Geometric and Topological Invariants of the Hypothesis Space

Carlos C. Rodríguez

Department of Mathematics and Statistics
The University at Albany, SUNY.
http://omega.albany.edu:8008/

Abstract.
The form and shape of a hypothesis space imposes natural objective constraints to any inferential process. This contribution summarizes what is currently known and the mathematics that are thought to be needed for new developments in this area. For example, it is well known that the quality of best possible estimators deteriorates with increasing volume, dimension and curvature of the hypothesis space. It is also known that regular statistical parametric models are finite dimensional Riemannian manifolds admitting a family of dual affine connections. Fisher information is the metric induced on the hypothesis space by the Hellinger distance. Nonparametric models are infinite dimensional manifolds. Global negative curvature implies asymptotic inadmissibility of uniform priors. When there is uncertainty about the model and the prior, entropic methods are more robust than standard Bayesian inference. The presence of some types of singularities allow the existence of faster than normal estimators ..., etc. The large number of fundamental statistical concepts with geometric and topological content suggest to try to look at Riemannian Geometry, Algebraic Geometry, K-theory, Algebraic Topology, Knot-theory and other branches of current mathematics, not as empty esoteric abstractions but as allies for statistical inference.

Keywords: Bayesian Inference, Information Geometry, Ignorance Priors, Empirical Bayes, Geometric Theory of Ignorance, Statistical Manifold
PACS: 02.40.-k,02.50.Tt

INTRODUCTION

Why there seems to be geometry everywhere?
I say it's Inference 'cause Inference is what brains do.

In the beginning was Eve, naked!

I have a brain
I obs. x
I want to understand x
I want to predict future x'
Problem: How?

You may complain that this is *just* talking, and I reply that you may have a point and that I plan to talk about that later but for now let's look at the naked facts: x and x'. Wait, did I say naked? oops! I am sorry, please scratch that! no, *there are no naked facts*, there is no

inmaculate *obs.*, there is no theoretical vacuum. Data and theory are always entangled. The reason is categorically trivial. In order for x to be the argument of a probability function it must be a logical proposition in a given domain of discourse. It must have *meaning* and by meaning I mean theory. Concretely, a theory θ is an explanation for the data x. A theory is a code for compressing x. The parameter θ is only a label, a name for a complete function; the probability distribution for x fixing the numerical values of $p_\theta(x) = p(x|\theta)$ for all the possible observations x. My trivial starting point is that there is never a clear demarcation between data and theory because there cannot be any! It is as if data and theory were the proverbial two sides of a single θx coin. The two sides are not symmetric though. The coin only shows its x side. The θ side is unobservable, impalpable. It is hidden but it is always there with the x. With all x and x'.

Smells like QM? Well, not just QM. Please, read again the previous lines. The logical impossibility of disentangling data from theory changes the book of science at the genesis.

The Gospel According to Herb

Herb Robbins tried to tell me (and the rest of us) something like this many times [1]. But I never *really* listened until a couple of years ago!.. Let me reenact one of Herb's favorite plays: Suppose $(\theta_1,x_1),(\theta_2,x_2),\ldots,(\theta_n,x_n)$ are independent samples from a random vector (θ,x) with joint distribution fixed by requiring that $x|\theta \sim \text{Poisson}(\theta)$ and $\theta \sim \text{G}$ where G is some unknown distribution. i.e., conditionally on θ, x is Poisson with mean θ and θ follows some fix but unknown distribution G. Think about car insurance and interpret (θ_i,x_i) as the impalpable accident proneness θ_i for the i-th observed driver together with her (!) last year record of observed number of accidents x_i. It is easy to check that these distributional assumptions are equivalent to assuming that $\theta_1,\ldots,\theta_n$ are iid with common distribution G and that for $i = 1,2,\ldots n$, we have $x_i|\theta_1,\ldots,\theta_n$ i.e. x_i conditionally on all the thetas is Poisson with mean θ_i. It seems obviously clear that, since the vectors (θ_i,x_i) are assumed independent for different values of the index i, that only x_i can tell us something about θ_i and that the best guess for θ_i must be x_i. But contrary to naive intuition it pays off to combine the assumed independent vectors to gather the information about the unknown G that is indirectly contained in the record of observations $x^n = (x_1,\ldots,x_n)$. Under the assumed joint distribution for (θ,x) the function[2] $u = u(x)$ that minimizes the quadratic loss $L(u) = E(u(x)-\theta)^2$ is the posterior mean $u_G(x) = E(\theta|x)$ which is a functional of G that expands to,

$$u_G(x) = \frac{\int \theta f(x|\theta)\, dG(\theta)}{\int f(x|\theta)\, dG(\theta)}$$

---

[1] In the *special olympics* of Robbins Seminar, I have the world record of attending weekly for more than ten years!

[2] Robbins preferred the notation $t(x)$ instead of $u(x)$ but after Vijay Balasabrumanian's paper I reserve t for truth.

Had G been known, the Bayes risk $L(u_G)$ would be the best anyone could hope for. However we do not know G we only know the records x^n. The little miracle is that for large n the data x^n is enough! The Poisson kernel,

$$f(x|\theta) = e^{-\theta}\frac{\theta^x}{x!}$$

behaves nicely and allows us to write,

$$u_G(x) = \frac{(x+1)f_G(x+1)}{f_G(x)}$$

in terms of the marginal distribution of x, given by $f_G(x) = \int f(x|\theta)\,dG(\theta)$. We are in luck since the iid assumption for the vectors (θ_i, x_i) for $i = 1,\ldots,n$ implies that $x_1, x_2, \ldots, x_n$ are iid from f_G and by the classic law of large numbers the empirical frequencies,

$$f_n(x) = \frac{\text{number of } x_i = x}{n}$$

will go without fuss a.s. to the unknown probabilities $f_G(x)$ which in turn will make $u_n(x) = (x+1)f_n(x+1)/f_n(x)$ merrily go along a.s. to $u_G(x)$, the best anyone could theoretically hope for. In fact, in the 1950's M.V. Johns, Jr. wrote a Ph.D. thesis under Robbins dotting all the i's showing that in this case, if $\int \theta^2\,dG(\theta) < \infty$ as n increases, $(u_n(x_1),\ldots,u_n(x_n))$ estimates the impalpables $(\theta_1,\ldots,\theta_n)$ as well as if G was actually known! Now recall that the θ_i are iid from G. Thus, the $u_n(x_i)$ allow to estimate G(θ) empirically by the proportion of $u_n(x_i) \le \theta$ and by the classic Glivenko-Cantelli theorem we have a strong uniform approximation of the whole impalpable G. This, often made Robbins howl to the faithful subjectivists: *you can't learn from data! I can!*. Indeed, who cares what you, I, them or anyone subjectively thinks about the prior distribution G if the observations can actually estimate it!! I know that Herb died beleiving that in the future every rational being would be an Empirical Bayesian. He howled that to us many times in his lectures. We are now in 2010, officially in the future, and that does not seem to be the case. Why not?

Progress Report

That was the good news from Empirical Bayes (E.B.). Bayesianism without cheating. The parameter θ is assumed to have a distribution G but this G is allowed to be anything as long as $\int \theta^2\,dG(\theta) < \infty$. That's an example of an objective notion of ignorance. The problem is that allowing G to be anything doesn't always work in practice. For many problems assuming an ad-hoc parametric class for G turns out to be more efficient. But when the parameters of G are directly estimated from the data (which is the E.B. recipe) there is often observed overfitting resembling the typical problems of maximum likelihood estimators. For this reason, plain E.B. is dismissed by the faithful bayesians as an old fashion remnant from the frequentist era used by timids unable to ad-hoc-ly code hyper or hyper-hyper prios into their MCMC procedures. More over, the label *Empirical*

Bayes is deprecated and *Hierarchical Bayes* is recomended instead, presumably because Empirical Bayes is linguistically appealing and it threatens to expose the unjustified a priori beliefs of the non-empirical faithful bayesians.

Back to Empirical Bayes with Maximum Entropy

During the past ten to fifteen years a quiet evolution has been underway. There is now a new understanding of some old problems of inference. We now have *A Geometric Theory of Ignorance* standing on solid mathematical grounds that produces, priors, posteriors and likelihoods and a unification of statistical inference based on honesty as the only principle. It boils down to the well known Jaynesian slogan of *maximum entropy subject to a constraint* in its equivalent potent form of, *maximally noncommittal w.r.t. the missing information* or in my own terms: *maximum honesty*, i.e.,

maximum ignorance subject to whatever is known

Of course these aphorisms are only motivational devices that need to be properly spelled out in order to be useful. But once you understand their content, there is no aspect of statistical inference left untouched. I will make my point with a simple example.

Let $x_1, \ldots, x_n$ be a sample from a $N(\mu, \sigma^2)$ with $\theta = (\mu, \sigma) \in \Theta$. Everything else unknown. Estimate θ.

This problem may look ridiculously simple to you but I have been unable to find a satisfactory solution in the current literature. We are particularly interested in the case when $n = 1$ or $n = 2$. Of course, for n sufficiently large ($n > 3$ is usually enough) the solutions obtained by following the textbook recipes turn out to be adequate FAPP (for all practical purposes) only a little bit missleading. However, even for large values of n the discrepancies with the unique honest solution, even though they are small, they are still there and in principle there is still a problem.

Yes, there has been progress in this problem. However, before using the new ideas from the geometric theory of ignorance, I'd like to use this opportunity to mock the standard proposals by exposing their incorrect answers to this simple problem.

Case I: $\Theta = (0,10) \times (0.1, \infty)$**. i.e.,** $0 < \mu < 10$ **and** $\sigma > 0.1$ **and** $n = 1$ **with** $x_1 = x$
With everything else unknown, please estimate θ.

The cannon goes something like this. You are either a bayesian or a frequentist, you need to pick your religion. The frequentists will make you believe they are cooler because, they claim, they can solve the inference problems without the need of a prior distribution for the parameters. Just max the likelihood or the nonparametric likelihood or the partial likelihood or the pseudo-likelihood whichever works. They claim that diversity of procedures is a feature not a bug. They claim they are open minded and that they allow the data to speak for itself (or is it themselves?). For our simple problem if x is outside the range of μ, like $x = -20$, the MLE is $\hat{\theta} = (0, 20)$ and if instead of $x = -20$ you observe $0 < x < 10$ you'll find $\hat{\theta} = (x, 0.1)$. What's wrong with that? I say everything is wrong with that. It obviously overfits the data. When x is within the

range of μ the MLE picks x for the mean and the minimum possible value in the range of σ for the sd. Had the allowed minimum value for σ been $10^{-10^{10}}$ instead of just 10^{-1} the MLE would had been $(x, 10^{-10^{10}})$ instead. Likewise, when x is outside the range of μ the MLE picks the closest value from the range of μ and the distance from x to the range of μ as the sd. The MLE seems to follow the data blindly without using the actual information given in the problem. That's bad enough but what's even worse is that there is no measure of uncertainty attached to the point estimate in this case. We have no idea how good or how bad $\hat{\theta}$ is as an estimator. With only one observation and the variance unknown, the standard textbook recipe for computing the error bars for the confidence intervals breaks down. That's really bad news for the MLE since in this problem there is objective uncontroversial information about θ that goes beyond just the observation. I can show for example that $(|x-5|/102, 17|x-5|)$ has probability at least 95% of covering the true σ whatever the value of μ (see my *Confidence Intervals from One Observation*) without the need of any priors. The main reason we are given for using the MLE is that it is often consistent and second order efficient as $n \to \infty$. That's nice but hardly relevant here where $n = 1$! The other justification is that the MLE is easy to compute without the need of a prior or a loss function. Now that, I claim, is utterly incorrect. The MLE is the bayes estimator for 01-loss and a flat prior. Thus, the problem is not that the MLE does not need a prior. The problem is that the MLE uses the WRONG prior! [3].

So let's run out of the MLE temple and try our luck accross the street with the bayesian church... We are immediately welcome into the party of a priori freedom and the likelihood principle. We are quickly told to feed into holy bayes theorem the likelihood, prior, and data, and wait for the answers to our prayers to come out of the MCMC oracle. OK sounds good. Now how do we bayesians propose to solve the simple problem above? Well, we are told, we recommend independent priors for the individual components of $\theta = (\mu, \sigma)$. Also, it is ok to use improper (i.e. unnormalizable) priors as long as the posterior turns out proper. For location parameters like μ we recommend flat, prop to $d\mu$, and for scale parameters like σ we recommend Jeffreys prior prop to $d\sigma/\sigma$. In doubt go for proper but try to keep it as flat as possible to maximize "ignorance". For example student-t distributions that have heavy tails often work better than normal distributions and don't worry too much about the prior, if it matters then that's probably an indication that you need more data.

I don't know about you but to me all these semi-empirical authority based recipes for priors sound bogus. There should be a better way and in fact now there is, but let's check out the answers for our simple problem obtained by following the standard bayesian texbook recipes[4].

The standard Jeffreys prior recipe $p(\theta)\, d\theta \propto \sigma^{-1} d\mu\, d\sigma$ is improper for this problem since,

---

[3] follow the drama in my *Wrong Priors*

[4] Which textbook? It doesn't matter, no textbook that I know off can solve this problem without being **dishonest**, i.e., without assuming information that is **not** available.

$$\int_\Theta \frac{d\mu\, d\sigma}{\sigma} = 10 \int_{0.1}^{\infty} \frac{d\sigma}{\sigma} = \infty$$

here we enter shaky territory. An unnormalizable positive function does not become a probability density by the virtue of describing it with the LaTeX"\propto" $\propto$ symbol instead of with the $=$ sign no matter who's famous name from a famous university is attached to the famous recommendation. Interpreting the improper function as a limit of a proper normalizable sequence does not necessarily solve the problem either if the posterior is improper or if the limit depends on the sequence. There is also the popular claim that when the data comes from a normal distribution Jeffreys recipe $d\mu d\sigma/\sigma$ matches the answer from sampling theory. But what consolation is it to know that an answer coincides with another one known to be incorrect? The same criticism applies, to some extent, to all bayesian schemes attempting to match coverages obtainable from sampling theory using MLE.

But let's brush aside all criticisms and proceed formaly with the standard bayesian recipe. With a single observation x the posterior when $\theta \in \Theta$ becomes

$$p(\theta|x) \propto \frac{1}{\sigma}\varphi\left(\frac{\mu - x}{\sigma}\right) \cdot \frac{1}{\sigma}$$

where $\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$ is the standard density of the standard normal distribution. Integrating over the rage of σ we obtain,

$$\int_{0.1}^{\infty} p(\theta|x)\, d\sigma \propto \frac{\Phi(-5|\mu - x|)}{|\mu - x|}.$$

where Φ denotes the gaussian c.d.f.. When the observation is within the range of μ i.e., when $0 < x < 10$ then the posterior has a pole of order 1 at $\mu = x$ and the posterior is improper so the standard recipe breaks down in that case. When x lies outside the range of μ like $x = -20$ then the posterior is proper but still overfits and it is very different from the only honest answer for this problem.

The honest answer: The set of all gaussians labeled by $\theta \in \Theta$ constitute a Riemannian manifold of finite information volume (area in this case). The normalized volume form for this hypothesis space is given by,

$$dV = \frac{d\mu \wedge d\sigma}{\sigma^2}$$

Therefore, there exists a uniform distribution for this hypothesis space. If we chose the labels $\theta \in \Theta$ a priori according to the (non uniform) density $p(\theta) = 1/\sigma^2$ we'll be picking points (i.e. gaussians in this case) uniformly at random from all the available choices. This is the (corrected) Laplace Principle of Indiference. Nowadays, in 2010, this is a differential geometry triviality. However, it was unknown to Laplace and, I claim, it was also unknown to Jeffreys in 1939 [5]. With a single observation x the posterior

[5] I do not know who first thought about it but I arrived to this conlusion independently, in a more general context, at the end of my MaxEnt 1988 paper *The Metrics Induced by the Kullback Number*

from this proper prior is automatically proper. Integrating over σ we find the marginal posterior for μ given for $0 < \mu < 10$ by,

$$p(\mu|x) = \frac{1}{Z} \cdot \frac{1 - e^{-50(\mu - x)^2}}{(\mu - x)^2}$$

When $x = -20$ we obtain $Z = 1/60$ and the above density is (FAPP) numerically identical to $60(\mu + 20)^{-2}$. When $0 < x < 10$ the above marginal posterior for μ is very similar to a truncated Cauchy centered at x. For example when $x = 1$ the marginal posterior for μ is very close to a Cauchy located at 1 with scale 0.1616768 truncated to the range of $\mu \in [0, 10]$. Integrating the posterior over the range of μ we obtain the marginal posterior for σ given by,

$$p(\sigma|x) = \frac{1}{Z(x)} \left[\Phi\left(\frac{10 - x}{\sigma}\right) - \Phi\left(\frac{-x}{\sigma}\right) \right] \text{ for } \sigma > 0.1$$

For example when $x = 1$ this is monotonically decreasing with median ≈ 0.1915 and IQR $\approx (0.1314, 0.3531)$. It does not overfit.

Let us now consider a case when Θ has infinite information volume so that there is no uniform distribution over the hypothesis space.

Case II: $\Theta = (0, 10) \times (0, \infty)$. i.e., $0 < \mu < 10$ and $\sigma > 0$ and $n = 1$ with $x_1 = x$.
With everything else unknown, please estimate θ.

There are no bayesian nor frequentists solutions for this problem. Assuming a prior out of the blue is not allowed. However, there is an objective notion of most ignorant prior that depends on the value of some parameters. Namely, the scalar density field

$$\pi(p|t, \nu, \delta, \alpha) = \frac{1}{Z} \left[1 + \alpha \nu I_\delta(p : t)\right]^{\frac{-1}{\nu}} \tag{1}$$

where, t is the true distribution of the data (an infinite dimensional parameter), $\nu \in [0, 1], \delta \in [0, 1], \alpha > 0$ with α large enough so that $Z < \infty$. The I_δ is the δ information deviation, defined for unnormalized data distributions by,

$$I_\delta(p : t) = \frac{1}{\delta(1 - \delta)} \int [\delta p + (1 - \delta)t - p^\delta t^{1-\delta}]$$

One should think of this semi-parametric family, of most ignorant priors, as the information boundary of the inference problem. The numerical value of the ignorance action[6] achieved by the optimal scalar density field (1) provides a sticker price against which one should compare other priors. The value of the action provides a new objective procedure that allows to quantify the amount of extra information that it is added to the problem with a particular prior distribution. The geometric theory of ignorance does not always single out a unique prior but it always provides the optimal family of priors. Priors outside the optimal family can be ignored without penalty. This constitutes a

[6] see my *A Geometric Theory of Ignorance* for the exact definition.

clear improvement on the current state of statistical inference. There are still many open problems.

Let us finally consider the simplest case, still with infinite information volume, where the computations are easy and allow us to see some general points.

Case III: $\Theta = (-\infty, \infty) \times \{\sigma_0\}$. i.e., $-\infty < \mu < \infty, \sigma = \sigma_0$ given, $n = 1$ with $x_1 = x$.
With everything else unknown, please estimate θ.

The standard bayesian recipe $p(\mu)\,d\mu \propto d\mu$ even though improper does produce the posterior $N(x, \sigma_0^2)$ which is proper. The same answer is obtained with any member of the class of ignorant priors (1) as $\alpha \to 0$ provided that $t = N(\mu_0, \sigma_0^2)$ for any μ_0. When $\nu = 0$ with $\alpha > 0$ and $\delta = 0$ or $\delta = 1$, the standard conjugate prior $N(\mu_0, \sigma_0^2/\alpha)$ is obtained[7]. More interestingly with the same choices of parameters but with $\nu = 1$ instead of $\nu = 0$ the family of priors given by (1) is the family of Cauchy distributions with arbitrary location μ_0 and arbitrary scale $\sqrt{\frac{2}{\alpha}}\sigma_0$. But contrary to naive intuition the Cauchy family has infinite sticker price for this case. The gaussian conjugate family assumes less than the Cauchy as a family of priors for the unknown mean of a normal distribution with given variance. One needs $\nu \leq 2/3$ for the ignorance action to be finite.

Priors with heavy tails have been studied by the bayesians in relation with what is known as *bayesian robustness*. One of the major results from *bayesian robustness* is that priors with fat tails (like student-t distributions) produce more *robust* inferences than gaussian priors. The geometric theory of inference allows an objective and precise quantification of these type of results.

Parameters Models and Prior Information

A quick summary: For data x to be data it needs to be interpreted as a two sided object xp a dataTheory object. Honest inference needs to be fed with the uncertainty of the whole xp, i.e., we need a joint distribution for (x, p). In other words we need a likelihood $p(x)$ and a prior $\pi(p)$. Always. Thus, a range for (x, p) is needed. i.e., a set S, that I suggest to call, The Statistical Manifold specifying the support for dataTheories xp.

$$S = \{(x, p) : \text{Prob}(x, p) > 0\}$$

More generally if $\tilde{\mathscr{P}}$ denotes the cone of finite (i.e., normalizable) measures on a data space (manifold) $\mathscr{X}$ then, $S \subset \mathscr{X} \times \tilde{\mathscr{P}}$. This is the manifold of interest and the choice of this space is the main source of prior information. The manifold S specifies the *Subject Matter* of the inference problem. It should then be clear that the main source of prior information is not the prior distribution but the choice of S that includes a choice of model and even more important the choice of data space.

Statistical inference has concentrated mostly on S manifolds with simple product topology $S = \mathscr{X} \times M$. Where $x \in \mathscr{X}$ the data space and $p \in M$ a statistical model.

---

[7] see my *Antidata*

All the probabilities $p \in M$ are assumed to have the same essential[8] support $\mathscr{X}$ to fulfill the standard requirements for regularity. In this case we say that M is homogeneous. The most studied case of this type is the exponential family. More interesting topologies for S always show surprises for statistical inference.

The content of statistical inference is to be found in the group of symmetries of S. We should demand invariance under coordinate changes for S but also, and this is the most characteristic symmetry of inference, invariance under sufficient reductions of the data. Thus, two statistical manifolds, S and S' should be taken as equivalent for inference if there is a sufficient transformation $f : S \to S'$. A sufficient transformation is a map $f(x,p) = (y,q)$ that allows to recover data (x',p) equivalent to the original (x,p) from just knowledge of (y,q). Equivalent in the sense that x' and x are samples from the same distribution p, i.e. dataTheory points with the same hidden side. This defines a new category, first studied by Chentsov for the $S = \mathscr{X} \times M$ case. Let's refer to it as the S-category of Statistical inference. As it was mentioned above, this is reminescent of topology. The study of topology is the study of properties of sets that remain invariant under the group of continuous transformations. Two manifolds are topologically equivalent if they can be continuously deformed into each other. We say they are homeomorphic. Analogously, Statistical Inference is the study of the properties of S manifolds that remain invariant under sufficient maps[9].

One should expect the functors from the S-category to help move forward Statistical Inference into the really new territory of algebraic topology, K-theory and beyond.

I believe that new mathematics remeniscent of the discovery of topology in 1900 by Poincaré is just around the corner. When the information volume is infinite we need either real extra prior information beyond the Statistical Manifold S or a new way for extracting extra information from S. In principle there may be extra information in what I call *the abstraction sequence* [10]

$$S_0 \to S_1 \to S_2 \to \ldots,$$

where $S_0 = S$ is the Statistical Manifold containing the *meaningful* data space, and for $i = 1,2,\ldots$, the statistical manifold S_i is the information boundary (1) associated to S_{i-1}. This sequence provides the sequence of *explanations*. If we think of a probability distribution as providing an explanation for the data [11] then the model M containing the set of explanations should be thought as data at a higher level of abstraction that requires its own explanation provided by the manifold of ignorant priors, itself new data at a still more abstract level, etc. It is in principle posible for the sequence of explanations to finish at some $i < \infty$ if the information volume at level i is finite or there is a singularity of some kind. We enter here open territory where almost nothing is currently known.

---

[8] i.e., specified up to sets of measure zero.

[9] these were called *markovian morphisms* by Chentsov

[10] A rose is a rose is a rose... in my 1989 MaxEnt paper *Objective Bayesianism and Geometry*.

[11] Recall that probability distributions are in a one to one correspondance with codes. Check the MDL literature or T. Cover's book on Information Theory.

What's New?

The geometric theory of ignorance provides a radically new approach to statistical inference. It provides a better way. A much better way. It is the only honest way that we know off. This way works in any number of dimensions. It shows the optimality of conjugate priors for the exponential family when given relative to the volume form. It shows the optimality of priors with tails following power laws. The honest way, erases the bayes/frequentist dichotomy, it makes a dent at the mind/body puzzle, at the objective/subjective dichotomy, and it justifies the variational action used by Perelman to solve Thurston's geometrization conjecture (yes the Poincaré Conjecture, now a Theorem followed from it). In fact, this way does not even need bayes theorem at all. It produces it as a special case! It actually improves on it when there is uncertainty about the model. The honest way is built on the (unlikely united) shoulders of Edwin Jaynes MaxEnt, Herbert Robbins Empirical Bayes, and Shun-ichi Amari's Information Geometry. This way solves a 260 year old problem that goes all the way back to the beginings of probability with Bernoulli's Ars Conjectandi: Namely *The mathematical encoding of ignorance in statistical inference*. It turns out that All you need is *honesty*. That is all you need[12].

ACKNOWLEDGMENTS

Part of this research was done while I was on sabbatical leave at Riken's Brain Science Institute in Japan. I am in debt to Shun-ichi Amari for making it possible. I thank Ali Mohammad Djafari, the organizers of MaxEnt2010 and the Jaynes foundation for their financial support and to Ariel Caticha and Kevin Knuth for many stimulating conversations.

[12] apologies to The Beatles. tra-la-la. Cumbayá