The Basics of Information Geometry

Ariel Caticha

Department of Physics, University at Albany–SUNY, Albany, NY 12222, USA

Abstract.
To what extent can we distinguish one probability distribution from another? Are there quantitative measures of distinguishability? The goal of this tutorial is to approach such questions by introducing the notion of the "distance" between two probability distributions and exploring some basic ideas of such an "information geometry".

Keywords: Entropic dynamics, Quantum Theory, Maximum Entropy

Einstein, 1949: "[The basic ideas of General Relativity were in place] ... *in 1908. Why were another seven years required for the construction of the general theory of relativity? The main reason lies in the fact that it is not so easy to free oneself from the idea that coordinates must have an immediate metrical meaning.*" [1]

INTRODUCTION

A main concern of any theory of inference is the problem of updating probabilities when new information becomes available. We want to pick a probability distribution from a set of candidates and this immediately raises many questions. What if we had picked a neighboring distribution? What difference would it make? What makes two distributions similar? To what extent can we distinguish one distribution from another? Are there quantitative measures of distinguishability? The goal of this tutorial is to address such questions by introducing methods of geometry. More specifically the goal will be to introduce a notion of "distance" between two probability distributions.

A parametric family of probability distributions is a set of distributions $p_\theta(x)$ labeled by parameters $\theta = (\theta^1 \ldots \theta^n)$. Such a family forms a *statistical manifold*, namely, a space in which each point, labeled by coordinates θ, represents a probability distribution $p_\theta(x)$. Generic manifolds do not come with an intrinsic notion of distance; such additional structure has to be supplied separately in the form of a metric tensor. Statistical manifolds are, however, an exception. One of the main goals of this chapter is to show that statistical manifolds possess a uniquely natural notion of distance — the so-called information metric. This metric is not an optional feature; it is inevitable. *Geometry is intrinsic to the structure of statistical manifolds.*

The distance $d\ell$ between two neighboring points θ and $\theta + d\theta$ is given by Pythagoras' theorem which, written in terms of a metric tensor g_{ab}, is[1]

$$d\ell^2 = g_{ab} d\theta^a d\theta^b \ . \tag{1}$$

The singular importance of the metric tensor g_{ab} derives from a theorem due to N. Čencov that states that the metric g_{ab} on the manifold of probability distributions is essentially unique: up to an overall scale factor there is only one metric that takes into account the fact that these are not distances between simple structureless dots but distances between probability distributions. [2]

We will not develop the subject in all its possibilities[2] but we do wish to emphasize one specific result. Having a notion of distance means we have a notion of volume and this in turn implies that *there is a unique and objective notion of a distribution that is uniform over the space of parameters* — equal volumes are assigned equal probabilities. Whether such uniform distributions are maximally non-informative, or whether they define ignorance, or whether they reflect the actual prior beliefs of any rational agent, are all important issues but they are quite beside the specific point that we want to make, namely, that they are uniform — and this is not a matter of subjective judgment but of objective mathematical proof.

EXAMPLES OF STATISTICAL MANIFOLDS

An n-dimensional manifold $\mathscr{M}$ is a smooth, possibly curved, space that is locally like $\mathscr{R}^n$. What this means is that one can set up a coordinate frame (that is a map $\mathscr{M} \to \mathscr{R}^n$) so that each point $\theta \in \mathscr{M}$ is identified or labelled by its coordinates, $\theta = (\theta^1 \ldots \theta^n)$. A statistical manifold is a manifold in which each point θ represents a probability distribution $p_\theta(x)$. As we shall later see, a very convenient notation is $p_\theta(x) = p(x|\theta)$. Here are some examples:

The **multinomial distributions** are given by

$$p(\{n_i\}|\theta) = \frac{N!}{n_1! n_2! \ldots n_m!} (\theta^1)^{n_1} (\theta^2)^{n_2} \ldots (\theta^m)^{n_m} \ , \tag{2}$$

where $\theta = (\theta^1, \theta^2 \ldots \theta^m)$, $N = \sum_{i=1}^m n_i$ and $\sum_{i=1}^m \theta^i = 1$. They form a statistical manifold of dimension $(m-1)$ called a simplex, S_{m-1}. The parameters $\theta = (\theta^1, \theta^2 \ldots \theta^m)$ are a convenient choice of coordinates.

The **multivariate Gaussian distributions** with means μ^a, $a = 1 \ldots n$, and variance σ^2,

$$p(x|\mu, \sigma) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp -\frac{1}{2\sigma^2} \sum_{a=1}^{n} (x^a - \mu^a)^2 \ , \tag{3}$$

---

[1] The use of superscripts rather than subscripts for the indices labelling coordinates is a standard and very convenient notational convention in differential geometry. We adopt the standard convention of summing over repeated indices, for example, $g_{ab} f^{ab} = \sum_a \sum_b g_{ab} f^{ab}$.

[2] For a more extensive treatment see [3][4]. Here we follow closely the presentation in [5].

form an $(n+1)$-dimensional statistical manifold with coordinates $\theta = (\mu^1, \ldots, \mu^n, \sigma^2)$.

The **canonical distributions**,

$$p(i|F) = \frac{1}{Z} e^{-\lambda_k f_i^k} \ , \tag{4}$$

are derived by maximizing the Shannon entropy $S[p]$ subject to constraints on the expected values of n functions $f_i^k = f^k(x_i)$ labeled by superscripts $k = 1, 2, \ldots n$,

$$\left\langle f^k \right\rangle = \sum_i p_i f_i^k = F^k \ . \tag{5}$$

They form an n-dimensional statistical manifold. As coordinates we can either use the expected values $F = (F^1 \ldots F^n)$ or, equivalently, the Lagrange multipliers, $\lambda = (\lambda_1 \ldots \lambda_n)$.

DISTANCE AND VOLUME IN CURVED SPACES

The basic intuition behind differential geometry derives from the observation that curved spaces are locally flat: curvature effects can be neglected provided one remains within a sufficiently small region. The idea then is rather simple: within the close vicinity of any point x we can always transform from the original coordinates x^a to new coordinates $\hat{x}^\alpha = \hat{x}^\alpha(x^1 \ldots x^n)$ that we *declare* to be locally Cartesian (here denoted with a hat and with Greek superscripts, $\hat{x}^\alpha$). An infinitesimal displacement is given by

$$d\hat{x}^\alpha = X_a^\alpha \, dx^a \quad \text{where} \quad X_a^\alpha = \frac{\partial \hat{x}^\alpha}{\partial x^a} \tag{6}$$

and the corresponding infinitesimal distance can be computed using Pythagoras theorem,

$$d\ell^2 = \delta_{\alpha\beta} d\hat{x}^\alpha d\hat{x}^\beta \ . \tag{7}$$

Changing back to the original frame

$$d\ell^2 = \delta_{\alpha\beta} d\hat{x}^\alpha d\hat{x}^\beta = \delta_{\alpha\beta} X_a^\alpha X_b^\beta \, dx^a dx^b \ . \tag{8}$$

Defining the quantities

$$g_{ab} \equiv \delta_{\alpha\beta} X_a^\alpha X_b^\beta \ , \tag{9}$$

we can write the infinitesimal Pythagoras theorem in generic coordinates x^a as

$$d\ell^2 = g_{ab} dx^a dx^b \ . \tag{10}$$

The quantities g_{ab} are the components of the metric tensor. One can easily check that under a coordinate transformation g_{ab} transforms according to

$$g_{ab} = X_a^{a'} X_a^{b'} g_{a'b'} \quad \text{where} \quad X_a^{a'} = \frac{\partial x^{a'}}{\partial x^a} \, , \tag{11}$$

so that the infinitesimal distance $d\ell$ is independent of the choice of coordinates.

To find the finite length between two points along a curve $x(\lambda)$ one integrates along the curve,

$$\ell = \int_{\lambda_1}^{\lambda_2} d\ell = \int_{\lambda_1}^{\lambda_2} \left(g_{ab} \frac{dx^a}{d\lambda} \frac{dx^b}{d\lambda} \right)^{1/2} d\lambda \ . \tag{12}$$

Once we have a measure of distance we can also measure angles, areas, volumes and all sorts of other geometrical quantities. To find an expression for the n-dimensional volume element dV_n we use the same trick as before: transform to locally Cartesian coordinates so that the volume element is simply given by the product

$$dV_n = d\hat{x}^1 d\hat{x}^2 \ldots d\hat{x}^n \ , \tag{13}$$

and then transform back to the original coordinates x^a using eq.(6),

$$dV_n = \left| \frac{\partial \hat{x}}{\partial x} \right| dx^1 dx^2 \ldots dx^n = |\det X_a^\alpha| \, d^n x \ . \tag{14}$$

This is the volume we seek written in terms of the coordinates x^a but we still have to calculate the Jacobian of the transformation, $|\partial \hat{x}/\partial x| = |\det X_a^\alpha|$. The transformation of the metric from its Euclidean form $\delta_{\alpha\beta}$ to g_{ab}, eq.(9), is the product of three matrices. Taking the determinant we get

$$g \equiv \det(g_{ab}) = [\det X_a^\alpha]^2 \ , \tag{15}$$

so that

$$|\det (X_a^\alpha)| = g^{1/2} \ . \tag{16}$$

We have succeeded in expressing the volume element in terms of the metric $g_{ab}(x)$ in the original coordinates x^a. The answer is

$$dV_n = g^{1/2}(x) d^n x \ . \tag{17}$$

The volume of any extended region on the manifold is

$$V_n = \int dV_n = \int g^{1/2}(x) d^n x \ . \tag{18}$$

Example: A uniform distribution over such a curved manifold is one which assigns equal probabilities to equal volumes,

$$p(x) d^n x \propto g^{1/2}(x) d^n x \ . \tag{19}$$

Example: For Euclidean space in spherical coordinates (r, θ, ϕ),

$$d\ell^2 = dr^2 + r^2 d\theta^2 + r^2 \sin^2\theta d\phi^2 \ , \tag{20}$$

and the volume element is the familiar expression

$$dV = g^{1/2} dr d\theta d\phi = r^2 \sin\theta \, dr d\theta d\phi \ . \tag{21}$$

TWO DERIVATIONS OF THE INFORMATION METRIC

The distance $d\ell$ between two neighboring distributions $p(x|\theta)$ and $p(x|\theta+d\theta)$ or, equivalently, between the two points θ and $\theta+d\theta$, is given by the metric g_{ab}. Our goal is to compute the tensor g_{ab} corresponding to $p(x|\theta)$. We give a couple of derivations which illuminate the meaning of the information metric, its interpretation, and ultimately, how it is to be used. Other derivations based on asymptotic inference are given in [6] and [7].

At this point a word of caution (and encouragement) might be called for. Of course it is possible to be confronted with sufficiently singular families of distributions that are not smooth manifolds and studying their geometry might seem a hopeless enterprise. Should we give up on geometry? No. The fact that statistical manifolds can have complicated geometries does not detract from the value of the methods of information geometry any more than the existence of surfaces with rugged geometries detracts from the general value of geometry itself.

Derivation from distinguishability

We seek a quantitative measure of the extent that two distributions $p(x|\theta)$ and $p(x|\theta+d\theta)$ can be distinguished. The following argument is intuitively appealing. [8][9] The advantage of this approach is that it clarifies the interpretation — *the metric measures distinguishability*. Consider the relative difference,

$$\Delta = \frac{p(x|\theta+d\theta)-p(x|\theta)}{p(x|\theta)} = \frac{\partial \log p(x|\theta)}{\partial\theta^a}\,d\theta^a. \tag{22}$$

The expected value of the relative difference, $\langle\Delta\rangle$, might seem a good candidate, but it does not work because it vanishes identically,

$$\langle\Delta\rangle = \int dx\, p(x|\theta)\,\frac{\partial \log p(x|\theta)}{\partial\theta^a}\,d\theta^a = d\theta^a\,\frac{\partial}{\partial\theta^a}\int dx\, p(x|\theta) = 0. \tag{23}$$

(Depending on the problem the symbol $\int dx$ may represent either discrete sums or integrals over one or more dimensions; its meaning should be clear from the context.) However, the variance does not vanish,

$$d\ell^2 = \langle\Delta^2\rangle = \int dx\, p(x|\theta)\,\frac{\partial \log p(x|\theta)}{\partial\theta^a}\,\frac{\partial \log p(x|\theta)}{\partial\theta^b}\,d\theta^a d\theta^b\ . \tag{24}$$

This is the measure of distinguishability we seek; a small value of $d\ell^2$ means that the relative difference Δ is small and the points θ and $\theta+d\theta$ are difficult to distinguish. It suggests introducing the matrix g_{ab}

$$g_{ab}(\theta) \stackrel{\text{def}}{=} \int dx\, p(x|\theta)\,\frac{\partial \log p(x|\theta)}{\partial\theta^a}\,\frac{\partial \log p(x|\theta)}{\partial\theta^b} \tag{25}$$

called the Fisher information *matrix* [10], so that

$$d\ell^2 = g_{ab}\, d\theta^a d\theta^b \ . \tag{26}$$

Up to now no notion of distance has been introduced. Normally one says that the reason it is difficult to distinguish two points in say, the three dimensional space we seem to inhabit, is that they happen to be too close together. It is tempting to invert this intuition and assert that two points θ and $\theta + d\theta$ are *close* together whenever they are difficult to distinguish. Furthermore, being a variance, the quantity $d\ell^2 = \langle \Delta^2 \rangle$ is positive and vanishes only when $d\theta$ vanishes. Thus, it is natural to introduce distance by interpreting g_{ab} as the metric tensor of a Riemannian space. [8] This is the *information metric*. The recognition by Rao that g_{ab} is a metric in the space of probability distributions gave rise to the subject of information geometry [3], namely, the application of geometrical methods to problems in inference and in information theory.

The coordinates θ are quite arbitrary; one can freely relabel the points in the manifold. It is then easy to check that g_{ab} are the components of a tensor and that the distance $d\ell^2$ is an invariant, a scalar under coordinate transformations. Indeed, the transformation

$$\theta^{a'} = f^{a'}(\theta^1 \ldots \theta^n) \tag{27}$$

leads to

$$d\theta^a = \frac{\partial \theta^a}{\partial \theta^{a'}} d\theta^{a'} \quad \text{and} \quad \frac{\partial}{\partial \theta^a} = \frac{\partial \theta^{a'}}{\partial \theta^a} \frac{\partial}{\partial \theta^{a'}} \tag{28}$$

so that, substituting into eq.(25),

$$g_{ab} = \frac{\partial \theta^{a'}}{\partial \theta^a} \frac{\partial \theta^{b'}}{\partial \theta^b} g_{a'b'} \tag{29}$$

Derivation from relative entropy

Elsewhere we argued for the concept of relative entropy $S[p,q]$ as a tool for updating probabilities from a prior q to a posterior p when new information in the form of constraints becomes available. (For a detailed development of the Method of Maximum Entropy see [5] and references therein.) The idea is to use $S[p,q]$ to rank those distributions p relative to q so that the preferred posterior is that which maximizes $S[p,q]$ subject to the constraints. The functional form of $S[p,q]$ is derived from very conservative design criteria that recognize the value of information: what has been learned in the past is valuable and should not be disregarded unless rendered obsolete by new information. This is expressed as a *Principle of Minimal Updating*: beliefs should be revised only to the extent required by the new evidence. According to this interpretation those distributions p that have higher entropy $S[p,q]$ are *closer* to q in the sense that they reflect a less drastic revision of our beliefs.

The term 'closer' is very suggestive but it can also be dangerously misleading. On one hand, it suggests there is a connection between entropy and geometry. As shown below, such a connection does, indeed, exist. On the other hand, it might tempt us to

identify $S[p,q]$ with distance which is, obviously, incorrect: $S[p,q]$ is not symmetric, $S[p,q] \neq S[q,p]$, and therefore it cannot be a distance. There is a relation between entropy and distance but the relation is not one of identity.

In curved spaces the distance between two points p and q is the length of the shortest curve that joins them and the length ℓ of a curve, eq.(12), is the sum of *local* infinitesimal lengths $d\ell$ lying between p and q. On the other hand, the entropy $S[p,q]$ is a *non-local* concept. It makes no reference to any points other than p and q. Thus, the relation between entropy and distance, if there is any all, must be a relation between two infinitesimally close distributions q and $p = q + dq$. Only in this way can we define a distance without referring to points between p and q. (See also [11].)

Consider the entropy of one distribution $p(x|\theta')$ relative to another $p(x|\theta)$,

$$S(\theta',\theta) = -\int dx\, p(x|\theta') \log \frac{p(x|\theta')}{p(x|\theta)} \ . \tag{30}$$

We study how this entropy varies when $\theta' = \theta + d\theta$ is in the close vicinity of a given θ. It is easy to check – recall the Gibbs inequality, $S(\theta',\theta) \leq 0$, with equality if and only if $\theta' = \theta$ — that the entropy $S(\theta',\theta)$ attains an absolute maximum at $\theta' = \theta$. Therefore, the first nonvanishing term in the Taylor expansion about θ is second order in $d\theta$

$$S(\theta + d\theta, \theta) = \frac{1}{2} \frac{\partial^2 S(\theta',\theta)}{\partial \theta'^a \partial \theta'^b}\bigg|_{\theta'=\theta} d\theta^a d\theta^b + \ldots \leq 0 \ , \tag{31}$$

which suggests defining a distance $d\ell$ by

$$S(\theta + d\theta, \theta) = -\frac{1}{2} d\ell^2 \ . \tag{32}$$

A straightforward calculation of the second derivative gives the information metric,

$$-\frac{\partial S(\theta',\theta)}{\partial \theta'^a \partial \theta'^b}\bigg|_{\theta'=\theta} = \int dx\, p(x|\theta) \frac{\partial \log p(x|\theta)}{\partial \theta^a} \frac{\partial \log p(x|\theta)}{\partial \theta^b} = g_{ab} \ . \tag{33}$$

UNIQUENESS OF THE INFORMATION METRIC

A most remarkable fact about the information metric is that it is essentially unique: except for a constant scale factor it is the only Riemannian metric that adequately takes into account the nature of the points of a statistical manifold, namely, that these points represent probability distributions, that they are not "structureless". This theorem was first proved by N. Čencov within the framework of category theory [2]; later Campbell gave an alternative proof that relies on the notion of Markov mappings. [12] Here I will describe Campbell's basic idea in the context of a simple example.

We can use binomial distributions to analyze the tossing of a coin (with probabilites $p(\text{heads}) = \theta$ and $p(\text{tails}) = 1 - \theta$). We can also use binomials to describe the throwing of a special die. For example, suppose that the die is loaded with equal probabilities for three faces, $p_1 = p_2 = p_3 = \theta/3$, and equal probabilities for the other three faces,

$p_4 = p_5 = p_6 = (1-\theta)/3$. Then we use a binomial distribution to describe the coarse outcomes low $= \{1,2,3\}$ or high $= \{4,5,6\}$ with probabilities θ and $1-\theta$. This amounts to mapping the space of coin distributions to a subspace of the space of die distributions.

The embedding of the statistical manifold of $n = 2$ binomials, which is a simplex $\mathscr{S}_1$ of dimension one, into a subspace of the statistical manifold of $n = 6$ multinomials, which is a simplex $\mathscr{S}_5$ of dimension five, is called a Markov mapping.

Having introduced the notion of Markov mappings we can now state the basic idea behind Campbell's argument: whether we talk about heads/tails outcomes in coins or we talk about low/high outcomes in dice, binomials are binomials. Whatever geometrical relations are assigned to distributions in $\mathscr{S}_1$, exactly the same geometrical relations should be assigned to the distributions in the corresponding subspace of $\mathscr{S}_5$. Therefore, these Markov mappings are not just embeddings, they are congruent embeddings — distances between distributions in $\mathscr{S}_1$ should match the distances between the corresponding images in $\mathscr{S}_5$.

Now for the punch line: the goal is to find the Riemannian metrics that are invariant under Markov mappings. It is easy to see why imposing such invariance is extremely restrictive: The fact that distances computed in $\mathscr{S}_1$ must agree with distances computed in subspaces of $\mathscr{S}_5$ introduces a constraint on the allowed metric tensors; but we can always embed $\mathscr{S}_1$ and $\mathscr{S}_5$ in spaces of larger and larger dimension which leads to more and more constraints. It could very well have happened that no Riemannian metric survives such restrictive conditions; it is quite remarkable that some do survive and it is even more remarkable that (up to an uninteresting scale factor) the surviving Riemannian metric is unique. Details of the proof are given in [5].

THE METRIC FOR SOME COMMON DISTRIBUTIONS

The statistical manifold of **multinomial distributions**,

$$P_N(n|\theta) = \frac{N!}{n_1!\dots n_m!}\theta_1^{n_1}\dots\theta_m^{n_m}\,, \tag{34}$$

where

$$n = (n_1\dots n_m) \quad \text{with} \quad \sum_{i=1}^{m} n_i = N \quad \text{and} \quad \sum_{i=1}^{m}\theta_i = 1\,, \tag{35}$$

is the simplex $\mathscr{S}_{m-1}$. The metric is given by eq.(25),

$$g_{ij} = \sum_n P_N \frac{\partial \log P_N}{\partial\theta_i}\frac{\partial \log P_N}{\partial\theta_j} \quad \text{where} \quad 1 \le i,j \le m-1\,. \tag{36}$$

The result is

$$g_{ij} = \left\langle (\frac{n_i}{\theta_i} - \frac{n_m}{\theta_m})(\frac{n_j}{\theta_j} - \frac{n_m}{\theta_m}) \right\rangle = \frac{N}{\theta_i}\delta_{ij} + \frac{N}{\theta_m}\,, \tag{37}$$

where $1 \le i,\, j \le m-1$. A somewhat simpler expression can be obtained writing $d\theta_m = -\sum_{i=1}^{m-1} d\theta_i$ and extending the range of the indices to include $i,\, j = m$. The result is

$$d\ell^2 = \sum_{i,j=1}^{m} g_{ij}\, d\theta_i d\theta_j \quad \text{with} \quad g_{ij} = \frac{N}{\theta_i}\delta_{ij}\,. \tag{38}$$

A uniform distribution over the simplex $\mathscr{S}_{m-1}$ assigns equal probabilities to equal volumes,

$$P(\theta)d^{m-1}\theta \propto g^{1/2}d^{m-1}\theta \quad \text{with} \quad g = \frac{N^{m-1}}{\theta_1\theta_2\ldots\theta_m} \tag{39}$$

In the particular case of binomial distributions $m = 2$ with $\theta_1 = \theta$ and $\theta_2 = 1-\theta$ we get

$$g = g_{11} = \frac{N}{\theta(1-\theta)} \tag{40}$$

so that the uniform distribution over θ (with $0 < \theta < 1$) is

$$P(\theta)d\theta \propto [\frac{N}{\theta(1-\theta)}]^{1/2}d\theta \ . \tag{41}$$

Canonical distributions: Let z denote the microstates of a system (*e.g.*, points in phase space) and let $m(z)$ be the underlying measure (*e.g.*, a uniform density on phase space). The space of macrostates is a statistical manifold: each macrostate is a canonical distribution obtained by maximizing entropy $S[p,m]$ subject to n constraints $\langle f^a\rangle = F^a$ for $a = 1\ldots n$, plus normalization,

$$p(z|F) = \frac{1}{Z(\lambda)}m(z)e^{-\lambda_a f^a(z)} \quad \text{where} \quad Z(\lambda) = \int dz\, m(z)e^{-\lambda_a f^a(z)} \ . \tag{42}$$

The set of numbers $F = (F^1\ldots F^n)$ determines one point $p(z|F)$ on the statistical manifold so we can use the F^a as coordinates.

First, here are some useful facts about canonical distributions. The Lagrange multipliers λ_a are implicitly determined by

$$\langle f^a\rangle = F^a = -\frac{\partial \log Z}{\partial \lambda_a}, \tag{43}$$

and it is straightforward to show that a further derivative with respect to λ_b yields the covariance matrix,

$$C^{ab} \equiv \langle (f^a - F^a)(f^b - F^b)\rangle = -\frac{\partial F^a}{\partial \lambda_b} \ . \tag{44}$$

Furthermore, from the chain rule

$$\delta^c_a = \frac{\partial \lambda_a}{\partial \lambda_c} = \frac{\partial \lambda_a}{\partial F^b}\frac{\partial F^b}{\partial \lambda_c} \ , \tag{45}$$

it follows that the matrix

$$C_{ab} = -\frac{\partial \lambda_a}{\partial F^b} \tag{46}$$

is the inverse of the covariance matrix, $C_{ab}C^{bc} = \delta^c_a$.

The information metric is

$$g_{ab} = \int dz\, p(z|F)\, \frac{\partial \log p(z|F)}{\partial F^a}\, \frac{\partial \log p(z|F)}{\partial F^b}$$
$$= \frac{\partial \lambda_c}{\partial F^a}\frac{\partial \lambda_d}{\partial F^b} \int dz\, p\, \frac{\partial \log p}{\partial \lambda_c}\, \frac{\partial \log p}{\partial \lambda_d}\ . \tag{47}$$

Using eqs.(42) and (43),

$$\frac{\partial \log p(z|F)}{\partial \lambda_c} = F^c - f^c(z) \tag{48}$$

therefore,

$$g_{ab} = C_{ca} C_{db} C^{cd} \implies g_{ab} = C_{ab}\ , \tag{49}$$

so that the metric tensor g_{ab} is the inverse of the covariance matrix C^{ab}.

Instead of the expected values F^a we could have used the Lagrange multipliers λ_a as coordinates. Then the information metric is the covariance matrix,

$$g^{ab} = \int dz\, p(z|\lambda)\, \frac{\partial \log p(z|\lambda)}{\partial \lambda_a}\, \frac{\partial \log p(z|\lambda)}{\partial \lambda_b} = C^{ab}\ . \tag{50}$$

Therefore the distance $d\ell$ between neighboring distributions can written in either of two equivalent forms,

$$d\ell^2 = g_{ab} dF^a dF^b = g^{ab} d\lambda_a d\lambda_b\ . \tag{51}$$

The uniform distribution over the space of macrostates assigns equal probabilities to equal volumes,

$$P(F) d^n F \propto C^{-1/2} d^n F \quad \text{or} \quad P'(\lambda) d^n \lambda \propto C^{1/2} d^n \lambda\ , \tag{52}$$

where $C = \det C^{ab}$.

Gaussian distributions are a special case of canonical distributions — they maximize entropy subject to constraints on mean values and correlations. Consider Gaussian distributions in D dimensions,

$$p(x|\mu, C) = \frac{c^{1/2}}{(2\pi)^{D/2}} \exp\left[-\frac{1}{2} C_{ij} (x^i - \mu^i)(x^j - \mu^j) \right]\ , \tag{53}$$

where $1 \le i \le D$, C_{ij} is the inverse of the correlation matrix, and $c = \det C_{ij}$. The mean values μ^i are D parameters μ^i, while the symmetric C_{ij} matrix is an additional $\frac{1}{2}D(D+1)$ parameters. Thus, the dimension of the statistical manifold is $\frac{1}{2}D(D+3)$.

Calculating the information distance between $p(x|\mu, C)$ and $p(x|\mu + d\mu, C + dC)$ is a matter of keeping track of all the indices involved. Skipping all details, the result is

$$d\ell^2 = g_{ij} d\mu^i d\mu^j + g^{ij}_k dC_{ij} d\mu^k + g^{ijkl} dC_{ij} dC_{kl}\ , \tag{54}$$

where

$$g_{ij} = C_{ij}\,, \quad g^{ij}_k = 0\,, \quad \text{and} \quad g^{ijkl} = \frac{1}{4}(C^{ik}C^{jl} + C^{il}C^{jk})\ , \tag{55}$$

where C^{ik} is the correlation matrix, that is, $C^{ik}C_{kj} = \delta^i_j$. Therefore,

$$d\ell^2 = C_{ij}dx^i dx^j + \frac{1}{2}C^{ik}C^{jl}dC_{ij}dC_{kl} \ . \tag{56}$$

To conclude we consider a couple of special cases. For Gaussians that differ only in their means the information distance between $p(x|\mu,C)$ and $p(x|\mu+d\mu,C)$ is obtained setting $dC_{ij} = 0$, that is,

$$d\ell^2 = C_{ij}dx^i dx^j \ , \tag{57}$$

which is an instance of eq.(49). Finally, for spherically symmetric Gaussians,

$$p(x|\mu,\sigma) = \frac{1}{(2\pi\sigma^2)^{D/2}} \exp\left[-\frac{1}{2\sigma^2}\delta_{ij}(x^i-\mu^i)(x^j-\mu^j)\right] \ . \tag{58}$$

The covariance matrix and its inverse are both diagonal and proportional to the unit matrix,

$$C_{ij} = \frac{1}{\sigma^2}\delta_{ij} \ , \quad C^{ij} = \sigma^2\delta^{ij} \ , \quad \text{and} \quad c = \sigma^{-2D} \ . \tag{59}$$

Substituting

$$dC_{ij} = d\frac{1}{\sigma^2}\delta_{ij} = -\frac{2\delta_{ij}}{\sigma^3}d\sigma \tag{60}$$

into eq.(56), the induced information metric is

$$d\ell^2 = \frac{1}{\sigma^2}\delta_{ij}d\mu^i d\mu^j + \frac{1}{2}\sigma^4\delta^{ik}\delta^{jl}\frac{2\delta_{ij}}{\sigma^3}d\sigma\frac{2\delta_{kl}}{\sigma^3}d\sigma \tag{61}$$

which, using

$$\delta^{ik}\delta^{jl}\delta_{ij}\delta_{kl} = \delta^k_j\delta^j_k = \delta^k_k = D \ , \tag{62}$$

simplifies to

$$d\ell^2 = \frac{\delta_{ij}}{\sigma^2}d\mu^i d\mu^j + \frac{2D}{\sigma^2}(d\sigma)^2 \ . \tag{63}$$

CONCLUSION

With the definition of the information metric we have only scratched the surface. Not only can we introduce lengths and volumes but we can make use of all sorts of other geometrical concepts such geodesics, normal projections, notions of parallel transport, covariant derivatives, connections, and curvature. The power of the methods of information geometry is demonstrated by the vast number of applications. For a very incomplete point of entry to the enormous literature in mathematical statistics see [4][13][14][15]; in model selection [16][17]; in thermodynamics [18]; and for the extension to a quantum information geometry see [19][20].

The ultimate range of these methods remains to be explored. In this tutorial we have argued that information geometry is a natural and inevitable tool for reasoning with

incomplete information. One may perhaps conjecture that to the extent that science consists of reasoning with incomplete information, then we should expect to find probability, and entropy, and also geometry in all aspects of science. Indeed, I would even venture to predict that once we understand better the physics of space and time we will find that even that old and familiar first geometry — Euclid's geometry for physical space — will turn out to be a manifestation of information geometry. But that is work for the future.

REFERENCES

1. A. Einstein, p. 67 in "*Albert Einstein: Philosopher-Scientist*", ed. by P. A. Schilpp (Open Court 1969).
2. N. N. Čencov: *Statistical Decision Rules and Optimal Inference*, Transl. Math. Monographs, vol. 53, Am. Math. Soc. (Providence, 1981).
3. S. Amari, *Differential-Geometrical Methods in Statistics* (Springer-Verlag, 1985).
4. S. Amari and H. Nagaoka, *Methods of Information Geometry* (Am. Math. Soc./Oxford U. Press, 2000).
5. A. Caticha, *Entropic Inference and the Foundations of Physics* (USP Press, São Paulo, Brazil 2012); online at http://www.albany.edu/physics/ACaticha-EIFP-book.pdf.
6. W. K. Wootters, "Statistical distance and Hilbert space", Phys. Rev. **D**, 357 (1981).
7. V. Balasubramanian, "Statistical inference, Occam's razor, and statistical mechanics on the space of probability distributions", Neural Computation **9**, 349 (1997).
8. C. R. Rao, "Information and the accuracy attainable in the estimation of statistical parameters", Bull. Calcutta Math. Soc. **37**, 81 (1945).
9. C. Atkinson and A. F. S. Mitchell, "Rao's distance measure", Sankhyā **43**A, 345 (1981).
10. R. A. Fisher, "Theory of statistical estimation", Proc. Cambridge Philos. Soc. **122**, 700 (1925).
11. C. C. Rodríguez, "The metrics generated by the Kullback number", *Maximum Entropy and Bayesian Methods*, J. Skilling (ed.) (Kluwer, Dordrecht 1989).
12. L. L. Campbell, "An extended Čencov characterization of the information metric", Proc. Am. Math. Soc. **98**, 135 (1986).
13. B. Efron, Ann. Stat. **3**, 1189 (1975).
14. C. C. Rodríguez, "Entropic priors", *Maximum Entropy and Bayesian Methods*, edited by W. T. Grandy Jr. and L. H. Schick (Kluwer, Dordrecht 1991).
15. R. A. Kass and P. W. Vos, *Geometric Foundations of Asymptotic Inference* (Wiley, 1997).
16. J. Myung, V. Balasubramanian, and M.A. Pitt, Proc. Nat. Acad. Sci. **97**, 11170 (2000).
17. C. C. Rodríguez, "The ABC of model selection: AIC, BIC and the new CIC", *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, ed. by K. Knuth *et al.*, AIP Conf. Proc. Vol. **803**, 80 (2006) (omega.albany.edu:8008/CIC/me05.pdf).
18. G. Ruppeiner, Rev. Mod. Phys. **67**, 605 (1995).
19. R. Balian, Y. Alhassid and H. Reinhardt, Phys Rep., **131**, 2 (1986).
20. R. F. Streater, Rep. Math. Phys., **38**, 419-436 (1996).