Information Geometry of Estimating Functions in Parametric Statistical Models

Masayuki Henmi The Institute of Statistical Mathematics, Tokyo, Japan henmi@ism.ac.jp

【Introduction】

In information geometry, a parametric statistical model (a family of probability density functions) is treated as a differentiable manifold, where the Riemannian metric called Fisher metric and the pair of two torsion-free dual affine connections called the exponential and mixture connections play essential roles for statistical inference. For example, the maximum likelihood estimation in an exponential family is understood as the orthogonal projection of the geodesic defined by the mixture connection. This comes from the fact that an exponential family is a dually flat space, where the curvature and the torsion tensors of the two dual affine connections are all equal to zero. Recently, it has been found by the authors that a general estimating function naturally induces a similar geometric structure on a statistical model, that is, a Riemannian metric and a pair of dual affine connections. However, one of the affine connections is not necessarily torsion-free, especially when the estimating function is not integrable with respect to the parameter of the statistical model. In this presentation, we explain this geometry with some basic knowledge of information geometry.

【Statistical manifold】

In this presentation,we assume that all geometrical objects on differentiable manifolds are smooth (C^∞).

Definition 1 (Statistical manifold)

For a Riemannian manifold (M, g) and an affine connection ∇ on M, we call (M, g, ∇) a **statistical manifold** if and only if both of ∇ and its dual connection ∇^* with respect to g are torsion-free.

Remark

1. The *dual connection* ∇^* of ∇ with respect to g is defined by

$$Xg(Y,Z) = g(\nabla_X Y, Z) + g(Y, \nabla^*_X Z) \quad (\forall X, \forall Y, \forall Z \in \mathcal{X}(M)),$$

where $\mathcal{X}(M)$ is the set of all vector fields on M.

2. For an affine connection ∇ on M, its curvature tensor field R and torsion tensor field T are defined by

$$R(X,Y)Z := \nabla_X\nabla_Y Z - \nabla_Y\nabla_X Z - \nabla_{[X,Y]}Z, \quad T(X,Y) := \nabla_X Y - \nabla_Y X - [X,Y]$$
$$(\forall X, \forall Y, \forall Z \in \mathcal{X}(M)).$$

An affine connection ∇ is said to be *torsion-free* if $T = 0$. For a torsion-free affine connection ∇, $\nabla^* = \nabla$ implies that ∇ is the Levi-Civita connection with respect to g. when we let R^* and T^* be the curvature and torsion tensor fields of ∇^*, respectively. $R = 0$ always implies $R^* = 0$, but $T = 0$ does not necessarily imply $T^* = 0$. For a statistical manifold (M, g, ∇), $R = 0$ implies that ∇ and ∇^* are both flat (*i.e.* $T = 0, R = 0$ and $T^* = 0, R^* = 0$). In this case, (M, g, ∇, ∇^*) is called a *dually flat space*.

【Contrast function】

For a real-valued function ϕ on the direct product $M \times M$ of a manifold M and vector fields $X_1, ..., X_i, Y_1, ..., Y_j$ on M, the functions $\phi[X_1, ..., X_i|Y_1, ..., Y_j]$, $\phi[X_1, ..., X_i|\ ]$ and $\phi[\ |Y_1, ..., Y_j]$ on M are defined by

$$\begin{aligned}\phi[X_1, \ldots, X_i|Y_1, \ldots, Y_j](r) &:= (X_1)_p \cdots (X_i)_p (Y_1)_q \cdots (Y_j)_q \phi(p,q)|_{p=r,q=r},\\ \phi[X_1, \ldots, X_i|\ ](r) &:= (X_1)_p \cdots (X_i)_p \phi(p,r)|_{p=r},\\ \phi[\ |Y_1, \ldots, Y_j](r) &:= (Y_1)_q \cdots (Y_j)_q \phi(r,q)|_{q=r}\end{aligned}$$

for any $r \in M$, respectively.

Definition 2 (Contrast function)

A real-valued function ϕ on $M \times M$ is called a **contrast function** on M if and only if it satisfies

(a) $\phi(p,p) = 0 \quad (\forall p \in M)$,
(b) $\phi[X|\ ] = \phi[\ |X] = 0 \quad (\forall X \in \mathcal{X}(M))$,
(c) $g(X,Y) := -\phi[X|Y] \quad (\forall X, \forall Y \in \mathcal{X}(M))$ is a Riemannian metric on M.

These conditions imply that

$$\phi(p,q) \geq 0, \quad \phi(p,q) = 0 \Longleftrightarrow p = q$$

in some neighborhood of the diagonal set $\{(r,r)|r \in M\}$ in $M \times M$. Although a contrast function is not necessarily symmetric, this property means that a contrast function measures some discrepancy between two points on M (at least locally). For a given contrast function ϕ, the two affine connections ∇ and ∇^* are defined by

$$g(\nabla_X Y, Z) = -\phi[XY|Z], \ g(Y, \nabla^*_X Z) = -\phi[Y|XZ] \quad (\forall X, \forall Y, \forall Z \in \mathcal{X}(M)).$$

In this case, ∇ and ∇^* are both torsion-free and dual to each other with respect to g. This means that both of (M, g, ∇) and (M, g, ∇^*) are statistical manifolds. In particular, (M, g, ∇) is called the statistical manifold induced from the contrast function ϕ.

【Geometry induced from Kullback-Leibler divergence】

Let $S = \{p(\boldsymbol{x};\boldsymbol{\theta}) \mid \boldsymbol{\theta} = (\theta^1, ..., \theta^d) \in \Theta \subset \boldsymbol{R}^d\}$ be a regular parametric statistical model, that is, a set of probability density functions with respect to a dominating measure ν on a sample space Ω, each element of which is indexed by a parameter vector $\boldsymbol{\theta}$ in an open subset Θ of $\boldsymbol{R}^d$, and which satisfies some regularity conditions.

The **Kullback-Leibler divergence** of the two density functions $p_1(\boldsymbol{x}) = p(\boldsymbol{x};\boldsymbol{\theta}_1)$ and $p_2(\boldsymbol{x}) = p(\boldsymbol{x};\boldsymbol{\theta}_2)$ in S,

$$\phi_{KL}(p_1, p_2) := \int_\Omega p_2(\boldsymbol{x}) \log \frac{p_2(\boldsymbol{x})}{p_1(\boldsymbol{x})} \nu(d\boldsymbol{x})$$

is a contrast function on S. Its induced Riemannian metric and dual affine connectins are called **Fisher metric** g^F, the **exponential connection** $\nabla^{(e)}$ and **mixture connection** $\nabla^{(m)}$, respectively, which are given as follows:

$$\begin{aligned} g^F_{jk}(\boldsymbol{\theta}) &:= g^F(\partial_j, \partial_k) = E_{\boldsymbol{\theta}}\{s^j(\boldsymbol{x},\boldsymbol{\theta}) s^k(\boldsymbol{x},\boldsymbol{\theta})\},\\ &\begin{cases} \Gamma^{(e)}_{ij,k}(\boldsymbol{\theta}) := g^F(\nabla^{(e)}_{\partial_i}\partial_j, \partial_k) = E_{\boldsymbol{\theta}}[\{\partial_i s^j(\boldsymbol{x},\boldsymbol{\theta})\} s^k(\boldsymbol{x},\boldsymbol{\theta})] \\ \Gamma^{(m)}_{ik,j}(\boldsymbol{\theta}) := g^F(\partial_j, \nabla^{(m)}_{\partial_i}\partial_k) = \int_\Omega s^j(\boldsymbol{x},\boldsymbol{\theta}) \partial_i\partial_k p(\boldsymbol{x};\boldsymbol{\theta}) \nu(d\boldsymbol{x}) \end{cases}, \end{aligned}$$

where $E_{\boldsymbol{\theta}}$ indicates that the expectation is taken with respect to $p(\boldsymbol{x};\boldsymbol{\theta})$, $\partial_i = \frac{\partial}{\partial\theta^i}$ and $s^i(\boldsymbol{x};\boldsymbol{\theta}) = \partial_i \log p(\boldsymbol{x};\boldsymbol{\theta})$ $(i = 1, \ldots, d)$. In particular, if S is an exponential family, $(S, g^F, \nabla^{(e)}, \nabla^{(m)})$ is a dually flat space. This geometrical structure plays the most fundamental and important role in the differential geometry of statistical inference. For example, the maximum likelihood estimator in an exponential family is obtained by the orthogonal projection of the geodesic with respect to the mixture connection $\nabla^{(m)}$.

【Statistical manifold admitting torsion】

Definition 3 (Statistical manifold admitting torsion)

For a Riemannian manifold (M, g) and an affine connection ∇ on M, we call (M, g, ∇) a **statistical manifold admitting torsion** (or **SMAT** for short) if and only if the dual connection ∇^* of ∇ with respect to g is torsion-free.

For a SMAT (M, g, ∇), the affine connection ∇ possibly has torsion (*i.e.* $T \neq 0$). Hence, $R = 0$ does not necessarily imply that ∇ is flat, but it implies that ∇^* is flat since $R^* = 0$ and $T^* = 0$. In this case, we call (M, g, ∇, ∇^*) a *partially flat space*.

【Pre-contrast function】

For a real-valued function ρ on the direct product $TM \times M$ of a manifold M and its tangent bundle TM, and vector fields$X_1, ..., X_i, Y_1, ..., Y_j, Z$ on M, the function $\rho[X_1, ..., X_i Z|Y_1, ..., Y_j]$ on M is defined by

$$\rho[X_1, \ldots, X_i Z|Y_1, \ldots, Y_j](r) := (X_1)_p \cdots (X_i)_p (Y_1)_q \cdots (Y_j)_q \rho(Z_p, q)|_{p=r,q=r} \quad (\forall r \in M)$$

The functions $\rho[X_1, ..., X_i Z|\ ]$ and $\rho[\ |Y_1, ..., Y_j]$ are also defined in the same way as above.

Definition 4 (Pre-contrast function)

A real-valued function ρ on $TM \times M$ is called a **pre-contrast function** on M if and only if it satisfies

(a) $\rho(f_1X_1 + f_2X_2, q) = f_1\rho(X_1, q) + f_2\rho(X_2, q)$
$(\forall f_1, \forall f_2 \in C^\infty(M), \ \forall X_1, \forall X_2 \in \mathcal{X}(M), \ \forall q \in M)$.
(b) $\rho[X|\ ] = 0 \quad (\forall X \in \mathcal{X}(M)) \quad \big(i.e.\ \rho(X_p, p) = 0 \quad (\forall p \in M)\big)$.
(c) $g(X,Y) := -\rho[X|Y] \quad (\forall X, \forall Y \in \mathcal{X}(M))$ is a Riemannian metric on M.

For a given pre-contrast function ρ, two affine connections ∇ and ∇^* are defined by

$$g(\nabla_X Y, Z) = -\rho[XY|Z], \quad g(Y, \nabla^*_X Z) = -\rho[Y|XZ] \quad (\forall X, \forall Y, \forall Z \in \mathcal{X}(M)).$$

In this case, ∇ and ∇^* are dual to each other with respect to g and ∇^* is torsion-free. However, the affine connection ∇ possibly has torsion. This means that (M, g, ∇) is a SMAT and it is called the SMAT induced from the pre-contrast function ρ.

For any contrast function ϕ on M, the function ρ_ϕ, which is defined by

$$\rho_\phi(X_p, q) := X_p\phi(p, q) \quad (\forall p, \forall q \in M, \ \forall X_p \in T_p(M)),$$

is a pre-contrast function on M. The notion of pre-contrast function is obtained by taking the fundamental properties of the first derivative of a contrast function as axioms.

【Geometry induced from estimating functions】

Let $S = \{p(\boldsymbol{x};\boldsymbol{\theta}) \mid \boldsymbol{\theta} = (\theta^1, ..., \theta^d) \in \Theta \subset \boldsymbol{R}^d\}$ be a regular parametric statistical model. An estimating function on S, which we consider here, is a $\boldsymbol{R}^d$-valued function $\boldsymbol{u}(\boldsymbol{x},\boldsymbol{\theta})$ satisfying

$$E_{\boldsymbol{\theta}}\{\boldsymbol{u}(\boldsymbol{x},\boldsymbol{\theta})\} = \boldsymbol{0}, \quad E_{\boldsymbol{\theta}}\{\|\boldsymbol{u}(\boldsymbol{x},\boldsymbol{\theta})\|^2\} < \infty, \quad \det\left[E_{\boldsymbol{\theta}}\left\{\frac{\partial \boldsymbol{u}}{\partial \boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{\theta})\right\}\right] \neq 0 \quad (\forall \boldsymbol{\theta} \in \Theta).$$

For an estimating function $\boldsymbol{u}(\boldsymbol{X},\boldsymbol{\theta})$ and a random sample $X_1, \ldots, X_n$ from an unknown probability distribution $p(\boldsymbol{x};\boldsymbol{\theta}_0)$ in S, an M-estimator $\hat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}_0$ is obtained as a solution to the estimating equation

$$\sum_{i=1}^n \boldsymbol{u}(\boldsymbol{X}_i, \boldsymbol{\theta}) = \boldsymbol{0}.$$

The M-estimator $\hat{\boldsymbol{\theta}}$ has the consistency and asymptotic normality

$$\hat{\boldsymbol{\theta}} \longrightarrow \boldsymbol{\theta}_0 \text{ (in probability)}, \quad \sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \longrightarrow N\left(\boldsymbol{0}, \mathrm{Avar}\left(\hat{\boldsymbol{\theta}}\right)\right) \text{ (in distribution)}$$

as $n \to \infty$ under some additional regularity conditions.
The matrix $\mathrm{Avar}(\hat{\boldsymbol{\theta}})$ is the asymptotic variance-covariance matrix of $\hat{\boldsymbol{\theta}}$ and is given by

$$\mathrm{Avar}(\hat{\boldsymbol{\theta}}) = \{A(\boldsymbol{\theta}_0)\}^{-1} B(\boldsymbol{\theta}_0) \{A(\boldsymbol{\theta}_0)^{-1}\}^T = \{G(\boldsymbol{\theta}_0)\}^{-1},$$

where $A(\boldsymbol{\theta}) := E_{\boldsymbol{\theta}}\{(\partial \boldsymbol{u}/\partial \boldsymbol{\theta})(\boldsymbol{x},\boldsymbol{\theta})\}$, $B(\boldsymbol{\theta}) := E_{\boldsymbol{\theta}}\left\{\boldsymbol{u}(\boldsymbol{x},\boldsymbol{\theta})\boldsymbol{u}(\boldsymbol{x},\boldsymbol{\theta})^T\right\}$ and $G(\boldsymbol{\theta})$ is the asymptotic variance-covariance matrix of the *standardized estimating function* $\boldsymbol{u}_*(\boldsymbol{x},\boldsymbol{\theta})$ of $\boldsymbol{u}(\boldsymbol{X},\boldsymbol{\theta})$, which is defined by

$$\boldsymbol{u}_*(\boldsymbol{x},\boldsymbol{\theta}) := E_{\boldsymbol{\theta}}\left\{\boldsymbol{s}(\boldsymbol{x},\boldsymbol{\theta})\boldsymbol{u}(\boldsymbol{x},\boldsymbol{\theta})^T\right\}\left[E_{\boldsymbol{\theta}}\left\{\boldsymbol{u}(\boldsymbol{x},\boldsymbol{\theta})\boldsymbol{u}(\boldsymbol{x},\boldsymbol{\theta})^T\right\}\right]^{-1}\boldsymbol{u}(\boldsymbol{x},\boldsymbol{\theta}),$$

where $\boldsymbol{s}(\boldsymbol{x},\boldsymbol{\theta}) := (\partial/\partial\boldsymbol{\theta}) \log p(\boldsymbol{x};\boldsymbol{\theta})$ is the score function for $\boldsymbol{\theta}$. Geometrically, the i-th component of the standardized estimating function $\boldsymbol{u}_*(\boldsymbol{x},\boldsymbol{\theta})$ is the orthogonal projection of the i-th component of the score function $\boldsymbol{s}(\boldsymbol{x},\boldsymbol{\theta})$ onto the linear space spanned by all components of the estimating function $\boldsymbol{u}(\boldsymbol{x},\boldsymbol{\theta})$ in the Hilbert space

$$\mathcal{H}_{\boldsymbol{\theta}} := \{a(\boldsymbol{x}) \mid E_{\boldsymbol{\theta}}\{a(\boldsymbol{x})\} = 0, \ E_{\boldsymbol{\theta}}\{a(\boldsymbol{x})^2\} < \infty\}$$

with the inner product $< a(\boldsymbol{x}), b(\boldsymbol{x}) >_{\boldsymbol{\theta}} := E_{\boldsymbol{\theta}}\{a(\boldsymbol{x})b(\boldsymbol{x})\}$ $(\forall a(\boldsymbol{x}), \forall b(\boldsymbol{x}) \in \mathcal{H}_{\boldsymbol{\theta}})$. The matrix $G(\boldsymbol{\theta})$ is called a *Godambe information matrix*, which can be seen as a generalization of the Fisher information matrix.

Since the Kullback-Leibler divergence ϕ_{KL} is a contrast function on S, we obtain a pre-contrast function ρ_{KL} on S from the first derivative of ϕ_{KL}:

$$\rho_{KL}((\partial_j)_{p_1}, p_2) := (\partial_j)_{p_1}\phi_{KL}(p_1, p_2) = -\int_\Omega s^j(\boldsymbol{x},\boldsymbol{\theta}_1) p(\boldsymbol{x};\boldsymbol{\theta}_2) \nu(d\boldsymbol{x}) \quad (j = 1, \ldots, d)$$

for any two probability distributions $p_1(\boldsymbol{x}) = p(\boldsymbol{x};\boldsymbol{\theta}_1)$ and $p_2(\boldsymbol{x}) = p(\boldsymbol{x};\boldsymbol{\theta}_2)$ in S. This observation leads to the following proposition.

Proposition (Pre-contrast function associated with an estimating function)

For an estimating function $\boldsymbol{u}(\boldsymbol{x},\boldsymbol{\theta})$ on the parametric model S, a pre-contrast function $\rho_{\boldsymbol{u}} : TS \times S \to \boldsymbol{R}$ is defined by

$$\rho_{\boldsymbol{u}}((\partial_j)_{p_1}, p_2) := -\int_\Omega u^j_*(\boldsymbol{x},\boldsymbol{\theta}_1) p(\boldsymbol{x};\boldsymbol{\theta}_2) \nu(d\boldsymbol{x}) \quad (j = 1, \ldots, d)$$

for any two probability distributions $p_1(\boldsymbol{x}) = p(\boldsymbol{x};\boldsymbol{\theta}_1)$ and $p_2(\boldsymbol{x}) = p(\boldsymbol{x};\boldsymbol{\theta}_2)$ in S, where $u^j_*(\boldsymbol{x},\boldsymbol{\theta})$ is the j-th component of the standardized estimating function $\boldsymbol{u}_*(\boldsymbol{x},\boldsymbol{\theta})$ of $\boldsymbol{u}(\boldsymbol{x},\boldsymbol{\theta})$.

The use of $\boldsymbol{u}_*(\boldsymbol{x},\boldsymbol{\theta})$ instead of $\boldsymbol{u}(\boldsymbol{x},\boldsymbol{\theta})$ ensures that the definition of the function $\rho_{\boldsymbol{u}}$ does not depend on the choice of coordinate system (parameter) of S.

The proof of Proposition is straightforward. In particular, the condition (b) in the definition of pre-contrast function follows from the unbiasedness of the (standardized) estimating function. The Riemannian metric g, dual connections ∇ and ∇^* induced from the pre-contrast function $\rho_{\boldsymbol{u}}$ are given as follows:

$$\begin{aligned} g_{jk}(\boldsymbol{\theta}) &:= g(\partial_j, \partial_k) = E_{\boldsymbol{\theta}}\{u^j_*(\boldsymbol{x},\boldsymbol{\theta}) u^k_*(\boldsymbol{x},\boldsymbol{\theta})\} = G_{jk}(\boldsymbol{\theta}),\\ &\begin{cases} \Gamma_{ij,k}(\boldsymbol{\theta}) := g(\nabla_{\partial_i}\partial_j, \partial_k) = E_{\boldsymbol{\theta}}[\{\partial_i u^j_*(\boldsymbol{x},\boldsymbol{\theta})\} s^k(\boldsymbol{x},\boldsymbol{\theta})] \\ \Gamma^*_{ik,j}(\boldsymbol{\theta}) := g(\partial_j, \nabla^*_{\partial_i}\partial_k) = \int_\Omega u^j_*(\boldsymbol{x},\boldsymbol{\theta}) \partial_i\partial_k p(\boldsymbol{x};\boldsymbol{\theta}) \nu(d\boldsymbol{x}) \end{cases}, \end{aligned}$$

where $G_{jk}(\boldsymbol{\theta})$ is the (j,k) component of the Godambe information matrix $G(\boldsymbol{\theta})$. Note that ∇^* is always torsion-free since $\Gamma^*_{ik,j} = \Gamma^*_{ki,j}$, whereas ∇ is not necessarily torsion-free unless $\boldsymbol{u}_*(\boldsymbol{x},\boldsymbol{\theta})$ is integrable with respect to $\boldsymbol{\theta}$ (*i.e.* there exists a function $\psi(\boldsymbol{x},\boldsymbol{\theta})$ satisfying $\partial_j\psi(\boldsymbol{x},\boldsymbol{\theta}) = u^j_*(\boldsymbol{x},\boldsymbol{\theta})$ $(j = 1, \ldots, d)$).

【Reference】

Henmi, M. and Matsuzoe, H. (2018). Statistical Manifolds Admitting Torsion and Partially Flat Spaces. *Geometric Structures of Information* Springer, 37-50.