The α-representations of the Fisher Information Matrix
— On gauge freedom of the FIM —

Frank Nielsen
Frank.Nielsen@acm.org

19 September 2017
Revised September 2020

The *Fisher Information Matrix* [1] (FIM) for a family of *parametric* probability models $\{p(x;\theta)\}_{\theta\in\Theta}$ (densities $p(x;\theta)$ expressed with respect to a positive base measure ν) indexed by a D-dimensional parameter vector $\theta := (\theta^1, \ldots, \theta^D)$ is historically defined by

$$I(\theta) := [I_{ij}(\theta)], \quad I_{ij}(\theta) := E_{p(x;\theta)}\left[\partial_i l(x;\theta)\partial_j l(x;\theta)\right], \tag{1}$$

where $l(x;\theta) := \log p(x;\theta)$ is the *log-likelihood function*, and $\partial_i :=: \frac{\partial}{\partial\theta^i}$ (by notational convention). The FIM is a $D \times D$ positive semi-definite matrix for a D-order parametric family.

The FIM is a cornerstone in statistics and occurs in many places, like for example the celebrated *Cramér-Rao lower bound* [3] for an unbiased estimator $\hat{\theta}$:

$$\mathrm{Var}_{p(x;\theta)}[\hat{\theta}] \succeq I^{-1}(\theta),$$

where $\succeq$ denotes the Löwner @artial ordering of positive semi-definite matrices: $A \succeq B$ iff. $A - B \succ 0$ is positive semi-definite. Another use of the FIM is in gradient descent method using the *natural gradient* (see [6] for its use in deep learning).

Yet, it is common to encounter another equivalent expression of the FIM in the literature [3, 1]:

$$I'_{ij}(\theta) := 4\int \partial_i\sqrt{p(x;\theta)}\partial_j\sqrt{p(x;\theta)}\mathrm{d}\nu(x) \tag{2}$$

This form of the FIM is well-suited to prove that the FIM is always positive semi-definite matrix [1]: $I(\theta) \succeq 0$.

It turns out that one can define a family of equivalent representations of the FIM using the *α-embeddings* of the parametric family. We define the *α-representation* of densities $l^{(\alpha)}(x;\theta) := k_\alpha(p(x;\theta))$ with

$$k_\alpha(u) := \begin{cases} \frac{2}{1-\alpha}u^{\frac{1-\alpha}{2}}, & \text{if } \alpha \neq 1 \\ \log u, & \text{if } \alpha = 1. \end{cases} \tag{3}$$

The function $l^{(\alpha)}(x;\theta)$ is called the *α-likelihood function*.

The α-representation of the FIM (or α-FIM for short) is

$$\boxed{I_{ij}^{(\alpha)}(\theta) := \int \partial_i l^{(\alpha)}(x;\theta)\partial_j l^{(-\alpha)}(x;\theta)\mathrm{d}\nu(x)} \tag{4}$$

In compact notation, we have $I_{ij}^{(\alpha)}(\theta) = \int \partial_i l^{(\alpha)} \partial_j l^{(-\alpha)} \mathrm{d}\nu(x)$ (this is the α-FIM). We can expand the α-FIM expressions as follows

$$I_{ij}^{(\alpha)}(\theta) = \begin{cases} \frac{1}{1-\alpha^2} \int \partial_i p(x;\theta)^{\frac{1-\alpha}{2}} \partial_j p(x;\theta)^{\frac{1+\alpha}{2}} \mathrm{d}\nu(x) & \text{for } \alpha \neq \pm 1 \\ \int \partial_i \log p(x;\theta) \partial_j p(x;\theta) \mathrm{d}\nu(x) & \text{for } \alpha \in \{-1, 1\} \end{cases}$$

The proof that $I_{ij}^{(\alpha)}(\theta) = I_{ij}(\theta)$ follows from the fact that

$$\partial_i l^{\alpha} = p^{-\frac{\alpha+1}{2}} \partial_i p = p^{\frac{1-\alpha}{2}} \partial_i l,$$

since $\partial_i l = \frac{\partial_i p}{p}$.

Therefore we get

$$\partial_i l^{(\alpha)} \partial_j l^{(-\alpha)} = p \partial_i l \partial_j l,$$

and $I_{ij}^{(\alpha)}(\theta) = E[\partial_i l \partial_j l] = I_{ij}(\theta)$.

Thus Eq. 1 and Eq. 2 where two examples of the α-representation, namely the 1-representation and the 0-representation, respectively. The 1-representation of Eq. 1 is called the logarithmic representation, and the 0-representation of Eq. 2 is called the square root representation.

Note that $I_{ij}(\theta) = E[\partial_i l \partial_j l] = \int p \partial_i l \partial_j l \mathrm{d}\nu(x) = \int \partial_i p \partial_j l \mathrm{d}\nu(x) = I_{ij}^{(1)}(\theta)$ since $\partial_i l = \frac{\partial_i p}{p}$

In information geometry [1], $\{\partial_i l^{(\alpha)}\}_i$ plays the role of tangent vectors, the *α-scores*. Geometrically speaking, the tangent plane $T_{p(x;\theta)}$ can be described using any α-base. The statistical manifold $M = \{p(x;\theta)\}_\theta$ is imbedded into the function space $\mathbb{R}^{\mathcal{X}}$, where $\mathcal{X}$ denotes the support of the densities.

Under regular conditions [3, 1], the α-representation of the FIM for $\alpha \neq -1$ can further be rewritten as

$$I_{ij}^{(\alpha)}(\theta) = -\frac{2}{1+\alpha} \int p(x;\theta)^{\frac{1+\alpha}{2}} \partial_i \partial_j l^{(\alpha)}(x;\theta) \mathrm{d}\nu(x). \tag{5}$$

Since we have

$$\partial_i \partial_j l^{(\alpha)}(x;\theta) = p^{\frac{1-\alpha}{2}} \left(\partial_i \partial_j l + \frac{1-\alpha}{2} \partial_i l \partial_j l \right),$$

it follows that

$$I_{ij}^{(\alpha)}(\theta) = -\frac{2}{1+\alpha} \left(-I_{ij}(\theta) + \frac{1-\alpha}{2} I_{ij} \right) = I_{ij}(\theta).$$

Notice that when $\alpha = 1$, we recover the equivalent expression of the FIM (under mild conditions)

$$I_{ij}^{(1)}(\theta) = -E[\nabla^2 \log p(x;\theta)].$$

In particular, when the family is an exponential family [5] with cumulant function $F(\theta)$, we have

$$I(\theta) = \nabla^2 F(\theta) \succ 0.$$

Similarly, the coefficients of the α-connection can be expressed using the α-representation as

$$\Gamma_{ij,k}^{(\alpha)} = \int \partial_i \partial_j l^{(\alpha)} \partial_k^{(-\alpha)} \mathrm{d}\nu(x).$$

The Riemannian metric tensor g_{ij} (a geometric object) can be expressed in matrix form $I_{ij}^{(\alpha)}(\theta)$ using the α-base, and this tensor is called the Fisher metric tensor.

Gauge freedom of the Riemannian metric tensor has been investigated under the framework of (ρ, τ)-monotone embeddings [2] in information geometry: Let ρ and τ be two strictly increasing functions, and f a strictly convex function such that $f'(\rho(u)) = \tau(u)$ (with f^* denoting its convex conjugate). Let us write $p_\theta(x) = p(x; \theta)$.

The (ρ, τ)-metric tensor ${}^{\rho,\tau}g(\theta) = [{}^{\rho,\tau}g_{ij}(\theta)]_{ij}$ can be derived [4] from the (ρ, τ)-divergence:

$$D_{\rho,\tau}(p : q) = \int \left(f(\rho(p(x))) + f^*(\tau(q(x))) - \rho(p(x))\tau(q(x))\right) \mathrm{d}\nu(x) \tag{6}$$

We have:

$$\begin{aligned}
{}^{\rho,\tau}g_{ij}(\theta) &= \int (\partial_i \rho(p_\theta(x)))\,(\partial_j \tau(p_\theta(x)))\,\mathrm{d}\nu(x), && (7)\\
&= \int \rho'(p_\theta(x))\tau'(p_\theta(x))\,(\partial_i p_\theta(x))\,(\partial_j p_\theta(x))\,\mathrm{d}\nu(x), && (8)\\
&= \int f''(\rho(p_\theta(x)))\,(\partial_i \rho(p_\theta(x)))\,(\partial_j \rho(p_\theta(x)))\,\mathrm{d}\nu(x), && (9)\\
&= \int (f^*)''(\tau(p_\theta(x)))\,(\partial_i \tau(p_\theta(x)))\,(\partial_j \tau(p_\theta(x)))\,\mathrm{d}\nu(x). && (10)
\end{aligned}$$

The second equation shows that there is a gauge function freedom $\rho'(u)\tau'(u)$ when calculating the (ρ, τ)-Riemannian metric.

Initially created 19th September 2017 (last updated September 2, 2020).

References

[1] O. Calin and C. Udrişte. *Geometric Modeling in Probability and Statistics*. Mathematics and Statistics. Springer International Publishing, 2014.

[2] Jan Naudts and Jun Zhang. Rho–tau embedding and gauge freedom in information geometry. *Information geometry*, 1(1):79–115, 2018.

[3] Frank Nielsen. Cramér-Rao lower bound and information geometry. *arXiv preprint arXiv:1301.3578*, 2013.

[4] Frank Nielsen. An elementary introduction to information geometry. *arXiv preprint arXiv:1808.08271*, 2018.

[5] Frank Nielsen and Vincent Garcia. Statistical exponential families: A digest with flash cards. *arXiv preprint arXiv:0911.4863*, 2009.

[6] Ke Sun and Frank Nielsen. Relative Fisher information and natural gradient for learning large modular models. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3289–3298, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.