Information Geometry of Relative α-Entropy

M. Ashok Kumar[1] and Kumar Vijay Mishra[2]

[1]Indian Institute of Technology Palakkad, Kerala, India
[2]The University of Iowa, Iowa City, IA 52246 USA

June 18, 2018

OUTLINE

- Information Geometry - study of statistical models from a Riemannian geometric perspective

OUTLINE

- Information Geometry - study of statistical models from a Riemannian geometric perspective
- Riemannian Metric from Divergence Functions

OUTLINE

- Information Geometry - study of statistical models from a Riemannian geometric perspective
- Riemannian Metric from Divergence Functions
- Fisher Information Metric from Kullback-Leibler Divergence

OUTLINE

- Information Geometry - study of statistical models from a Riemannian geometric perspective
- Riemannian Metric from Divergence Functions
- Fisher Information Metric from Kullback-Leibler Divergence
- Estimation Error Bounds from Information Geometry

INTRODUCTION

- **Manifolds** - statistical models

INTRODUCTION

- **Manifolds** - statistical models
  - Every point on the manifold is a probability distribution from the model.

INTRODUCTION

- **Manifolds** - statistical models
  - Every point on the manifold is a probability distribution from the model.
  - In general, a manifold $S = \{p_\theta : \theta = (\theta_1, \ldots, \theta_k) \in \Theta\}$, where Θ is an open subset of $\mathbb{R}^k$, called parameter space.

INTRODUCTION

- **Manifolds** - statistical models
  - Every point on the manifold is a probability distribution from the model.
  - In general, a manifold $S = \{p_\theta : \theta = (\theta_1, \ldots, \theta_k) \in \Theta\}$, where Θ is an open subset of $\mathbb{R}^k$, called parameter space.
  - Dimension of a manifold is the “dimension" of the parameter space

INTRODUCTION

- **Manifolds** - statistical models
  - Every point on the manifold is a probability distribution from the model.
  - In general, a manifold $S = \{p_\theta : \theta = (\theta_1, \ldots, \theta_k) \in \Theta\}$, where Θ is an open subset of $\mathbb{R}^k$, called parameter space.
  - Dimension of a manifold is the “dimension" of the parameter space
  - **Example**: $\{\mathcal{N}(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma \in (0, \infty)\}$ is a 2-dimensional manifold.

INTRODUCTION

- **Manifolds** - statistical models
  - Every point on the manifold is a probability distribution from the model.
  - In general, a manifold $S = \{p_\theta : \theta = (\theta_1, \ldots, \theta_k) \in \Theta\}$, where Θ is an open subset of $\mathbb{R}^k$, called parameter space.
  - Dimension of a manifold is the “dimension" of the parameter space
  - **Example**: $\{\mathcal{N}(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma \in (0, \infty)\}$ is a 2-dimensional manifold.
- **Riemannian metric**: An inner product defined at every point on the manifold.

DIVERGENCES

Let $\mathcal{P}$ denote the space of all probability measures defined on $\mathcal{X}$ and $\mathcal{P}$ denote the space of all (positive) measures on $\mathcal{X}$.

DIVERGENCES

Let $\mathcal{P}$ denote the space of all probability measures defined on $\mathcal{X}$ and $\mathcal{P}$ denote the space of all (positive) measures on $\mathcal{X}$.

- **Divergence** is a non-negative function that measures the "distance" between two measures u and v satisfying

$$D(u,v) = 0 \text{ if and only if } u = v.$$

DIVERGENCES

Let $\mathcal{P}$ denote the space of all probability measures defined on $\mathcal{X}$ and $\mathcal{P}$ denote the space of all (positive) measures on $\mathcal{X}$.

- **Divergence** is a non-negative function that measures the "distance" between two measures u and v satisfying

$$D(u,v) = 0 \text{ if and only if } u = v.$$

- **Example**: The Kullback-Leibler divergence
  - For probability measures p and q,

$$I(p,q) = \sum_x p(x) \log\left(\frac{p(x)}{q(x)}\right).$$

DIVERGENCES

Let $\mathcal{P}$ denote the space of all probability measures defined on $\mathcal{X}$ and $\mathcal{P}$ denote the space of all (positive) measures on $\mathcal{X}$.

- **Divergence** is a non-negative function that measures the "distance" between two measures u and v satisfying

$$D(u,v) = 0 \text{ if and only if } u = v.$$

- **Example**: The Kullback-Leibler divergence
  - For probability measures p and q,

$$I(p,q) = \sum_x p(x) \log\left(\frac{p(x)}{q(x)}\right).$$

  - For measures u and v,

$$I(u,v) = \sum_x u(x) \log\left(\frac{u(x)}{v(x)}\right) - \sum_x u(x) + \sum_x v(x).$$

EGUCHI'S THEORY

- Eguchi (1992) defines a Riemannian metric on S by the matrix

$$G^{(D)}(\theta) = \left[g_{i,j}^{(D)}(\theta) \right],$$

where

$$g_{i,j}^{(D)}(\theta) := -\frac{\partial}{\partial \theta_i} \frac{\partial}{\partial \theta'_j} D(p_\theta, p_{\theta'}) \Big|_{\theta=\theta'}$$

and $\theta = (\theta_1, \ldots, \theta_k)$, $\theta' = (\theta'_1, \ldots, \theta'_k)$.

EGUCHI'S THEORY

- Eguchi (1992) defines a Riemannian metric on S by the matrix

$$G^{(D)}(\theta) = \left[g_{i,j}^{(D)}(\theta) \right],$$

where

$$g_{i,j}^{(D)}(\theta) := -\frac{\partial}{\partial \theta_i} \frac{\partial}{\partial \theta'_j} D(p_\theta, p_{\theta'}) \Big|_{\theta=\theta'}$$

and $\theta = (\theta_1, \ldots, \theta_k)$, $\theta' = (\theta'_1, \ldots, \theta'_k)$.

- This defines an inner product at every point p_θ of S.

Fisher Information Matrix

- Let $S = \{p_\theta : \theta \in \Theta\}$ be a statistical manifold.
- Applying Eguchi's theory to KL-divergence, we get a Riemannian metric $G^{(I)}(\theta) = [g^{(I)}_{i,j}(\theta)]$ on S, where

$$\begin{aligned} g^{(I)}_{i,j}(\theta) &= E_\theta[\partial_i \log p_\theta(X) \cdot \partial_j \log p_\theta(X)] \\ &= \mathrm{Cov}_\theta[\partial_i \log p_\theta(X), \partial_j \log p_\theta(X)], \end{aligned}$$

where the last equality follows from the fact that the expectation of the score function is zero, that is, $E_\theta[\partial_i \log p_\theta(X)] = 0$.

Information Geometry Results

Theorem [Amari, Nagaoka 2000]

Let f be a real valued differentiable function on S. The *differential* of f at p, $(\mathrm{d}f)_p$, satisfies

$$\|(\mathrm{d}f)_p\|_p^2 = \sum_{i,j} (g^{i,j})^{(e)} \partial_j(f) \partial_i(f),$$

where $[G^{(e)}]^{-1} = (g^{i,j})^{(e)}$ is the inverse of $G^{(e)}$.

Information Geometry Results

Theorem [Amari, Nagaoka 2000]

Let $A : \mathbb{X} \to \mathbb{R}$ be any mapping (that is, a vector in $\mathbb{R}^{\mathbb{X}}$.) Let $E[A] : \mathcal{P} \to \mathbb{R}$ be the mapping $p \mapsto E_p[A]$. We then have

$$\mathrm{Var}_p(A) = \|(\mathrm{d}E[A])\|_p^2.$$

Corollary

If S is a submanifold of $\mathcal{P}$, then

$$\mathrm{Var}_p[A] \geq \|(\mathrm{d}E[A]|_S)\|_p^2.$$

CRAMER-RAO BOUND

- Let $S = \{p_\theta : \theta \in \Theta\}$ be a statistical model and let $\widehat{\theta} := (\widehat{\theta}_1, \ldots, \widehat{\theta}_n)$ be an unbiased estimator of θ.
- Taking $A = \sum_{i=1}^{k} c_i \widehat{\theta}_i$ in the previous corollary, we have

$$\sum_{i,j} c_i c_j \mathrm{Cov}_\theta(\widehat{\theta}_i, \widehat{\theta}_j) \geq \sum_{i,j} c_i c_j (g^{(I)}(\theta))^{i,j}(\theta).$$

where $\mathrm{Var}_\theta(\widehat{\theta}) = [\mathrm{Cov}_\theta(\widehat{\theta}_i(X), \widehat{\theta}_j(X))]$ is the covariance matrix.

THEOREM

If $\widehat{\theta}$ is an unbiased estimator of θ, then

$$\mathrm{Var}_\theta[\widehat{\theta}] \geq [G^{(I)}(\theta)]^{-1}.$$

RELATIVE α-ENTROPY

Relative α-Entropy (also known as *Sundaresan's divergence or I_α-divergence*) between two probability distributions p and q is defined as

$$
\begin{aligned}
& I_\alpha(p,q) \\
& := \frac{\alpha}{1-\alpha} \log \sum_x p(x) \left(\frac{q(x)}{\|q\|_\alpha} \right)^{\alpha-1} - \frac{1}{1-\alpha} \log \sum_x p(x)^\alpha \\
& = \underbrace{\frac{\alpha}{1-\alpha} \log \sum_x p(x) \left(\frac{q(x)}{\|q\|_\alpha} \right)^{\alpha-1}}_{\text{cross-entropy term}} - \underbrace{H_\alpha(p)}_{\text{entropy term}}
\end{aligned}
$$

An α-Fisher Information Metric

$G^{(\alpha)}(\theta) = [g_{i,j}^{(\alpha)}(\theta)]$, where

$$
\begin{aligned}
g_{i,j}^{(\alpha)}(\theta) &:= g_{i,j}^{(I_\alpha)}(\theta) \\
&= -\frac{\partial}{\partial \theta'_j}\frac{\partial}{\partial \theta_i} I_\alpha(p_\theta, p_{\theta'})\Big|_{\theta'=\theta} \\
&= \frac{\alpha}{\alpha-1}\sum_x \partial_i p_\theta(x)\cdot \partial'_j\left(\frac{p_{\theta'}(x)^{\alpha-1}}{\sum_y p_\theta(y) p_{\theta'}(y)^{\alpha-1}}\right)\Big|_{\theta'=\theta} \\
&= \frac{1}{\alpha}\mathrm{Cov}_{\theta(\alpha)}[\partial_i \log p_\theta^{(\alpha)}(X), \partial_j \log p_\theta^{(\alpha)}(X)],
\end{aligned}
$$

where $p_\theta^{(\alpha)}$ is the α-escort distribution associated with p_θ,

$$
p_\theta^{(\alpha)}(x) := \frac{p_\theta(x)^\alpha}{\sum_y p_\theta(y)^\alpha}.
$$

An α-version of Cramer-Rao Bound

$$\begin{aligned}\partial_i^{(\alpha)}(p_\theta) &:= \frac{1}{\alpha-1}\partial_i'\left(\frac{{p_{\theta'}}^{\alpha-1}}{\sum_y p_\theta(y)p_{\theta'}(y)^{\alpha-1}}\right)\Bigg|_{\theta'=\theta} \\ &= \left[\frac{{p_\theta}^{(\alpha)}}{p_\theta}\partial_i(\log p_\theta) - \frac{{p_\theta}^{(\alpha)}}{p_\theta}E_{\theta^{(\alpha)}}[\partial_i(\log p_\theta)]\right].\end{aligned}$$

With this, the α-information metric can be written as

$$g_{i,j}^{(\alpha)}(\theta) = \alpha\sum_x \partial_i p_\theta(x)\cdot\partial_i^{(\alpha)}(p_\theta(x)).$$

Observe that $E_\theta[\partial_i^{(\alpha)}(p_\theta)] = 0$ as $\partial_i^{(\alpha)}p_\theta = \frac{p_\theta^{(\alpha)}}{p_\theta}\partial_i \log p_\theta^{(\alpha)}$.

When $\alpha = 1$, the r.h.s of the above reduces to $\partial_i(\log p_\theta)$.

AN α-VERSION OF CRAMER-RAO BOUND CONTD..

The α-*representation of a tangent vector* X *at* p can be defined as

$$\begin{aligned} X_p^{(\alpha)}(x) &:= \left[\frac{p^{(\alpha)}(x)}{p(x)} X_p^{(e)}(x) - \frac{p^{(\alpha)}(x)}{p(x)} E_{p^{(\alpha)}}[X_p^{(e)}]\right] \\ &= \left[\frac{p^{(\alpha)}(x)}{p(x)} \left(X_p^{(e)}(x) - E_{p^{(\alpha)}}[X_p^{(e)}]\right)\right]. \end{aligned}$$

The collection of all such α-representations is

$$T_p^{(\alpha)}(\mathcal{P}) := \{X_p^{(\alpha)} : X \in T_p(\mathcal{P})\}.$$

Clearly $E_p[X_p^{(\alpha)}] = 0$. Also, since any $A \in \mathbb{R}^{\mathbb{X}}$ with $E_p[A] = 0$ can be written as

$$A = \left[\frac{p^{(\alpha)}}{p} \left(B - E_{p^{(\alpha)}}[B]\right)\right]$$

REFERENCES

- Rao, C. R. (1945). “Information and the accuracy attainable in the estimation of statistical parameters", *Bulletin of Calcutta Mathematical Society*, **37**, pp. 81–91.
- Amari, S. and Nagaoka, H. (2000). “Methods of information geometry", *American Mathematical Society, Oxford University Press*, vol. 191.
- Eguchi, S. (1992). “Geometry of minimum contrast", *Hiroshima Mathematical Journal*, **22(3)**, pp. 631–647.
- Groves, T. and Rothenberg, T. (1969). “A Note on the Expected Value of an Inverse Matrix", *Biometrika*, **56**, pp. 690–691.