frac{x}{\alpha}\right)^{\beta-1}\exp\left\{-\left(\frac{x}{\alpha}\right)^{\beta}\right\}, \quad x \in \mathbb{R}^{+}, \theta = (\alpha,\beta) \in \mathbb{R}^{+} \times \mathbb{R}^{+}$$

Example. (Weibull distributions.)

$$p(x;\theta) = \frac{\beta}{\alpha}\left(\frac{x}{\alpha}\right)^{\beta-1}\exp\left\{-\left(\frac{x}{\alpha}\right)^{\beta}\right\}, \quad x \in \mathbb{R}^+, \theta = (\alpha,\beta) \in \mathbb{R}^+ \times \mathbb{R}^+$$

- Lie group structure (non-Abelian):

$$(\alpha_1,\beta_1)\cdot(\alpha_2,\beta_2) = \left(\alpha_1\alpha_2^{1/\beta_1}, \beta_1\beta_2\right)$$

  - Identity:

$$e = (1,1)$$

  - Inversion:

$$\rho^{-1} = \left(\frac{1}{\alpha^{\beta}}, \frac{1}{\beta}\right), \quad \rho = (\alpha,\beta)$$

Example. (Weibull distributions.)

$$p(x;\theta) = \frac{\beta}{\alpha}\left(\frac{x}{\alpha}\right)^{\beta-1}\exp\left\{-\left(\frac{x}{\alpha}\right)^{\beta}\right\}, \quad x \in \mathbb{R}^+, \theta = (\alpha,\beta) \in \mathbb{R}^+ \times \mathbb{R}^+$$

- Lie group structure (non-Abelian):

$$(\alpha_1,\beta_1)\cdot(\alpha_2,\beta_2) = \left(\alpha_1\alpha_2^{1/\beta_1}, \beta_1\beta_2\right)$$

  - Identity:

$$e = (1,1)$$

  - Inversion:

$$\rho^{-1} = \left(\frac{1}{\alpha^{\beta}}, \frac{1}{\beta}\right), \quad \rho = (\alpha,\beta)$$

- Group action $\rho \circ x$: *Do not know.*

How to use the IIC

- First of all, we can compute the η matrix from the group operation:

$$\eta_1^1 = \alpha, \quad \eta_2^1 = -\alpha \ln \alpha, \quad \eta_1^2 = 0, \quad \eta_2^2 = \beta$$

How to use the IIC

- First of all, we can compute the η matrix from the group operation:

$$\eta_1^1 = \alpha, \quad \eta_2^1 = -\alpha \ln \alpha, \quad \eta_1^2 = 0, \quad \eta_2^2 = \beta$$

- Solving the infinitesimal invariance criterion:

$$\xi_1 = x, \quad \xi_2 = -x \ln x,$$

namely

$$\mathbf{v}_1 = x\partial_x + \alpha\partial_\alpha, \quad \mathbf{v}_2 = -x \ln x \partial_x - \alpha \ln \alpha \partial_\alpha + \beta\partial_\beta.$$

How to use the IIC

- First of all, we can compute the η matrix from the group operation:

$$\eta_1^1 = \alpha, \quad \eta_2^1 = -\alpha \ln \alpha, \quad \eta_1^2 = 0, \quad \eta_2^2 = \beta$$

- Solving the infinitesimal invariance criterion:

$$\xi_1 = x, \quad \xi_2 = -x \ln x,$$

namely

$$\mathbf{v}_1 = x\partial_x + \alpha\partial_\alpha, \quad \mathbf{v}_2 = -x \ln x\partial_x - \alpha \ln \alpha\partial_\alpha + \beta\partial_\beta.$$

- The group action generated by $\mathbf{v}_1$ and $\mathbf{v}_2$ (using Lie series):

$$\rho \circ x \sim \exp\left(\left[\alpha x - \beta x \ln x\right]\partial_x\right)(x), \quad \rho = (\alpha, \beta)$$

How to use the IIC

- First of all, we can compute the η matrix from the group operation:

$$\eta_1^1 = \alpha, \quad \eta_2^1 = -\alpha \ln \alpha, \quad \eta_1^2 = 0, \quad \eta_2^2 = \beta$$

- Solving the infinitesimal invariance criterion:

$$\xi_1 = x, \quad \xi_2 = -x \ln x,$$

namely

$$\mathbf{v}_1 = x\partial_x + \alpha\partial_\alpha, \quad \mathbf{v}_2 = -x \ln x \partial_x - \alpha \ln \alpha \partial_\alpha + \beta\partial_\beta.$$

- The group action generated by $\mathbf{v}_1$ and $\mathbf{v}_2$ (using Lie series):

$$\rho \circ x \sim \exp\left(\left[\alpha x - \beta x \ln x\right]\partial_x\right)(x), \quad \rho = (\alpha, \beta)$$

Result: The model of Weibull distributions is a transformation model. It has constant curvature since its dimension is 2.

- Recall that the Fisher information metric is

$$g(\theta) = \begin{pmatrix} \frac{\beta^2}{\alpha^2} & \frac{\gamma-1}{\alpha} \\ \frac{\gamma-1}{\alpha} & \frac{(\gamma-1)^2}{\beta^2} + \frac{\pi^2}{6\beta^2} \end{pmatrix}, \quad g(e) = \begin{pmatrix} 1 & \gamma - 1 \\ \gamma - 1 & (\gamma - 1)^2 + \frac{\pi^2}{6} \end{pmatrix}$$

- Recall that the Fisher information metric is

$$g(\theta) = \begin{pmatrix} \frac{\beta^2}{\alpha^2} & \frac{\gamma-1}{\alpha} \\ \frac{\gamma-1}{\alpha} & \frac{(\gamma-1)^2}{\beta^2} + \frac{\pi^2}{6\beta^2} \end{pmatrix}, \quad g(e) = \begin{pmatrix} 1 & \gamma - 1 \\ \gamma - 1 & (\gamma - 1)^2 + \frac{\pi^2}{6} \end{pmatrix}$$

- The matrix $B(\theta)$ turns out to be diagonal

$$B(\theta) = \begin{pmatrix} \frac{\beta}{\alpha} & 0 \\ 0 & \frac{1}{\beta} \end{pmatrix}$$

such that $g(\theta) = B(\theta) g(e) B^T(\theta)$

- Recall that the Fisher information metric is

$$g(\theta) = \begin{pmatrix} \frac{\beta^2}{\alpha^2} & \frac{\gamma-1}{\alpha} \\ \frac{\gamma-1}{\alpha} & \frac{(\gamma-1)^2}{\beta^2} + \frac{\pi^2}{6\beta^2} \end{pmatrix}, \quad g(e) = \begin{pmatrix} 1 & \gamma - 1 \\ \gamma - 1 & (\gamma-1)^2 + \frac{\pi^2}{6} \end{pmatrix}$$

- The matrix $B(\theta)$ turns out to be diagonal

$$B(\theta) = \begin{pmatrix} \frac{\beta}{\alpha} & 0 \\ 0 & \frac{1}{\beta} \end{pmatrix}$$

  such that $g(\theta) = B(\theta)g(e)B^T(\theta)$

- Matrix inversion (e,g., in the natural gradient descent method) can be replaced by

$$g^{-1}(\theta) = B^{-T}(\theta)g^{-1}(e)B^{-1}(\theta)$$

Summary

- A brief introduction to information geometry, group actions and transformation models
- The main result: An infinitesimal invariance criterion for determining a transformation model

Summary

- A brief introduction to information geometry, group actions and transformation models
- The main result: An infinitesimal invariance criterion for determining a transformation model

- Future work
  - Other concrete examples
  - Applications to practical problems: To simplify the natural gradient descent method, in particular, simplify the computations of matrix inversion
  - etc.

Thanks very much for your attention.

Return!