Beyond Bayes

Carlos C. Rodríguez

http://omega.albany.edu:8008/
Department of Mathematics and Statistics
The University at Albany, SUNY
Albany, NY 12222
USA

Abstract. The Geometric Theory of Ignorance produces posterior distributions from priors and likelihoods without invoking Bayes Theorem at all. The standard bayesian posteriors minimize the ignorance action with parameters $\delta = \nu = 0$ when the truth t is replaced by the empirical distribution of n independent observations. Different values for the parameters produce new ways for processing the data obtaining posterior distributions that are maximally honest with respect to the explicitly available information and showing remarkable finite sample and asymptotic properties.

INTRODUCTION

There is no data in a true theoretical vacuum. By this I mean that it is impossible to have completely atheoretical data. For example a pure number, like $x = 2.7$ say, is not data unless it is understood as a logical proposition in a given domain of discourse e.g. *the result of a well defined experiment produced* $x = 2.7$. Thus, it is *TRUE* (or it is *FALSE*) that $x = 2.7$. Data is a logical proposition in a theoretical background. A theory is an explanation for the data. In a concrete and pragmatic way, a theory is an algorithm for compressing data. A theory is a code. A notion of likelihood is naturally identified with a notion of code length (e.g. $-\log p(x)$ being the standard length of the code for the symbol x occurring with probability $p(x)$). In this way a theory (i.e. a code) is identified with a probability distribution for the data. Every meaningful notion of expected code length defines a one-to-one correspondance between codes and probability distributions for the data. Thus, the best theory gives the shortest expected length for an encoded message. In other words in this approach maximum likelihood is best by definition. But of course this is tautological. It depends on what is likelihood and what is best, or equivalently, it depends on what we accept as expected code length. To realize that there is plenty of room for alternative notions of code length besides the standard $-\log p(x)$ (and thus likelihood) consider an information source with infinite Shannon entropy. For example, the distribution over the positive integers $k = 1, 2, \ldots$ with probabilities given by,

$$p(k) = \frac{1}{\log_2(k+1)} - \frac{1}{\log_2(k+2)}.$$

To find a "good" code for this distribution we either have to change the standard notion of code length or, equivalently, the standard notion of likelihood producing alternative,

but still meaningful, notions of entropy. Several questions immediately pop up: How much freedom is there about these choices of entropy, code length, likelihood?. How can we do inference with these more general notions?. When two or more choices are available how should we choose among them?. Are there advantages in using the more general notions in cases where the standard choices are still applicable?.

I show in this paper that the geometric theory of ignorance provides natural answers for some of these questions.

QUANTIFYING IGNORANCE WITH INFORMATION GEOMETRY

The basic space is the space of distributions for the combined (DATA,THEORY) vector (x,p). A statistical model is a set of theories, $M=\{p\}$, i.e. a set of probability distributions for the data. A prior is a probability distribution $\pi=\pi(p)$ over the model M. We denote by $\mathscr{P}$ the set of all possible probability distributions for the data x and by $\tilde{\mathscr{P}}$ and $\tilde{M}=\{\tilde{p}\}$ the corresponding cones of unnormalized distributions, i.e. $\tilde{p}=cp$ for some $c>0$. We often ommit the tildes and simply denote by p an unnormalized probability distribution. For $0<\delta<1$ we define the vector space $L_{1/\delta}$ of δ-powers of distributions containing all objects of the form $p^\delta f$ where p is a probability distribution and $f\in L_{1/\delta}(p)$ i.e., $\|p^\delta f\|^\delta=\int|f|^{1/\delta}p<\infty$. With this norm, the obvious definition of addition and multiplication by scalar, and the identification of equivalent objects, $L_{1/\delta}$ becomes a Banach space. Notice that $p^\delta\equiv q^\delta f$ iff $f^{1/\delta}=p/q$ is the density of p with respect to q. We also define the δ-coordinates of p by the vector in $L_{1/\delta}$ given by $l_\delta(p)=p^\delta/\delta$. Also, let $l_0(p)=l(p)=\log p$. For $\delta\in(0,1)$ we define the δ-information deviation for (unnormalized) probability distributions p and q by the finite non negative number,

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

For $\delta\in\{0,1\}, I_\delta$ is taken as the corresponding limit obtaining,

$$I_1(p:q)=\int\left(q-p+p\log\frac{p}{q}\right)=I_0(q:p)$$

coinciding with the Kullback number when $\int p=\int q$. The arithmetic mean $\delta p+(1-\delta)q$ is never smaller than the geometric mean $p^\delta q^{1-\delta}$ and they are both equal only when $p=q$. Hence, $I_\delta(p:q)\geq 0$ with equality iff $p\equiv q$. The space $L_{1/(1-\delta)}$ is the topological dual of the space $L_{1/\delta}$ and the only Hilbert space is the self-dual L_2 associated to $\delta=1/2$. This is the space of wave functions in quantum mechanics when the field of scalars is the complex plane. By using the δ-coordinates we can always regard the statistical model M as isometrically embedded in the Banach space $L_{1/\delta}$. All the δ-concepts, such as, δ-geodesic, δ-flat, δ-convex, etc are attributed to M when the δ-coordinates form a straight line, a flat space, a convex set, etc in the vector space $L_{1/\delta}$.

In particular, Fisher information is the metric induced on M by L_2. We say that M is regular when M forms a Riemannian manifold with Fisher information as the metric. The cone $\tilde{\mathscr{P}}$ of unnormalized probability distributions is δ-flat for all δ.

The Actions of Ignorance

Let π and η be priors on $M = \{p\}$. Let t be a probability distribution for the data x and, let $\beta > 0, 0 \leq \delta \leq 1, 0 \leq \nu < 1$. Define the positive scalar,

$$\mathscr{A} = \beta < I_\delta(p:t) >_\eta + I_{1-\nu}(\eta:\pi)$$

where we denote by $< f >_\eta = \int f\eta$ which is the expectation of f w.r.t. η when η is normalized. We call $\mathscr{A}$ an action of Ignorance since it defines a risk functional on the space of joint distributions of (x,p) where the data x is independent of the theory p. i.e., it ranks the *ignorant* distributions of the factorized form $t\eta = t(x)\eta(p)$. To see this consider the following two facts:

$$\mathscr{A} = \beta I_\delta(p\eta : t\eta) + I_{1-\nu}(\eta:\pi) \tag{1}$$

where we denote by $p\eta = p(x)\eta(p)$ the joint distribution on (DATA,THEORY) that picks p according to η and then picks x according to p. And,

$$I_1(t\eta : p\pi) = < I_0(p:t) >_\eta + c_t I_1(\eta:\pi) + (c_p - c_t)(c_\pi - c_\eta) \tag{2}$$

where $c_f = \int f$.

Fact (1) shows that for a given t, the η that minimizes $\mathscr{A}$ is most ignorant in the sense that makes $p(x)\eta(p)$ close to $t(x)\eta(p)$ and η close to π. When M has finite information volume (i.e. finite volume in L_2) then a simple choice for π is the uniform distribtion over M. Thus, in this case the best η is a compromise between making x and p independent (i.e. concentrating the mass about the δ-projection of t on M) and spreading the mass over M. Fact (2) shows that when $\delta = \nu = 0$ and $c_\eta = c_\pi$ minimizing $\mathscr{A}$ is equivalent to making $t\eta$ as close as possible to $p\pi$.

NEW POSTERIORS

The joint minimization of $\mathscr{A}$ over unnormalized distributions (t,η) is a trivial exercise in the calculus of variations. The optimality conditions are:

$$\eta(p) = [1 + \beta\nu I_\delta(p:t)]^{-\frac{1}{\nu}} \pi(p) \tag{3}$$

$$t^\delta(x) = \int_M p^\delta(x)\eta(p) \tag{4}$$

These two coupled functional equations could be solved, in principle, by iterative substitution starting from,

$$\eta_0(p) = \pi(p) \quad (5)$$

$$t_0^\delta(x) = \int_M p^\delta(x)\pi(p) \quad (6)$$

There is an error decomposition formula available. For any t we have,

$$< I_\delta(p:t) >_\pi = < I_\delta(p:t_0) >_\pi + I_\delta(t_0:t) \quad (7)$$

δ-likelihood

Equation (4) gives the δ-coordinates of the optimal t as the η-average of the δ-coordinates of p. This has the form of the predictive distribution of x but with p replaced with p^δ. Thus, given a sample $x^n = (x_1, \ldots, x_n)$ of n independent observations we define the δ-likelihood of p by,

$$L_{x^n}^\delta(p) = p^\delta(x_1)p^\delta(x_2)\cdots p^\delta(x_n) \quad (8)$$

Direct posteriors

For a given t the most ignorant prior is given by (3). When a sample x^n is available we can estimate t with either the empirical distribution $\hat{t}_n$ or a predictive distribution obtaining a direct (unnormalized) posterior,

$$\pi(p|x^n) = [1 + \beta\nu I_\delta(p:\hat{t}_n)]^{-1/\nu}\pi(p) \quad (9)$$

Asymptotic consistency is obtained when we let $\beta \to \infty$ as $n \to \infty$. In general these posteriors do not agree with bayes theorem unless we call likelihood the posterior density w.r.t. the prior π. Therefore, we call (δ, ν, β)-Likelihood the ratio $\pi(p|x^n)/\pi(p)$ obtained from (9). We also call $\hat{p}_n^{(1-\delta)}$ the $p \in M$ (when it exists) that maximizes this likelihood. Thus,

$$\begin{aligned} \hat{p}_n^{(1-\delta)} &= \arg\min_{p\in M} I_\delta(p:\hat{t}_n) \\ &= \arg\max_{p\in M} \sum_{i=1}^n p^\delta(x_i) \end{aligned} \quad (10)$$

standard maximum likelihood is the special case $\delta = 0$. In fact, $(\delta = 0, \nu = 0, \beta = \alpha n)$-Likelihood is the same as α-likelihood as we now show.

EXAMPLES AND SPECIAL CASES

The special case $\delta = \nu = 0$ with $\beta = \alpha n$ and $0 < \alpha \leq 1$ is particularly interesting. The generalized posterior (9) becomes,

$$\begin{aligned} \pi^{\alpha}(p|x^n) &= e^{-\alpha \sum_{i=1}^{n} \log(1/p(x_i))} \pi(p) \\ &= p^{\alpha}(x_1) \cdots p^{\alpha}(x_n) \pi(p) \end{aligned} \tag{11}$$

which recovers bayes theorem with δ-likelihood (8) when $\delta = \alpha$. Standard inference is the case $\alpha = 1$. Further more this shows that β and δ are related since the direct posterior with $\delta = 0$ and $\beta = \alpha n$ coincides with the δ-posterior with $\delta = \alpha$. Thus, it is natural to consider the consistency of this posterior in terms of I_α information separation.

Example: *A gaussian mean*. For given $\sigma > 0$ let $M = \{N(\theta, \sigma^2) : \theta \in R\}$. For any continuous prior π the most honest δ-predictor has δ-coordinates in the *center* of M defined by (6),

$$t_{\pi}^{\delta}(x) = \int \exp\left\{ \frac{-\delta}{2\sigma^2}(x-\theta)^2 \right\} \pi(\theta) d\theta \tag{12}$$

Notice that the normalized δ-coordinates define a diffusion in a fictitious time $\tau = \sigma^2/(2\delta)$ inversely proportional to δ. This suggests that, for the gaussian location model ignorance increases as $\delta \to 0$. The normalized delta coordinates satisfy the heat equation $(\partial_\tau + \triangle)\pi(\tau, x) = 0$ where $\triangle = -\nabla^2$ is the geometer's laplacian. On the other hand for a given t the best prior is given by (3). Thus if we guess t with a point in M, say $t = N(\theta_0, \sigma^2)$ then, for $\delta \approx 0$ (i.e. after a long ficticious time τ) we obtain,

$$\eta(\theta|\theta_0) = \left[1 + \beta\nu \frac{(\theta - \theta_0)^2}{2\sigma^2} \right]^{-1/\nu} \tag{13}$$

This is a student-t distribution with $(2/\nu - 1)$ degrees of freedom centered at θ_0 with scale $\sqrt{\frac{2}{\beta(2-3\nu)}}\sigma$. This suggests that ignorance increases as $\nu \to 1$. In the limit of maximum ignorance it becomes a Lorentzian.

When $\nu = 0$ the prior (from (3)) is $\eta(\theta|\theta_0) \equiv N(\theta_0, \sigma^2/\beta)$. Thus, $(\theta - \theta_0) \sim N(0, \sigma^2\tau)$ where we define $\tau = 1/\beta$ interpreted as the lapse of fictitious time between initial location θ_0 and final location θ. This is **not** brownian motion unless we promote the increments $(\theta - \theta_0)$ to the increments of an stochastic process indexed by $\tau > 0$, i.e. unless we assume time homogeneity and independence for non overlapping time lapses for an arbitrary number of such increments. The data space is now the infinite dimensional manifold of trajectories $\theta(\tau)$ and the statistical model M contains all the Wiener processes indexed by drifts $\theta_0(t)$, i.e. $\theta(t+\tau) = \theta_0(t) + \sigma W_\tau$ where $\{W_\tau : \tau > 0\}$ is the standard Wiener process.

When again $\pi = N(\theta_0, \sigma^2/\beta)$, (12) simplifies to $t_\pi = N(\theta_0, \sigma^2(1+\frac{1}{\beta}))$ which is outside of M and the same for all values of δ. It is often the case that the manifold of predictive distributions for different priors π provide a useful enlargement $\hat{M}$ of the original model M. In general when t is guessed by one $p_0 \in M$ the extended $\hat{M}$ is labeled by M itself and $\beta > 0$. We may take $\beta = \beta(p_0)$ making M and $\hat{M}$ of the same dimension, or the other way around $p_0 = p_0(\beta)$ labeling $\hat{M}$ by paths in M i.e. by the set $\{(\beta, p_0(\beta)) : \beta > 0, p_0(\beta) \in M\}$.

The value of I_δ giving the separation between the (unnormalized) $N(\theta, \sigma^2)$ and the (unnormalized) $N(\theta_0, \sigma^2/\beta)$ is given by,

$$I_\delta = \frac{\sqrt{2\pi}\sigma}{\delta(1-\delta)}\left[\delta + \frac{(1-\delta)}{\sqrt{\beta}} - \frac{1}{\sqrt{\delta+(1-\delta)\beta}}\exp\left\{\frac{-\delta(1-\delta)\beta(\theta-\theta_0)^2}{2\sigma^2(\delta+(1-\delta)\beta)}\right\}\right] \quad (14)$$

thus, the δ-projection of t_π onto the model M is achieved when the exponent is the largest posible i.e. at $\theta = \theta_0$. Hence, for the normal location model M with any normal π the best predictor in M is when θ is the mean of the prior π and this is the case for all δ. A robust and consistent direct posterior is given by (9) with $\beta = \alpha n$ by,

$$\pi(\theta|x^n) = \left[1 + \alpha n \nu \frac{(\theta - \bar{x_n})^2}{2\sigma^2}\right]^{-1/\nu} \quad (15)$$

This is not asymptotically gaussian. It remains a student-t with $(2/\nu - 1)$ degrees of freedom for all n. It has mean equal to the sample mean $\bar{x_n}$ provided $\nu < 1$ and variance given by,

$$\sigma_n^2 = \frac{2\sigma^2}{\alpha(2-3\nu)n} \quad (16)$$

provided $\nu < 2/3$.

THE GROUND STATE OF MAXIMUM IGNORANCE

The minimum possible value of $\mathscr{A}$ is obtained for a given t when η is (3). The actual numerical value depends on whether we use (3) or its normalized expression but the qualitative conclusion will be the same for both cases. Let us start with the unnormalized case. Let Z_ν be the integral over M of η. We have,

$$\begin{aligned} \nu(1-\nu)I_{1-\nu}(\eta:\pi) &= (1-\nu)Z_\nu + \nu - Z_\nu - \beta\nu < I_\delta >_\eta \\ &= \nu\left(1 - Z_\nu - \beta < I_\delta >_\eta\right) \end{aligned}$$

Hence,

$$\begin{aligned}\mathscr{A} &= \beta < I_\delta >_\eta + I_{1-\nu}(\eta:\pi) \\ &= \frac{1}{1-\nu}\left[1 - Z_\nu - \beta\nu < I_\delta >_\eta\right]\end{aligned}$$

For the two extreme values of ν the maximum ignorance becomes,

$$\lim_{\nu\to 0}\mathscr{A} = 1 - Z_0 \tag{17}$$

$$\lim_{\nu\to 1}\mathscr{A} = \infty. \tag{18}$$

This shows that for $\nu \approx 0$ the t that maximizes ignorance must maximize the evidence Z_ν. Equation (18) shows that the case $\nu = 1$ should either be excluded or $(1-\nu)\mathscr{A}$ instead of $\mathscr{A}$ be used.

Let us now consider the normalized case: $\eta^* = \eta/Z$. It is possible to show that for a given t, this η^* minimizes the action $\mathscr{A}^*$ with $\beta^* = \beta Z^\nu$. We have,

$$\nu(1-\nu)I_{1-\nu}(\eta^*:\pi) = 1 - Z^\nu - \beta\nu Z^\nu < I_\delta >_{\eta^*}$$

thus,

$$\mathscr{A}^* = \frac{1}{\nu(1-\nu)}\left[1 - Z^\nu - \nu^2\beta Z^\nu < I_\delta >_{\eta^*}\right] \tag{19}$$

and,

$$\lim_{\nu\to 0}\mathscr{A}^* = -\log Z \tag{20}$$

$$\lim_{\nu\to 1}\mathscr{A}^* = \infty. \tag{21}$$

ZHANG'S THEOREM

In [1, Theorem 4.1] Tong Zhang proved the following remarkable result about the asymptotic performance of the normalized direct α-posteriors introduced in (11). We use the notation,

$$\pi_n^\alpha(p|x^n) = \frac{p^\alpha(x_1)\cdots p^\alpha(x_n)\pi_n(p)}{\int_{M_n} p^\alpha(x_1)\cdots p^\alpha(x_n)\pi_n(p)}$$

Theorem 1 (Zhang 2003) *For any sequences π_n of priors on M_n and, ε_n of positive numbers converging to 0, if*

$$\sup_n \frac{-1}{n\varepsilon_n}\log\int_{I_0(p:t)\le\varepsilon_n}\pi_n(p) \;<\; \infty \tag{22}$$

then for $s_n \to \infty$ *and* $0 < \alpha < 1$ *we have,*

$$\int_{I_\alpha(p:t) \geq \varepsilon_n s_n} \pi_n^\alpha(p|x^n) \to 0 \tag{23}$$

in t-probability as $n \to \infty$*.*

Translation: The direct α-posterior π_n^α concentrates in an I_α ball of radius ε_n centered at t at rate $O_p(\varepsilon_n)$ provided only that the priors π_n put mass $O(e^{-cn\varepsilon_n})$ on the I_0 ball of radius ε_n centered at t. In other words, the only requirement for consistency is a local condition about the thinness of the priors π_n around t. The direct α-posteriors are thus robust against incorrect assignments of prior mass away from t. This is not always true for standard bayesian inference, (see the references in [1]) i.e. this is not true when $\alpha = 1$.

CONCLUSION

When the information source is t, the δ-information deviation $I_\delta(p:t)$ quantifies a redundancy (or opportunity loss), in terms of expected code length, in the code built from p instead of t. The problem is that the information source is often a moving target and almost never known with certainty. One way to handle the uncertainty about t is to provide a probabilistic model for t, i.e., a prior probability distribution π over $\mathscr{P}$ concentrating its mass in a subset $M \subset \mathscr{P}$. Since there is truly no meaningful data x without some kind of theory p, we can only attempt to transmit $(DATA, THEORY)$ as a whole and we need a code for (x,p) not just x. Several current data compressors (e.g. MP4) are based on this approach. To transmit a digital recording of a piano sonata, with the help of some parametric model M for the piano and the acoustics of the room, we transmit first the values of the parameters identifying $p \in M$ and then use that p to encode the long binary vector x. The ignorant actions $\mathscr{A}$ provide all the statistical invariant ways for measuring expected code length redundancies for transmitting $(DATA, THEORY)$. The $(1-\nu)$-information deviation $I_{1-\nu}(\eta:\pi)$ gives the redundancy in coding the theory with the prior η instead of the prior π and thus, $\mathscr{A}$ quantifies the total prize for transmiting the theory p and then the data x with the help of p. The actions $\mathscr{A}$ are true statistical invariants preserving all the symmetries of inference. These are: invariance under choice of coordinates for (x,p), invariance under choice of dominating measures for defining densities p, and invariance under sufficient reductions of the data. What seems most appealing in the code theoretical interpretations of statistical inference is the revelation that probability distributions, just like parameters, are *not true*, but only useful assumptions for encoding the data.

The actions $\mathscr{A}$ are given in non-parametric form and the model M does not need to be finite dimensional or regular. Thus, the theory of inference based on the minimization of $\mathscr{A}$ is extremely general with innumerable applications. Nevertheless, the bayesian fundamentalists are likely to reject direct posteriors or any other procedures manipulating the observations without the direct application of bayes theorem and strict adherence to the likelihood principle. I remind the fundamentalists that bayes theorem is a logical necesity only if we assume we know the prior and the model but that is almost never

the case in practice. Besides, the posteriors obtained from bayes theorem are one of the direct posteriors, namely the special case with parameters $\delta = \nu = 0$ and $\beta = n$.

Maximum Entropy is more general than Bayesian Inference. In fact, there is a trivial way in which this statement is correct not just for the standard entropy (associated to $\delta = 0$) but for all the $\delta = \alpha \in [0,1]$. To see this compute $Q^* = \arg\min I_\alpha(P:Q)$ where P and Q are joint distributions of (x,p) and the minimization is over all Q subject to the constraints that for all x, $\int_M Q(x,p) = \delta(x - x_0)$ i.e., subject to the constraint of observing x_0. The straight forward solution is $Q^*(x,p) = \delta(x - x_0)P(p|x)$ for all values of α. This result was first obtained in [2] for the case of $\alpha = 0$. Hence, minimization of I_α is also more general than bayesian inference.

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.

REFERENCES

1. T. Zhang, "Learning Bounds for a Generalized Family of Bayesian Posterior Distributions", *NIPS 2003*. Online at `http://stat.rutgers.edu/~tzhang/papers/nips03-bayes.pdf`
2. Caticha A. and Giffin A., "Updating Probabilities", *MaxEnt 2006*. Online at `http://arxiv.org/abs/physics/0608185v1`