**FROM EUCLID TO ENTROPY**

C. C. Rodríguez
Department of Mathematics and Statistics
SUNY at Albany
Albany, NY 12222 USA

ABSTRACT. This paper is about non-euclidean geometry in the theory of inference.

## 1. Introduction

Inference and Geometry get along so well that their offspring appear much before the union is formalized. It suffices to think of the odds ratio of Bayesian inference as the cross-ratio of projective geometry.

At this general level the statement is no more than an informal analogy. However, as I argue below, it is pregnant with true implications and suggestive interpretations. I guess that a similar conclusion will be inescapable to anyone reading side by side Courant and Robbins, 1941 (chap. IV) along with the first pages of *Information Theory and Statistics* Kullback, 1959. With this connection in mind let me re-create the introduction to the standard stories of geometry and inference.

## 2. The Invention of the New Geometry

The fable of geometry goes something like this... Once upon a time there was a greek man named Euclid who organized the mathematical knowledge in his time in six books known as The Elements. Euclid's work was so influential that it shaped intelectual creativity in Europe for over twenty two centuries.

In the Elements, Eulcid attempted to derive all the statements of elementary geometry as provable theorems in a system of a few axioms. But according to historians of mathematics: "Euclid's title to immortality is based on something quite other than the supposed logical perfection which is still sometimes erroneously ascribed to him. This is *his* recognition that the fifth of his postulates (his Axiom XI) is a pure assumption" (Bell, 1937 p.299).

The fifth postulate of Euclid asserts that through a point $P$ not on a line $l$ passes exactly one line $l'$ parallel to $l$. This and equivalent statements were the source of a 2,200 year long head ache for mathematics.

Based on the a priori assumption that Euclid's fifth postulate was at a much higher level of complexity than the others, generations of mathematicians tried, without success, to derive it from the other euclidean postulates. But the identification of euclidean geometry with physical reality was so deeply rooted in the conceptual system of the past that it took the tremendous intelectual courage (or naiveté ) of Lobachevsky to stand firm to an alternative of the Euclidean system. In fact, the *discovery* of the non-euclidean plane is regarded as one of the greatest liberations of the human mind. It forever opened the possibility of imagining, in a logically consistent way, other (non-euclidean) realities. For a graphic analogy: the non-euclidean plane is to mathematics as *Les Demoiselles d'Avignon* is to art.

It is possible to visualize the non-euclidean plane as the set of interior points of a circle. This is known as Klein's model of the Lobachevskian plane. By calling (non-euclidean) lines the chords of the circle and (non-euclidean) displacements the projective transformations of the plane that map the circle into itself, we obtain a system of *points* and *lines* that satisfy all the postulates of Euclid except the fifth. Now, instead of only one *parallel* we have an infinite number of them!

The violation of the euclidean parallel postulate is due to a change in the concept of distance. The rigid motions are no longer rotations and translations i.e. the group of transformations preserving the euclidean metric. The rigid displacements are now projective transformations and therefore the distance must be a function of a quantity invariant under projections; for non-euclidean rigid motions should leave non-euclidean distances invariant. The invariant quantity is the cross-ratio of projective geometry. It is a number associated to any four collinear points but here we only need a special case. Condider two arbitrary (non-euclidean) points $H_0$ and $H_1$ inside the circle and the segment $H_0H_1$ joining them. Prolonging the segment $H_0H_1$ we obtain the chord (non-euclidean line) $OH_0H_1D$ where $O$ and $D$ are points on the circumference. The cross-ratio of the points $O, H_0, H_1, D$ is given by

$$[O, H_0, H_1, D] = \frac{H_0D/H_1D}{H_0O/H_1O} \tag{1}$$

It follows from (1) that if $H_0, H_1$ and $H_2$ are on a line then,

$$[O, H_0, H_1, D] \cdot [O, H_1, H_2, D] = [O, H_0, H_2, D]. \tag{2}$$

Equation (2) shows that in order to preserve additivity we must define the non-euclidean distance in this model as

$$\overline{H_0H_1} = \log[O, H_0, H_1, D]. \tag{3}$$

Notice that (3) is never negative (since (1) is never less than 1) and the chords (e.g. $OD$) have infinite non-euclidean length, as they should. Let us leave geometry at this point and turn the attention to the theory of inference.

## 3. The Theorem Between Theories and Things

The subject of inference deals fundamentally with three kinds of propositions:

(i) Data. Denoted by $D$ (the facts or *things*).
(ii) Hypotheses. Denoted by $H, H_0, H_1, \ldots$ etc (the *theories*).
(iii) Prior information. Denoted by $O$ (*everything else*).

There is now substantial evidence (see the preliminary chapters of E.T. Jaynes' forthcomming book) that if we want real numbers to represent partial truth then these numbers should satisfy the product and sum rules of probability theory. Thus, if $P(A|B)$ denotes the partial truth of proposition A given proposition B then

$$P(AB|C) = P(A|BC) \cdot P(B|C) \tag{4}$$

and

$$P(A|B) + P(\overline{A}|B) = 1 \tag{5}$$

where $\overline{A}$ denotes negation of A and AB denotes the conjunction of propositions A and B. From the commutativity of conjunctions we have

$$P(HD|O) = P(DH|O) \tag{6}$$

and applying the product rule (4) on both sides of (6) we obtain the familiar Bayes theorem,

$$P(H|DO) = \frac{P(D|HO) \cdot P(H|O)}{P(D|O)}. \tag{7}$$

It is difficult to overemphasize the importance of this simple relation but we may try: *He first created Bayes Theorem and then He drew his pistol!* But who knows; may be a gigantic application of (7) turns out to be the key to Quantum Cosmology after all (see Hawking, 1986 for inspiration).

Theories are constructed in such a way that the value of the likelihood $P(D|HO)$ is completely specified. Unfortunately the initial probabilities $P(H|O)$ and $P(D|O)$ are often unknown. They have to be estimated from the maximum entropy principle of Jaynes, the axiomatics of Skilling, the invariance arguments of Jeffreys, intuition, ad-hockeries, etc. However, if all we want is to compare two competing hypotheses $H_0$ and $H_1$ we may do so by computing the numeric value of their likelihood ratio. Bayes theorem tells us just what this number means

$$\frac{P(D|H_0O)}{P(D|H_1O)} = \frac{P(H_0|DO)/P(H_1|DO)}{P(H_0|O)/P(H_1|O)}. \tag{8}$$

There is a remarkable similarity between the right hand side of (8) and the right hand side of (1). To obtain the same formula we need only to use the logical equation $DO = D$ and replace the euclidean separation $AB$ in (1) with the logical proximity $P(A|B)$ of proposition $A$ from proposition $B$. Thus, from (3) it is very natural to interpret the log likelihood ratio as measuring the separation of the hypothesis $H_1$ from the hypothesis $H_0$ given by the observation of $D$. Different bases for the logarithm provide different distance scales. Two useful scales are *bits* and *decibels* obtained with base 2 and base $10^{0.1} \simeq 1.259$. However, unlike (3) the log likelihood ratio is not positive nor symmetric. Positivity is recovered if we replace the log likelihood ratio by its expected value under $H_0$ i.e. the Kullback number between $H_0$ and $H_1$ given by

$$I(H_0 : H_1) = \sum_D P(D|H_0) \log \frac{P(D|H_0)}{P(D|H_1)}. \tag{9}$$

Hence, we arrive to where Kullback started: $I(H_0 : H_1)$ is the mean (non-euclidean-like) distance of $H_1$ from $H_0$.

The strict concavity of the logarithm implies that (9) is never negative and that it is zero only when $P(D|H_0) = P(D|H_1)$ for all $D$ possible under $H_0$ i.e. when $H_1$ is at distance zero from $H_0$. This makes the Kullback number very similar to a metric but it is still not symmetric and it doesn't satisfy the triangular inequality. But symmetry may not be desired in this context since logical proximities are clearly **not** symmetric: $P(A|B) \neq P(B|A)$.

## 4. The Geometrization of Inference

The previous intuitive analysis indicates that spaces of hypotheses have intrinsic geometries of possibly non-euclidean type. The language of differential geometry provides a rigorous formalization of this intuitive idea. Without going into the technicalities I summarize below some recent developments in this area.

Start by identifying hypotheses with probability measures over a fix measurable space. This is not ad-hoc. We are just defining explicitly what is meant by a hypothesis: *a label to a probability measure*. For if a given proposition $A$ is not able to specify the likelihood $P(D|A)$ it should not qualify as a well defined hypothesis. Moreover, if two hypotheses in the same space have the same prior probability and always produce the same likelihood then (according to (7)) they should be regarded as equivalent for inference purposes. Hence, hypotheses spaces are collections of probability measures. To evade complications we restrict the analysis to *dominated* sets of probability measures that can be *smoothly* parametrized with a finite number of parameters. With our definition a hypothesis could be a parameter, a density, a cumulative distribution, a characteristic function, or any other piece of information $H$ from where we can obtain the values of $P(D|H)$ for all $D$'s. When we replace in (9) $H_0$ and $H_1$ with probability measures $P$ and $Q$ we write

$$I(P:Q) = \int P(dD) \log \frac{dP}{dQ}(D). \tag{10}$$

To fix the notation consider a general hypotheses space containing the probability measures $P_\theta$ with densities (with respect to a fix dominating measure) $p_\theta$ where $\theta$ ranges over a set of parameters $\Theta$. Under mild regularity conditions this is a Riemmanian manifold with metric tensor given by

$$g_{ij}(\theta) = -\int P_\theta(dD) \frac{\partial^2 \log p_\theta}{\partial\theta_i \partial\theta_j}(D) \tag{11}$$

known as Fisher's information matrix. The invariant measure in this metric space is then given by

$$\eta(d\theta) \propto g^{1/2} d\theta. \tag{12}$$

Where $g$ is the determinant of the matrix defined by (11). The invariant measure specifies the surface area of the model and therefore it plays the role of the Lebesgue measure on a flat space. Thus it is natural to call this reference measure the prior of total ignorance. It is equal to Jeffreys non informative prior.

Informative families of prior distributions can be obtained from similar geometric considerations. A theme started by Shore and Johnson, 1980 and extended by Skilling, 1988,1989 and Rodriguez, 1989, 1990. The idea behind these works is essentially the following: invariance under reparametrizations of the sample and parameter spaces plus the assumption of no a priori correlations restricts the form of the possible priors to

$$\pi(d\theta|\alpha, m) \propto \exp(-\alpha I(P_\theta : m)) g^{1/2} d\theta. \tag{13}$$

Hence, the prior probability density is **completely** specified up to the values of $\alpha$ and $m$. Where $m$ is an initial measure for the data (e.g. $P_{\theta_0}$) and $\alpha > 0$ corresponds to the *number of virtual observations* (or amount of prior information) supporting m. Notice that the total ignorance prior (12) is obtained from (13) in the limit of no prior information ($\alpha \to 0$) for m.

It is interesting to note that the three classic geometries of constant curvature: hyperbolic (Lobachevskian, negative curvature), elliptic (spherical, positive curvature), and euclidean (zero curvature) are connected to the classic hypotheses spaces of the one dimensional gaussians, the discrete distributions, and the multivariate gaussians with fix variance covariance matrix respectively. With the powerful machinery of differential geometry at hand the theory of inference is entering a new age of development.

## 5. If Physics is Geometry and Inference is Geometry.. Then Life is But a Dream

It is a remarkable irony, that nature often plays on us, that it finds itself best expressed in the new imaginary avenues previously opened by mathematics. The history of physics contains many of these *coincidences.* e.g. non-euclidean geometry made general relativity thinkable and Penrose non periodic tilings allowed to *see* the quasi-cristals to name only the classic and the recent examples. Some neo platonists (see Penrose, 1989 for a lucid popular exposition) feel entitled to revive from here the old metaphysical thesis that physical existence is the shadow of the imaginary platonic world of mathematics. But their real claim, I believe, hides behind their shadows. It is the wishful thinking that the imaginary has physical existence. No doubt a very reasuring thought for mathematicians who find themselves to be only one logical step behind the proof of the convergence of *their* theories to the ultimate truth i.e. to god's (platonic) heaven.

However, the same facts used by the platonists, can be used to undermine their transcendental expectations. That physics is often best expressed in the current mathematical language of the time does not show that we are making contact with god's platonic mind but with our own, or rather the one of our culture. Instead of making the imaginary real it makes the real more imaginary, uncertain, temporal and chaotic than the advocators of the platonic heaven would like us to believe.

These speculations may seem like a long unjustified digression from the subjects of inference and geometry but they are not. They touch a central point. That the geometrization of inference may have something new to say about the level of reality of our physical theories. I hope to be able to come back to this idea in the future with *proofs.* In the mean while you may feel entitled to believe that *"it is a safe rule to apply that, when a mathematical or philosophical author writes with misty profundity, he is talking nonsense"*- A. N. Whitehead (1911). Even though I agree with the spirit of this popular belief, I can't resist to add: e.g. Principia Mahtematica. A bit of nonsensical dadaism is probably healthier than its alternative.

## REFERENCES

Bell, E.T.: 1937, *Men of Mathematics.* Simon and Schuster, Inc., New York.

Courant, R. and Robbins, H.E.: 1941, *What is Mathematics?.* Oxford University Press, New York.

Hawking, S.W.: 1986, 'Lectures on Quantum Cosmology', in H.J. de Vega and N. Sánchez (eds.), *Field Theory, Quantum Gravity and Strings.* Lectures Notes in Physics, **246**. Springer-Verlag, Berlin.

Jaynes, E.T.: 1990, *Probability Theory- The Logic of Science.* Preliminary chapters in TEX files can be obtained from the author.

Kullback, S.: 1959, *Information Theory and Statistics.* John Wiley and Sons, Inc., New York.

Penrose, R.: 1989, *The Emperor's New Mind.* Oxford University Press, New York.

Rodríguez, C.C.: 1989, 'The Metrics Induced by the Kullback Number', in J. Skilling (ed.), *Maximum Entropy and Bayesian Methods*, Kluwer, Dordrecht.

Rodríguez, C.C.: 1990, 'Objective Bayesianism and Geometry', in P. Fougere (ed.), *Maximum Entropy and Bayesian Methods*, Kluwer, Dordrecht.

Shore, J.E. and Johnson, R.W.: 1980, 'Axiomatic Derivation of the Principle of Maximum Entropy and the Principle of Minimum Cross-entropy', *IEEE Trans. Info. Theory*, **IT-26**, 26-39 and **IT-29**, 942-943.

Skilling, J.: 1988, 'The Aximoms of Maximum Entropy', in G.J. Erikson and C.R. Smith (eds.), *Maximum Entropy and Bayesian Methods in Science and Engineering*, **1**, Kluwer, Dordrecht.

Skilling, J.: 1989, 'Classic Maximum Entropy', in J. Skilling (ed.), *Maximum Entropy and Bayesian Methods*, Kluwer, Dordrecht.