Fisher-Rao distance on the covariance cone

Joseph Wells, Mary Cook, Karleigh Cameron, Benjamin Robinson

March 5, 2018

1 Introduction

Information geometry is the application of ideas from differential geometry to the field of statistics. Rao [9] was the first to observe that the Fisher information matrix forms a Riemannian metric for certain models, inducing on them the structure of a Riemannian manifold. Amari [3, 2, 4, 1] extended his observations considerably, discussing models which additionally possess so-called flat dualistic affine connections. Perhaps the culmination of his work is a novel proof of the Expectation-Maximization Theorem within the geometric context. Another famous result from information geometry is Cencov's theorem [7, Theorem 11.1], which characterizes all Riemmanian metrics and affine connections which are invariant to sufficient statistics for the family of models of all nonzero densities on finite outcome spaces. Other notable authors who have studied information geometry are Murray and Rice [8]; Le, Ay, Jost, Schwachhofer [5]; and Bauer, Bruveris, and Michor [6]; and there have been a host of others claiming novel applications, although these claims usually are without proof, or are just reproofs of facts from statistics that were already well-known.

Another notable author who has studied information geometry is Smith [10]. Smith developed, using the structure of an affine connection on a statistical model, a family of Cramer-Rao-type lower bounds on expected squared geodesic distance for fairly general models. Smith focused in particular on the model of a multivariate zero-mean Gaussian. Identifying zero-mean Gaussians with their covariances naturally gives this model the structure of the cone of positive-definite symmetric (or Hermitian) matrices. For this model, Smith derives a Cramer-Rao-type lower bound on the expected squared Fisher-Rao distance—the infimum of the lengths of all smooth paths between two points, where differential length corresponds to the Fisher information Riemannian metric. Unlike the usual Cramer-Rao bound, this bound is independent of the true underlying parameter, making it potentially more useful since this parameter is in fact unknown. The question is why one should care about the expected squared Fisher-Rao distance, or about Fisher-Rao distance at all? This debate usually boils down to claims about "naturality" of the distance, but usually has little substance. In this article we do not engage in this discussion, but rather seek to prove explicitly a closed-form expression for Fisher-Rao distance sketched by Smith, leaving the application of this distance to future work. More precisely,

we seek to prove the following Theorem, which corresponds to the case of real scalars. (Smith's result concerns the complex case, and that case is similar.)

Theorem 1. *The Fisher-Rao distance between two covariance matrices* $\mathbf{R}$ *and* $\mathbf{S} \in \mathbb{R}^{p\times p}$ *is given by*

$$d(\mathbf{R},\mathbf{S})^2 = \frac{1}{2}\mathrm{tr}\left[\left(\log \mathbf{R}^{-1/2}\mathbf{S}\mathbf{R}^{-1/2}\right)^2\right].$$

2 Proof of the main theorem

Let $\mathcal{P}$ denote the set of $p \times p$ covariance matrices and M be the model of p-dimensional real Gaussians with mean 0:

$$M = \left\{ p(\mathbf{x};\mathbf{R}) = \frac{1}{(2\pi)^{p/2}|\mathbf{R}|^{1/2}} \exp\left(-\frac{1}{2}\mathbf{x}^\top \mathbf{R}^{-1}\mathbf{x}\right) : \mathbf{R} \in \mathcal{P} \right\}.$$

M and $\mathcal{P}$ are naturally identified. We give $\mathcal{P}$ the smooth structure of $\mathbb{R}^{p^2}$, and since it embeds smoothly into $\mathbb{R}^{p^2}$, tangent vectors are precisely derivatives of smooth paths through $\mathcal{P}$. This means that tangent vectors are precisely $p \times p$ symmetric matrices. To see this, suppose $\mathbf{D}$ is a symmetric $p \times p$ matrix and let $\gamma : (-\epsilon, \epsilon) \to \mathcal{P}$ be the path $\mathbf{R} + t\mathbf{D}$. The derivative of this path at zero is $\mathbf{D}$. Conversely, suppose γ is a smooth path through $\mathcal{P}$ originating from $\mathbf{R}$. Then the difference quotient

$$\frac{\gamma(t) - \gamma(0)}{t}$$

is always symmetric; thus, so is its limit.

The proof of the main theorem will rely on the following lemmas:

Lemma 1. *The Fisher information metric* g *for the model* M *at the point* $\mathbf{R}$ *is given by*

$$g_{\mathbf{R}}(\mathbf{A},\mathbf{B}) = \frac{1}{2}\mathrm{tr}\mathbf{R}^{-1}\mathbf{A}\mathbf{R}^{-1}\mathbf{B},$$

where $\mathbf{A}, \mathbf{B}$ *are real symmetric matrices.*

Proof. Given a parametric model $\{p(\cdot;\theta)d\mu : \theta \in \Theta \subset \mathbb{R}^N\}$ on a sample space $\mathcal{X}$, for some connected open set Θ, the Fisher information matrix is

$$(g_\theta)_{ij} = \int_{\mathcal{X}} \frac{\partial}{\partial\theta^i} \log p(x;\theta)\, \frac{\partial}{\partial\theta^j} \log p(x;\theta)\, p(x;\theta) d\mu(x),$$

when defined. Under suitable regularity conditions, which will hold here for the Gaussian model, this can be reexpressed as

$$(g_\theta)_{ij} = -\int_{\mathcal{X}} \frac{\partial^2}{\partial\theta^i\partial\theta^j} \log p(x;\theta)\, p(x;\theta) d\mu(x),$$

This extends by bilinarity to a 2-tensor

$$g_\theta(v,w) = -\int_{\mathcal{X}} vw(\log p(x;\cdot))\, p(x;\theta) d\mu(x), \tag{1}$$

where v and w are tangent vectors at θ.

For the covariance model, let us consider $g_{\mathbf{R}}(\mathbf{D},\mathbf{D})$, where $\mathbf{D}$ is a tangent vector at $\mathbf{R}$. We abuse notation and use the field of symmetric matrices $\mathbf{D}$ interchangeably with the derivation that corresponds to it.

We have

$$\mathbf{D}(\log p(\mathbf{x};\cdot)) = \left.\frac{d}{dt}\log p(\mathbf{x};\mathbf{R}+t\mathbf{D})\right|_{t=0}.$$

On the right side, we have

$$\log p(\mathbf{x};\mathbf{R}+t\mathbf{D}) = \log\frac{1}{(2\pi)^{p/2}} - \frac{1}{2}\mathrm{tr}\log\left(\mathbf{R}+t\mathbf{D}\right) - \frac{1}{2}\mathbf{x}^\top\left(\mathbf{R}+t\mathbf{D}\right)^{-1}\mathbf{x}.$$

Taking the derivative for t and setting $t=0$ gives

$$\mathbf{D}(\log p(\mathbf{x};\cdot)) = -\frac{1}{2}\mathrm{tr}\mathbf{R}^{-1}\mathbf{D} + \frac{1}{2}\mathbf{x}^\top\mathbf{R}^{-1}\mathbf{D}\mathbf{R}^{-1}\mathbf{x}.$$

Applying $\mathbf{D}$ again yields

$$\mathbf{D^2}(\log p(x;\cdot)) = \frac{1}{2}\mathrm{tr}\mathbf{R}^{-1}\mathbf{D}\mathbf{R}^{-1}\mathbf{D} - \mathbf{x}^\top\mathbf{R}^{-1}\mathbf{D}\mathbf{R}^{-1}\mathbf{D}\mathbf{R}^{-1}\mathbf{x}. \tag{2}$$

The latter term can be written

$$\mathrm{tr}\left(\mathbf{R}^{-1}\mathbf{D}\mathbf{R}^{-1}\mathbf{D}\mathbf{R}^{-1}\mathbf{x}\mathbf{x}^\top\right),$$

so integrating this term against $p(x;\mathbf{R})$ yields

$$\mathrm{tr}\left(\mathbf{R}^{-1}\mathbf{D}\mathbf{R}^{-1}\mathbf{D}\mathbf{R}^{-1}\mathbf{R}\right) = \mathrm{tr}\left((\mathbf{R}^{-1}\mathbf{D})^2\right).$$

Adding in the first term of (2) and applying the negative sign in (1) gives

$$g_{\mathbf{R}}(\mathbf{D},\mathbf{D}) = \frac{1}{2}\mathrm{tr}\left((\mathbf{R}^{-1}\mathbf{D})^2\right).$$

The result follows from polarization.

□

Although it is assumed that the reader is at least somewhat familiar with Riemannian geometry, we recall a few definitions and set the tone notationally. Using $\mathfrak{X}(M)$ to denote the smooth vector fields on a smooth manifold M, an *affine connection* is a map

$$\begin{aligned}\nabla : \mathfrak{X}(M)\times\mathfrak{X}(M) &\to \mathfrak{X}(M)\\ (X,Y) &\mapsto \nabla_X Y\end{aligned}$$

that is $C^\infty(M)$-linear in the first coordinate, $\mathbb{R}$-linear in the second coordinate, and for all $f \in C^\infty(M)$ satisfies

$$\nabla_X(fY) = f\nabla_X Y + (Xf)Y.$$

Given a Riemannian metric g, the *Levi-Civita connection* is an affine connection ∇ that also satisfies the following for all $X, Y, Z \in \mathfrak{X}(M)$:

1. $Xg(Y,Z) = g(\nabla_X Y, Z) + g(Y, \nabla_X Z)$, and
2. $\nabla_X Y - \nabla_Y X = XY - YX$.

As is well-known, the Levi-Civita connection for a given metric is the unique connection with these properties. (This is sometimes called the *Fundamental Theorem of Riemannian Geometry*.)

Given a local frame (∂_i) for the tangent bundle TM, we have that the connection coefficients satisfy

$$\nabla_{\partial_i}\partial_j = \Gamma_{ij}^k \partial_k$$

and these coefficients Γ_{ij}^k are called *Christoffel symbols*. A path $\gamma(t) = (\gamma_i(t))$ in M is called a *geodesic* if it satisfies the *geodesic equation*

$$\ddot{\gamma}_k(t) + \Gamma_{ij}^k \dot{\gamma}_i(t)\dot{\gamma}_j(t).$$

Here it is understood that Γ_{ij}^k is in $C^\infty(M)$ and that the above equation holds at $\boldsymbol{\gamma}(t)$. It is sometimes convenient to express the quadratic terms above in vector notation:

$$\boldsymbol{\Gamma}_{\boldsymbol{\gamma}(t)}\left(\dot{\boldsymbol{\gamma}}(t), \dot{\boldsymbol{\gamma}}(t)\right).$$

Given $p \in M$ and an affine connection ∇, we say that an open set U containing p is a *normal neighborhood* of p iff for all $q \in U$, the solution to the geodesic equation with boundary conditions p and q is unique. It is well known (see, e.g., Proposition 3.6 of [**?**]) that if ∇ is the Levi-Civita connection corresponding to a Riemannian metric g, the geodesics within a normal neighborhood are length-minimizing. It is also well-known that the covariance cone with the Fisher information metric has non-positive sectional curvature, making the whole cone a normal neighborhood of every point. Thus, Fisher-Rao distance, which we have defined as an infimum of path distances, is achieved by the geodesics corresponding to the Levi-Civita connection, and we need only find the lengths of these geodesics.

We note that it is well-known that if ∇ is an affine connection, $\boldsymbol{\Gamma}$ is the corresponding Christoffel symbol, and $X, Y \in \mathfrak{X}(M)$, then

$$\nabla_X Y = XY + \boldsymbol{\Gamma}(X, Y). \tag{3}$$

We are almost ready to prove our theorem, but first we need a couple of lemmas.

Lemma 2. *If* $\mathbf{X} \in \mathfrak{X}(M)$ *and* $f(\mathbf{R}) = \mathbf{R}^{-1}$, *then* $(\mathbf{X}f)(\mathbf{R}) = -\mathbf{R}^{-1}\mathbf{X}\mathbf{R}^{-1}$, *where* $\mathbf{X}$ *on the right side is considered as a field of symmetric matrices.*

Proof. Let a be an invertible square matrix, b be a square matrix of the same size, and let us compute the difference quotient for the derivative of $(a+hb)^{-1}$ at $h=0$:

$$\begin{aligned}\frac{1}{h}\left((a+hb)^{-1}-a^{-1}\right) &= \frac{1}{h}(a+hb)^{-1}\left(1-(a+hb)a^{-1}\right)\\ &= \frac{1}{h}-(a+hb)^{-1}hba^{-1}\end{aligned}$$

Taking the limit as $h\to 0$, we get $-a^{-1}ba^{-1}$.

The proof is completed by fixing a base point $p\in M$ and taking $a=\mathbf{R}_p$ and $b=\mathbf{X}_p$. □

Lemma 3. *The map $\boldsymbol{\gamma}:(-\varepsilon,\varepsilon)\to\mathcal{P}$ given by*

$$\boldsymbol{\gamma}(t)=\mathbf{R}^{1/2}\exp\Big(t\mathbf{R}^{-1/2}\mathbf{D}\mathbf{R}^{-1/2}\Big)\mathbf{R}^{1/2} \tag{4}$$

is a geodesic emanating from $\mathbf{R}$ in the direction of $\mathbf{D}$. (Here exp *is the usual matrix exponential.)*

Proof. In this proof and what follows we will denote $\boldsymbol{\gamma}$ without boldface, and it will be understood that it is matrix-valued.

We first claim that the Christoffel symbols of the Levi-Civita connection are given by

$$\boldsymbol{\Gamma}_{\mathbf{R}}(\mathbf{A},\mathbf{B})=-\frac{1}{2}\left(\mathbf{A}\mathbf{R}^{-1}\mathbf{B}+\mathbf{B}\mathbf{R}^{-1}\mathbf{A}\right).$$

(This is stated without proof in [10].) To do this, we must prove that $\boldsymbol{\Gamma}$ satisfies (3) for some affine connection and verify properties 1 and 2 above. Indeed, by defining ∇ as in (3) with our $\boldsymbol{\Gamma}$ as above, it is straightforward to show that ∇ satisfies the properties of an affine connection. Property 2 follows from the definition of ∇ and the fact that $\boldsymbol{\Gamma}_{\mathbf{R}}(\mathbf{A},\mathbf{B})-\boldsymbol{\Gamma}_{\mathbf{R}}(\mathbf{B},\mathbf{A})=0$. For property 1, let $\mathbf{X},\mathbf{Y},\mathbf{Z}\in\mathfrak{X}(\mathcal{P})$. We then have

$$\begin{aligned}\mathbf{X}g_{\mathbf{R}}(\mathbf{Y},\mathbf{Z}) = &-\frac{1}{2}\mathrm{tr}\left(\mathbf{R}^{-1}\mathbf{X}\mathbf{R}^{-1}\mathbf{Y}\mathbf{R}^{-1}\mathbf{Z}\right)+\frac{1}{2}\mathrm{tr}\left(\mathbf{R}^{-1}\mathbf{X}\mathbf{Y}\mathbf{R}^{-1}\mathbf{Z}\right)+\\ &-\frac{1}{2}\mathrm{tr}\left(\mathbf{R}^{-1}\mathbf{Y}\mathbf{R}^{-1}\mathbf{X}\mathbf{R}^{-1}\mathbf{Z}\right)+\frac{1}{2}\mathrm{tr}\left(\mathbf{R}^{-1}\mathbf{Y}\mathbf{R}^{-1}\mathbf{X}\mathbf{Z}\right)\end{aligned} \tag{5}$$

(Here we have used Lemma 2 and are again using derivations and vector fields interchangeably.) On the other hand

$$\begin{aligned}&g_{\mathbf{R}}\left(\nabla_{\mathbf{X}}\mathbf{Y},\mathbf{Z}\right)\\ &=g_{\mathbf{R}}(\mathbf{X}\mathbf{Y},\mathbf{Z})+g_{\mathbf{R}}(\boldsymbol{\Gamma}_{\mathbf{R}}(\mathbf{X},\mathbf{Y}),\mathbf{Z})\\ &=\frac{1}{2}\mathrm{tr}\left(\mathbf{R}^{-1}\mathbf{X}\mathbf{Y}\mathbf{R}^{-1}\mathbf{Z}\right)+\frac{1}{2}\mathrm{tr}\left(\mathbf{R}^{-1}\left(-\frac{1}{2}\right)\left(\mathbf{X}\mathbf{R}^{-1}\mathbf{Y}+\mathbf{Y}\mathbf{R}^{-1}\mathbf{X}\right)\mathbf{R}^{-1}\mathbf{Z}\right)\end{aligned}$$

and

$$\begin{aligned}
&g_{\mathbf{R}}\left(\mathbf{Y}, \nabla_{\mathbf{X}}\mathbf{Z}\right) \\
&= g_{\mathbf{R}}(\mathbf{Y}, \mathbf{XZ}) + g_{\mathbf{R}}(\mathbf{Y}, \mathbf{\Gamma}_{\mathbf{R}}(\mathbf{X}, \mathbf{Z})) \\
&= \frac{1}{2}\mathrm{tr}\left(\mathbf{R}^{-1}\mathbf{Y}\mathbf{R}^{-1}\mathbf{XZ}\right) + \frac{1}{2}\mathrm{tr}\left(\mathbf{R}^{-1}\mathbf{Y}\mathbf{R}^{-1}\left(-\frac{1}{2}\right)\left(\mathbf{X}\mathbf{R}^{-1}\mathbf{Z} + \mathbf{Z}\mathbf{R}^{-1}\mathbf{X}\right)\right)
\end{aligned}$$

Using the circulant property of trace, the sum of these two expressions is easily seen to equal (5).

We now wish to prove that $\gamma = \gamma(t)$ satisfies the geodesic equation

$$\ddot{\gamma} + \mathbf{\Gamma}_{\gamma}(\dot{\gamma}, \dot{\gamma}) = 0. \tag{6}$$

For simplicity, write $\mathbf{E} = \exp\left(t\mathbf{R}^{-1/2}\mathbf{D}\mathbf{R}^{-1/2}\right)$. Then we have that

$$\frac{d\mathbf{E}}{dt} = \mathbf{R}^{-1/2}\mathbf{D}\mathbf{R}^{-1/2}\mathbf{E} = \mathbf{E}\mathbf{R}^{-1/2}\mathbf{D}\mathbf{R}^{-1/2}$$

whence

$$\begin{aligned}
\gamma &= \mathbf{R}^{1/2}\mathbf{E}\mathbf{R}^{1/2}, \\
\dot{\gamma} &= \mathbf{D}\mathbf{R}^{-1/2}\mathbf{E}\mathbf{R}^{1/2} = \mathbf{R}^{1/2}\mathbf{E}\mathbf{R}^{-1/2}\mathbf{D}, \\
\ddot{\gamma} &= \mathbf{D}\mathbf{R}^{-1/2}\mathbf{E}\mathbf{R}^{-1/2}\mathbf{D}.
\end{aligned}$$

It is then straightforward to compute

$$\ddot{\gamma} + \mathbf{\Gamma}_{\gamma}(\dot{\gamma}, \dot{\gamma}) = \ddot{\gamma} - \dot{\gamma}\gamma^{-1}\dot{\gamma} = 0.$$

□

Proof of Theorem 1. Let $\gamma = \gamma(t)$ be a geodesic from $\mathbf{R}$ to $\mathbf{S}$. We may parameterize γ so that $\gamma(0) = \mathbf{R}$ and $\gamma(1) = \mathbf{S}$. Since γ takes the form of the map in Equation (4), we solve for $\mathbf{D}$ in the equation $\mathbf{S} = \gamma(1)$ to get

$$\mathbf{D} = \mathbf{R}^{1/2}\mathbf{L}\mathbf{R}^{1/2}$$

where

$$\mathbf{L} = \log \mathbf{R}^{-1/2}\mathbf{S}\mathbf{R}^{-1/2}.$$

We then have that

$$\begin{aligned} g_\gamma(\dot\gamma,\dot\gamma) &= \frac{1}{2}\mathrm{tr}\Big[\big(\gamma^{-1}\dot\gamma\big)^2\Big] \\ &= \frac{1}{2}\mathrm{tr}\Big[\big(\mathbf{R}^{-1}\mathbf{D}\big)^2\Big] \\ &= \frac{1}{2}\mathrm{tr}\Big[\Big(\mathbf{R}^{-1/2}\mathbf{L}\mathbf{R}^{1/2}\Big)^2\Big] \\ &= \frac{1}{2}\mathrm{tr}\Big[\mathbf{R}^{-1/2}\mathbf{L}^2\mathbf{R}^{1/2}\Big] \\ &= \frac{1}{2}\mathrm{tr}\big[\mathbf{L}^2\big] \\ &= \frac{1}{2}\mathrm{tr}\Big[\Big(\log\mathbf{R}^{-1/2}\mathbf{S}\mathbf{R}^{-1/2}\Big)^2\Big] \end{aligned}$$

The distance between $\mathbf{R}$ and $\mathbf{S}$ is just the length of the geodesic segment between these two covariance matrices

$$d(\mathbf{R},\mathbf{S}) = \int_0^1 \sqrt{g_\gamma(\dot\gamma,\dot\gamma)}\,dt = \sqrt{\frac{1}{2}\mathrm{tr}\Big[\big(\log\mathbf{R}^{-1/2}\mathbf{S}\mathbf{R}^{-1/2}\big)^2\Big]}.$$

□

Remark 1. Note that this distance is consistent (up to a factor of $1/\sqrt{2}$) with the distance for complex scalars derived in [10] for complex scalars. Smith's distance squared is

$$\sum_j (\log\lambda_j)^2,$$

where λ_j are the generalized eigenvalues of the pencil $\mathbf{S}-\lambda\mathbf{R}$. These may be defined as the roots of the following polynomial in λ: $\det(\mathbf{S}-\lambda\mathbf{R})$. But these roots are the same as those of $\det(\mathbf{R}^{-1/2}\mathbf{S}\mathbf{R}^{-1/2}-\lambda 1)$; thus, the sum of the squares of the logs of these eigenvalues is precisely the trace appearing in the last theorem.

References

[1] S-I Amari. Information geometry on hierarchy of probability distributions. *IEEE transactions on information theory*, 47(5):1701–1711, 2001.

[2] Shun-ichi Amari. Differential geometrical theory of statistics. *Amari et al. ABNK+87*, pages 19–94, 1987.

[3] Shun-ichi Amari. *Differential-geometrical methods in statistics*, volume 28. Springer Science & Business Media, 2012.

[4] Shun-ichi Amari and Hiroshi Nagaoka. *Methods of Information Geometry*, volume 191. American Mathematical Society, 2007.

[5] Nihat Ay, Jürgen Jost, Hông Vân Lê, and Lorenz Schwachhöfer. Information geometry and sufficient statistics. *Probability Theory and Related Fields*, 162(1-2):327–364, 2015.

[6] Martin Bauer, Martins Bruveris, and Peter W. Michor. Uniqueness of the Fisher–Rao metric on the space of smooth densities. *Bulletin of the London Mathematical Society*, 48(3):499–506, 2016.

[7] Nikolai Nikolaevich Cencov. *Statistical Decision Rules and Optimal Inference*. Number 53. American Mathematical Soc., 2000.

[8] Michael K. Murray and John W. Rice. *Differential Geometry and Statistics*, volume 48. CRC Press, 1993.

[9] C. Radhakrishna Rao. Information and accuracy attainable in the estimation of statistical parameters. *Bull. Calcutta Math. Soc.*, 37:81–91, 1945.

[10] Steven Thomas Smith. Covariance, subspace, and intrinsic Cramer-Rao bounds. *IEEE Transactions on Signal Processing*, 53(5):1610–1630, 2005.