A Geometric Theory of Ignorance

Carlos C. Rodríguez

Department of Mathematics and Statistics
The University at Albany, SUNY
Albany, NY

Abstract. This is an incomplete sketch of a theory that produces a Model and a prior on it, from observed data and other explicitly stated prior information. Such a theory shows the potential of explaining the universe around, and inside us. Such a theory is ultimately a theory of ignorance. I cry out loud: *it and bit from not!*.

Additional information is available online at `http://omega.albany.edu:8008/ignorance`

CONTENTS

[1] With respect to the operation of truth conjugation defined in the text

INTRODUCTION

Infandum, regina, iubes renovare dolorem, Troianas ut opes et lamentabile regnum cruerint Danai; quaeque ipse miserrima vidi, et quorum pars magna fui. Quis talia fando Myrmidonum Dolopumve aut duri miles Ulixi temperet a lacrimis? Et iam nox umida caelo praecipitat, suadentque cadentia sidera somnos.

THE MAIN PROBLEM OF INFERENCE

From prior information O and data $D = (x_1, x_2, \ldots, x_n) = x^n$ obtain a model $M = \{P_\theta : \theta \in \Theta\}$ and a prior $\pi = \pi(\theta)$ on M.

With the necessary ingredients: (O, D, M, π) we can cook a bayesian omelet on a large enough computer and instantiate,

$$\text{Inference} = \text{Bayes} + \text{Computing}$$

where Bayes stands for his theorem,

$$p(\theta|D,M,O) = \frac{p(D|\theta,M,O)p(\theta|M,O)}{p(D|M,O)} \tag{1}$$

In what follows we simply write $p = p(D|\theta,M,O)$ for the likelihood, $\pi = p(\theta|M,O)$ for the prior, and $Z = p(D|M,O)$ for the evidence.

The prior is half the problem

This paper concentrates primarily on a simplified version of the main task. We assume that a model M is available as part of the prior information O. Our problem is to find the prior distribution π over M. Hence, except for the last two sections where we consider model selection, we assume that $M \in O$.

We want to find an objective mechanism for producing prior distributions π from explicitly stated prior information O when the model M is part of O. Our only principle is honesty. We demand our π to be maximally ignorant about everything except what is explicitly contained in O.

IGNORANCE RELATIVE TO M AND O

To proceed, we need a concept of ignorance relative to a model M and possibly extra prior information O. Our proposal is based on the trivial realization that ignorance is nothing but uncertainty about truth. A quantity designed to measure the amount of ignorance contained in a given prior π over M must depend on the location of the true distribution, in the space of all distributions for the data. For the object of ignorance is truth; We are ignorant about truth. But ignorance is not just measured by the proximity of the true distribution $t(x)$ to the model M. A small model M far away from

t is not ignorant about truth, it is very knowledgeable, but about the wrong thing. It is precisely incorrect. Our initial task is to assess the amount of ignorance in a prior distribution π over a given model M, not the quality of M itself. Choosing M will be the subject of model selection that is treated at the end of this paper.

Our intuitive concept of ignorance is at the center of a tug of war between two forces pulling it in opposite directions. On the one hand we would like our ignorant prior π over M to be as spread as possible over M. On the other hand we would like π to make the parameter θ as stochastically independent as possible from the data x. In geometric terms π should be close to the uniform distribution ω over M and also the joint distribution $p(x,\theta) = p(x|\theta)\pi(\theta)$ (or $p(x)\pi(p)$ in parameter free notation) should be close to the independent model $t(x)\pi(\theta)$ (or $t(x)\pi(p)$). Even though the true distribution $t(x)$ is never known completely, it is often asymptotically learnable from observed data. Clearly, in the absence of actual observations or other restrictions, so that $O = \{M\}$ is all that is known, the concept of ignorance should reduce to plain uniformity over M. The other extreme, i.e., when $t(x)$ is precisely localized by O, the ignorant prior over M should reduce to a point mass at $\hat{p} \in M$ where $\hat{p}$ is the projection of t on M. The intermediate case should be a compromise between uniformity over M and concentration about $\hat{p}$. A trade-off between uniformity and independence.

To finish our attempt to quantify relative ignorance we need only a meaningful notion of separation between probability distributions. Thus, if D_1 is such a measure for joint distributions of (x,θ) and D_2 for distributions over M then, the maximum ignorant prior is defined as the π that minimizes the action,

$$\mathscr{A} = \alpha\, D_1(p\pi, t\pi) + D_2(\pi,\omega) \tag{2}$$

where α is a positive scalar parameter that measures the uncertainty about t and effectively weights the relative importance of independence versus uniformity.

In the next section we review the basic facts from information geometry. In particular, we identify the class of all the statistically meaningful measures of separation between unnormalized probability distributions as the family of δ-deviations.

INFORMATION GEOMETRY TOOLS

Differential geometry provides a powerful and natural language for statistical inference. We collect in this section some basic facts, definitions and notation for future reference. Amari's books are still the standard reference.

Let $\mathscr{P}$ be the space of all probability distributions on a given (Hausdorff) measurable space. We denote by $\tilde{\mathscr{P}}$ the cone of finite positive measures on the same measurable space, i.e. $\tilde{p} \in \tilde{\mathscr{P}} \iff \tilde{p} = cp$ with $p \in \mathscr{P}$ and $c > 0$. Notice that $\tilde{\mathscr{P}}$ is closed under addition and multiplication by a positive number but $\mathscr{P}$ is not. Recently, Zhou and Rohwer have demonstrated that there is nothing to be lost but a lot to be gained by considering objects in $\tilde{\mathscr{P}}$ rather than in $\mathscr{P}$. We follow here their advice and most of their notation.

A model (also known as regular statistical model or hypothesis space) M is a subset of $\mathscr{P}$ which is also a riemannian manifold with Fisher information as the metric. For $p \in \tilde{\mathscr{P}}$ and $\delta \in (0,1]$ we denote by,

$$l_\delta = l_\delta(p) = \frac{p^\delta}{\delta} \tag{3}$$

we also define $l_0 = l = \log p$. We call $l_\delta(p)$ the δ-coordinates of $p \in \tilde{\mathscr{P}}$. Notice that $l_\delta(p) \in L_{1/\delta}$ the space of δth power finite measures defined by,

$$L_{1/\delta} = \left\{ p^\delta f : p \in \tilde{\mathscr{P}} \text{ and } f \in L_{1/\delta}(p) \right\} \tag{4}$$

$L_{1/\delta}$ is a Banach space with the obvious identification of objects, $p^\delta f \equiv q^\delta g$ whenever the $g \in L_{1/\delta}(q)$ is actually $(p/q)^\delta f$. Notice that p/q is a Radon-Nikodým derivative, when it exists.

Fisher information on the whole $\tilde{\mathscr{P}}$ is given for all $\delta \in [0,1]$ by

$$I(p) = (g_{ij}(p)) = (<\partial_i, \partial_j>) = \left(\int p\; \partial_i l\; \partial_j l \right) = \left(\int \partial_i l_\delta\; \partial_j l_{1-\delta} \right) \tag{5}$$

where $g_{ij}(p)$ is given relative to a choice of an ordered Hilbert basis in L_2. Thus, ∂_i denotes the Gateaux derivative in the direction of the chosen ith coordinate basis in L_2. It is convenient to think of the infinite dimensional manifold,

$\tilde{\mathscr{P}}$ with the above metric, as the ambient external space for our models M. The concepts of δ-geodesic, δ-flat, and δ-convex M, as it is embedded in $L_{1/\delta}$, are then just defined with the help of the coordinates $l_\delta(p)$ in the Banach space $L_{1/\delta}$. We have these properties when the δ-coordinates form respectively a straight line, flat set, and convex set in $L_{1/\delta}$.

We denote by $dV = dV_M$ the volume form for the model M. In a given parameterization,

$$dV = \omega(\theta) = |I(\theta)|^{1/2} d\theta = \sqrt{\det I(\theta)}\, d\theta \tag{6}$$

For $\delta \in (0,1)$ and $p,q \in \tilde{\mathscr{P}}$ the δ-information-deviation (or just δ-deviation for short) between p and q is,

$$I_\delta(p:q) = \frac{1}{\delta(1-\delta)} \int [\delta p + (1-\delta)q - p^\delta q^{1-\delta}] \tag{7}$$

and for $\delta \in \{0,1\}$ just take limits of the expression above to obtain the Skilling actions,

$$I_1(p:q) = I_0(q:p) = \int \left(q - p + p \log \frac{p}{q} \right) \tag{8}$$

When $\delta = 1/2$,

$$I_{1/2}(p:q) = 2 \int (\sqrt{p} - \sqrt{q})^2 \tag{9}$$

is twice the square of the Hellinger distance, i.e. the familiar L_2 distance between wave functions. Notice that to compute $I_\delta(p:q)$ one picks any measure $r \in \tilde{\mathscr{P}}$ dominating both p and q (e.g. $p+q$ is always a choice), replace p and q by their densities with respect to r, i.e. p/r and q/r and carry out the integral with respect to the chosen measure r. The final result is independent of the choice of r. Thus, I_δ is truly a functional of the positive measures p and q and not just their densities.

Properties of δ-information

The family of δ-deviations has a number of remarkable properties:

1. homogeneity: $I_\delta(cp:cq) = c\, I_\delta(p:q)$ for all $c > 0$.

2. positivity: $I_\delta(p:q) \geq 0$ with equality iff $p \equiv q$.

3. duality: $I_\delta(p:q) = I_{1-\delta}(q:p)$

4. invariance: for any $T : \mathscr{X} \longmapsto \mathscr{Y}$ with positive jacobian we have,

$$I_\delta(p:q) = I_\delta(p_T:q_T)$$

where $p_T = p \circ T^{-1}$ is the transformed probability distribution in $\mathscr{Y}$.

5. sufficiency: $T : \mathscr{X} \longmapsto \mathscr{Y}$ is sufficient for discriminating probability distributions p and q (i.e. $\int p = \int q = 1$) iff

$$I_\delta(p:q) = I_\delta(p_T:q_T)$$

6. uniqueness: $I_\delta(p:q)$ are the only functions with the above properties.

7. topological equivalence: For $\delta \in (0,1)$ all the I_δ topologies are equivalent to the Hausdorff topology of the Hellinger metric, i.e. the topology generated by $I_{1/2}$.

8. Taylor expansion: In a given parameterization,

$$I_\delta(\theta + \varepsilon v : \theta) = \frac{1}{2} g_{ij} v^i v^j \varepsilon^2 + \frac{1}{6}[\overset{0}{\Gamma}_{ijk} + \overset{\delta}{\Gamma}_{kij} + \overset{1}{\Gamma}_{jki}] v^i v^j v^k \varepsilon^3 + o(\varepsilon^3)$$

where the Christoffel symbols are given by Amari's δ-connection,

$$\overset{\delta}{\Gamma}_{ijk} = \left\langle \overset{\delta}{\nabla}_{\partial_i} \partial_j, \partial_k \right\rangle = \int p\, [\partial_i \partial_j l + \delta\, \partial_i l \partial_j l] \partial_k l$$

The Levi-Civita metric connection corresponds to the self-dual case $\delta = 1/2$.

9. eguchi relations:

$$g_{ij} = -\partial_i \dot{\partial}_j I_\delta(p:\dot{q})|_{p=q} \quad \overset{\delta}{\Gamma}_{ijk} = -\partial_i \partial_j \dot{\partial}_k I_\delta(p:\dot{q})|_{p=q}$$

10. generalized cosine:

$$I_\delta(p:r) + I_\delta(r:q) = I_\delta(p:q) + \int (l_\delta(p) - l_\delta(r))(l_{1-\delta}(q) - l_{1-\delta}(r))$$

11. generalized pythagoras: If the δ-geodesic connecting p to r is orthogonal to the $(1-\delta)$-geodesic connecting r to q then,

$$I_\delta(p:r) + I_\delta(r:q) = I_\delta(p:q)$$

Special cases of δ-information-deviations have been discovered and re-discovered more, perhaps, than any other concept in the history of statistics. With the sole exception of $\delta = 1/2$, the δ-deviations are not symmetric, and do not satisfy the triangular inequality so they don't define distances, in the usual way. However, the above properties make them to be the only statistically meaningful way of measuring separation in the extended space of unnormalized probability distributions. The δ-information-deviations are the one and only one measures that are positive definite, and preserve all the fundamental symmetries of statistical inference, i.e. invariant under coordinate transformations of both the data and the parameter space and invariant under sufficient reductions of the data. If further more, one demands additivity over independent sources of information then, the only survivors are the Skilling actions which coincide with the Kullback numbers for normalized probability measures.

δ-flat Models

To finalize our quick over-view of IG we mention that intrinsically δ-flat models (i.e. with zero riemman tensor associated to the δ-connection) are always also $(1-\delta)$-flat and therefore admit mutually orthogonal geodesic coordinate systems (θ, η). The θ coordinates are δ-affine, i.e. make the δ-Christoffel symbols to be 0. The η coordinates are $(1-\delta)$-affine and make the $(1-\delta)$-Christoffel symbols 0. More over, in these special coordinate systems the metric g_{ij}, and its inverse g^{ij} are obtained by differentiating scalar potentials $\psi(\theta)$ and $\phi(\eta)$, i.e.,

$$g_{ij}(\theta) = \partial_i \partial_j \psi(\theta), \quad g^{ij}(\eta) = \partial^i \partial^j \phi(\eta). \tag{10}$$

and the two potentials are Legendre transforms of each other (just like entropy and free energy in Statistical Mechanics),

$$\psi(\theta) = \theta^i \eta_i - \phi(\eta) \tag{11}$$

with

$$\theta^i = \partial^i \phi(\eta), \quad \text{and} \quad \eta_i = \partial_i \psi(\theta). \tag{12}$$

In these coordinates,

$$I_\delta(p_\theta : q_\eta) = \psi(\theta) + \phi(\eta) - \theta^i \eta_i. \tag{13}$$

The extended space $\tilde{\mathscr{P}}$ is δ-flat and δ-convex for all $\delta \in [0,1]$. The space of all the discrete distributions on a finite set of atoms is also intrinsically δ-flat for all δ. Exponential family models are intrinsically $(0,1)$-flat.

ALL THE INVARIANT ACTIONS FOR IGNORANCE

As shown in the previous section, Information Geometry provides the needed class of measures of separation to be used in (2). Two are needed since we need to measure separation in two different spaces. Pick I_δ for deviations between joint distributions of (x, θ) and pick $I_{1-\nu}$ for quantifying deviations of distributions over M. We arrive to a three parameter family of invariant actions,

$$\mathscr{A}(\pi;\alpha,\delta,\nu) = \alpha\, I_\delta(p\pi : t\pi) + I_{1-\nu}(\pi : \omega) \tag{14}$$

where the parameters, $\alpha > 0$, $\delta \in [0,1]$ and $\nu \in [0,1]$ are free for now. The actions defined in (14) also depend on t and implicitly on M and O but it is not shown explicitly to keep the notation simple. Even though p appears on the right hand side of (14), it is integrated over inside I_δ so $\mathscr{A}$ itself is not a function of p.

The δ-deviations are well-defined in the extended space $\tilde{\mathscr{P}}$ of unnormalized positive measures and the extra freedom makes the optimization of actions of type (14) simpler. To take advantage of this benefit we adhere from now on to the following notational convention:

Assumption 1 *Distributions over M, like π and ω may appear unnormalized. Equations such us, $\pi(\theta) = f(\theta)$ could mean $\pi(\theta) \propto f(\theta)$ when needed and π may stand for a scalar density or a form depending on the context.*

We are now ready to formally define relative ignorance.

Definition 1 (relative (δ,ν)-ignorance) *We say that a π^* is (δ,ν)-ignorant at level $\alpha > 0$ relative to M and O iff*

$$\pi^* = \arg\min_{\pi \in O} \mathscr{A}(\pi;\alpha,\delta,\nu) \tag{15}$$

where, for lack of better notation, we write $\pi \in O$ to mean that the minimum is to be taken over all the π that agree with the prior information contained in O. Before writing the general solution for the case $O = \{M\}$, we consider the following:

Theorem 1 $I_\delta(p\pi : t\pi) = \int_M I_\delta(p : t)\ \pi(p) dV.$

Proof Just an application of Fubini's theorem. The case $\delta \in \{0,1\}$ follows by continuity from the case $\delta \in (0,1)$ for which,

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

Q.E.D.

In words: The δ-deviation between the joint distribution $p(x)\pi(p)$ and the independent model $t(x)\pi(p)$ is always equal to the expected δ-deviation between $p(x)$ and $t(x)$, where the expectation is taken with respect to the prior $\pi(p)$. Thus, the choice of δ is effectively a choice of distance for distributions over the data space. The choice of ν specifies the distance for distributions over M or, equivalently, the parameter space. With the help of this theorem let us re-write the actions (14) as,

$$\begin{aligned}\mathscr{A}(\pi;\alpha,\delta,\nu) &= \alpha\ \langle I_\delta(p:t)\rangle_\pi + I_{1-\nu}(\pi : \omega) & (16)\\ &= \alpha \left\{ I_\delta(p\pi : t\pi) + \frac{1}{\alpha} I_{1-\nu}(\pi : \omega) \right\} & (17)\end{aligned}$$

Using the standard arguments from optimization subject to inequality constraints (Khun-Tucker) we can show that the π^* minimizing (16) is the same as the π^* closer to the uniform over M (in the $(1-\nu)$-deviation) subject to a maximum average distance (in the δ-deviation) from t. In the case $\nu = 0$ this is literally maximum entropy subject to a constraint. In a similar way, by using (17) we can choose to optimize independence subject to an inequality constraint on uniformity. We write it as a theorem,

Theorem 2 *The following are all equivalent:*

1. *π^* is (δ,ν)-ignorant at level $\alpha > 0$.*
2. $$\begin{aligned}\pi^* = \ & \arg\max\ \{-I_{1-\nu}(\pi : \omega)\} \\ & s.t. \\ & < I_\delta(p:t) >_\pi\ \le\ E_{\max}\end{aligned} \tag{18}$$
3. $$\begin{aligned}\pi^* = \ & \arg\min\ I_\delta(p\pi : t\pi) \\ & s.t. \\ & I_\nu(\omega : \pi)\ \le C_{\max}\end{aligned} \tag{19}$$

where $E_{\max}, C_{\max}$, and α are functions of each other.

Theorem (2) provides a simple geometric characterization for our concept of ignorance. The ignorant prior is the one that maximizes spread over M subject to keeping the mean distance from a fix truth t less than a maximum acceptable value $E_{\max}$. Clearly, if $t \notin M$ then $E_{\max} \geq I_\delta(\hat{p} : t) > 0$ otherwise there will be no solution.

MAXIMUM IGNORANCE

We have found a path to ignorance so let us dare to walk the walk.

Theorem 3 *Let*

$$\alpha_{\min} = \inf\{\alpha > 0 : Z(\alpha) = \int_M [1 + \alpha\nu\, I_\delta(p:t)]^{-\frac{1}{\nu}}\ dV < \infty\} \tag{20}$$

Then, for $\alpha > \alpha_{min}$, when $O = \{M\}$ the (δ, ν)-ignorant prior at level $Z^\nu \alpha$ is,

$$\frac{\pi^*}{\omega} = \frac{1}{Z(\alpha)} [1 + \alpha\nu I_\delta]^{-\frac{1}{\nu}} \tag{21}$$

Proof Impose the normalization constraint and use (16) to write, $\mathscr{A} + (\lambda/\nu)\int \pi = \int \mathscr{L}$ with,

$$\mathscr{L} = Z^\nu \alpha I_\delta \pi + \frac{\pi}{\nu} + \frac{\omega}{1-\nu} - \frac{\pi^{1-\nu}}{1-\nu}\frac{\omega^\nu}{\nu} + \frac{\lambda}{\nu}\pi$$

where we have replaced α with $Z^\nu \alpha$. Thus,

$$\frac{\delta \mathscr{L}}{\delta \pi} = 0 \iff \frac{\pi^*}{\omega} = [1 + \lambda + Z^\nu \alpha\nu I_\delta]^{-\frac{1}{\nu}}$$

which coincides with the normalized density (21) when $\lambda = Z^\nu - 1$.
Q.E.D.

At this point the reader is encouraged to praise the author for his choice of notation. All the symbols in (21) are statistically and geometrically meaning-full. The left hand side of (21) is a Radon-Nikodým derivative. The right hand side is a normalized scalar density field over M. Several families of prior distributions that are routinely used in practical applications, as subjective priors, are special cases of (21). Hence, equation (21) may provide an objective justification for their practical use. In particular, equation (21) includes the family of multivariate Student-t distributions when e.g. M is a location model in the exponential family, $\delta \in \{0,1\}$ and $\nu \in (0,1]$. For this reason we may think of (21) as a generalized-t family. The case $\nu \in \{0,1\}$ is obtained from (21) by taking limits. As $\nu \to 0$, we obtain $(\pi^*/\omega) \to \exp(-\alpha I_\delta)$. Thus, equation (21) includes all the entropic priors as the special case $\nu = 0$ and $\delta \in \{0,1\}$. The $(0,0)$-ignorant priors include the standard[2] family of conjugate priors for the exponential family likelihoods (see below). The case $\nu = 1$ can be thought as a generalized multivariate Cauchy. Jeffreys' priors are the case $\alpha = 0$ but theorem (2) (see the paragraph right after it) shows that in some cases there may be an $\alpha_{\min} > 0$ beyond which the solution is meaningless. An $\alpha_{\min} > 0$ may be needed even if $t \in M$ for some models with infinite volume or with unbounded densities, in order to assure that the prior is normalizable. All the traditional maximum entropy distributions are also included as the limit $\alpha \to \infty$.

IGNORANCE IS NEGATIVE FREE ENERGY

Let $Z = Z(\alpha, \delta, \nu, t)$ be the normalizing partition function for the family of ignorant priors (21), i.e.,

$$Z = \int_M [1 + \alpha\nu\, I_\delta(p:t)]^{-\frac{1}{\nu}}\ dV. \tag{22}$$

[2] Only if they are interpreted as scalar densities on M

Definition 2 (Free Energy) *The function $F = F(\alpha, \delta, \nu, t)$ given by*

$$F = -\log Z \tag{23}$$

is the free energy associated to the (δ, ν)-ignorant prior (21).

We show that the actual minimal value of the action at the optimal (δ, ν)-ignorant prior is an increasing function of the free energy and it is exactly the free energy when $\nu = 0$.

Theorem 4 *Let the model M and the ignorant prior be normalized, i.e., $\int \omega = \int \pi^* = 1$. Then the minimal value of the action that (21) minimizes is $\mathscr{A}^*$ given by,*

$$\mathscr{A}^* = \frac{1}{\nu}(1 - e^{-\nu F}) + \nu I_{1-\nu} \tag{24}$$

Moreover,

$$\lim_{\nu \to 0} \mathscr{A}^* = F \tag{25}$$

$$\lim_{\nu \to 1} \mathscr{A}^* = 1 - e^{-F} + I_0 \tag{26}$$

$$\tag{27}$$

Proof From theorem (3) we have,

$$\mathscr{A}^* = Z^\nu \alpha < I_\delta > + I_{1-\nu} \tag{28}$$

using (21) and the normalization assumptions we compute,

$$\begin{aligned} I_{1-\nu} &= \frac{1}{\nu(1-\nu)}\left(1 - \int \left(\frac{\pi^*}{\omega}\right)^{1-\nu} \omega\right) \\ &= \frac{1}{\nu(1-\nu)} - \frac{Z^\nu}{\nu(1-\nu)} \int [1 + \alpha \nu I_\delta] \pi^* \\ &= \frac{1}{\nu(1-\nu)} - \frac{Z^\nu}{\nu(1-\nu)} (1 + \alpha \nu < I_\delta >) \end{aligned} \tag{29}$$

substituting (29) into (28),

$$\begin{aligned} \mathscr{A}^* &= \frac{1}{\nu(1-\nu)} \left\{1 - (1 + \nu^2 \alpha < I_\delta >) e^{-\nu F}\right\} \\ &= \frac{1}{\nu(1-\nu)} \left\{1 - (1 + \nu^2 [\mathscr{A}^* - I_{1-\nu}] e^{\nu F}) e^{-\nu F}\right\} \end{aligned}$$

solving for $\mathscr{A}^*$ we obtain (24).
Q.E.D.
Theorem (4) establishes a connection between amount of ignorance and free energy that naturally prompts us to ask new questions that theorem (4) is unable to answer. For example, (24) suggests, but it does not prove, that the optimal action for ignorance $\mathscr{A}^*$ is always an strictly increasing function of F for all $\nu \in [0, 1]$. We can only deduce from (24) and (25) that $\mathscr{A}^*$ is increasing in F for $\nu \in [0, \varepsilon]$ for some $\varepsilon > 0$. A more detailed analysis of $I_{1-\nu}$ as a function of F seems to support the conjecture that $\mathscr{A}^*$ is increasing in F for all $\nu \in [0, 1]$. A final proof is not available at the moment. No proof is really needed if we change the problem a little... if the mountain doesn't come to Mohammed...

The problem, I believe, is our original definition for free energy (23). Let us look into the explanation of the traditional thermodynamic quantity by means of statistical mechanics. The phenomenological concept of free energy $A = E - TS$, as originally defined by Helmholtz, is related to the statistical mechanical concept of sum-over-states (i.e.

partition function) Z by the usual formula $A \equiv -kT \log Z$ but only when $Z = \sum_n e^{-E_n/kT}$ i.e. only when the partition function is the normalizing constant of a model in the exponential family. But our family of $(\delta, \nu))$-ignorant priors is not in the exponential family unless $\nu = 0$. This suggests redefining free energy as the optimal value of our action for ignorance, in which case, ignorance is negative free energy by definition!

It is intuitively clear that ignorance is the other side of the coin of information. We should expect something like

$$\text{Information} + \text{Ignorance} = O. \tag{30}$$

We know from Shannon that negative entropy is information. Theorem (4) shows that, at least for $\nu = 0$, negative free energy is ignorance. Should we then expect the sum of the two to give the whole glass of knowledge?. What is the whole glass anyway?. Shouldn't the whole glass change somehow dynamically? For otherwise how are we going to be able to account for true evolutionary innovation?

IGNORANCE IS SELF-ADJOINT [3]

Let us denote by $\pi(p|t)$ the right hand side of (21). If we assume that $t \in M$ then we can regard $\pi(p|t)$ as a function of two arbitrary distributions in M. Let us also consider an arbitrary smooth relabeling of the elements of M. Thus, each distribution in M has two names, $p = p(q)$ and $q = q(p)$. Also, each distribution in M has two roles that it can play. It can act as the truth t or as just p. We denote the role-flipping operator with a hat on top of the name. Hence,

$$\pi(p|\hat{q}) = [1 + \alpha\nu I_\delta(p : \hat{q})]^{-\frac{1}{\nu}} \tag{31}$$

denotes the (δ, ν)-ignorant prior in the p-coordinates given truth $\hat{q}$ represented in the q-coordinates. Since ignorant priors are scalar functions which are invariant under relabelings we have,

$$\pi(p|\hat{q}) = \pi(q|\hat{p}). \tag{32}$$

Define the operation of *coordinate transposition*,

$$(p, q) \to (q, p) \tag{33}$$

that defines the transpose of π as,

$$\pi^\top(p|q) = \pi(q|p) \tag{34}$$

and define also the operation of *truth conjugation*,

$$p \to \hat{p} \tag{35}$$

that defines the conjugate of π as,

$$\bar{\pi}(p|\hat{q}) = \pi(\hat{p}|q). \tag{36}$$

If we define the adjoint of π as its conjugate transpose and we denote it by $\pi^\dagger$, then using (32) we obtain,

$$\pi(p|\hat{q}) = \pi^\top(\hat{q}|p) = \pi^\dagger(q|\hat{p}) = \pi(q|\hat{p}). \tag{37}$$

So $\pi = \pi^\dagger$ which is our definition of self-adjoint.

[3] WITH RESPECT TO THE OPERATION OF TRUTH CONJUGATION DEFINED IN THE TEXT

IGNORANT PRIORS ON δ-CONVEX MODELS

A δ-convex set of unnormalized distributions is defined, as most δ-properties are, by looking at its δ-coordinates. It is δ-convex when the set of the δ-coordinates of its members is convex as a subset of the Banach space $L_{1/\delta}$. Straight lines of δ-coordinates are δ-geodesics so a set S is δ-convex iff given two points in S, the delta geodesics connecting them are included in S. Models which are δ-convex, have the remarkable property of being able to represent truth faithfully with one of their members. But in general there is a price to be paid. A model M always entails a compromise between simplicity (small, computationally tractable) and accuracy (big, complex). There is in general no warranty that the true distribution $t(x)$ is in M. As the following theorem shows, when the model is δ-convex, we can replace $t \notin M$ by its $(1-\delta)$-projection $\hat{q} \in M$ without missing a bit.

Theorem 5 *Let M be closed (in the Hellinger topology) and δ-convex. Let $\pi(p|t,\alpha)$ denote the (δ,ν)-ignorant prior at level α when the true distribution is t. Then,*

$$\pi(p|t,\alpha) = \pi(p|\hat{q},\hat{\alpha}) \tag{38}$$

where $\hat{p}$ is the unique $(1-\delta)$-projection of t on M, i.e.,

$$I_{1-\delta}(t:\hat{q}) = \min_{p\in M} I_{1-\delta}(t:p) \tag{39}$$

and

$$\hat{\alpha} = \frac{\alpha}{1+\alpha\nu I_\delta(\hat{q}:t)} \tag{40}$$

Proof Use the generalized law of cosines and generalized pythagoras to show that the $(1-\delta)$-projection of t on a closed δ-convex M, exists, it is unique, and belongs to M. Just as it is the case in a Hilbert space (see Amari85 p99 theorem3.9). Now recall that $\tilde{\mathscr{P}}$ is δ-flat for all δ so the $(1-\delta)$-geodesic in $\tilde{\mathscr{P}}$ connecting t to the projection $\hat{q} \in M$ is orthogonal to the δ-geodesic connecting $\hat{q}$ to an arbitrary $p \in M$. Thus, by generalized pythagoras,

$$I_\delta(p:t) = I_\delta(p:\hat{q}) + I_\delta(\hat{q}:t). \tag{41}$$

Replacing (41) in the formula for the ignorant prior (21) and sticking to assumption (1) we have,

$$\begin{aligned}
\pi(p|t,\alpha) &= [1+\alpha\nu\, I_\delta(\hat{q}:t) + \alpha\nu\, I_\delta(p:\hat{q})]^{-\frac{1}{\nu}} \\
&= [1+\alpha\nu\, I_\delta(\hat{q}:t)]^{-\frac{1}{\nu}}\, [1+\hat{\alpha}\nu\, I_\delta(p:\hat{q})]^{-\frac{1}{\nu}} \\
&= \pi(p|\hat{q},\hat{\alpha})
\end{aligned} \tag{42}$$

Q.E.D.

Equation (40) is interesting. It quantifies the sticker price to be paid when replacing $t \notin M$ with $\hat{q} \in M$. It is clear that α measures, the amount of information in the ignorant prior. The smaller the α, the closer the ignorant prior is to the uniform over M. Now, equation (40) is telling us that if we want to guess t with some distribution in M we can do so. Prior and likelihood will stay the same and therefore the inferences will be exactly the same. However, the new prior needs to be less informative about the projected truth in order to be able to match the original. On the one hand equation (40) confirms the obvious: Information about t is worth less the farther away t is from M. On the other hand, equation (40) provides a precise quantification that, as far as I know, it was not known before. For example, if $t \notin M$ and $\nu > 0$, then as we collect more and more information about t so that α increases, $\hat{\alpha}$ approaches monotonically from below, the number $(\nu I_\delta(\hat{q}:t))^{-1} > 0$. Equation (40) also tells us that $\nu = 0$ is very special.

IGNORANCE ON δ-FLAT MODELS

Models that are intrinsically δ-flat, admit dual geodesic coordinate systems (see (11),(12),(13)), and they are obviously δ-convex. Thus, theorem (5) is applicable. We have,

Theorem 6 *Under the conditions and notation used in theorem (5). If M is δ-flat then, in the δ-affine coordinates θ,*

$$\pi(\theta|\hat{\theta},\hat{\alpha}) = \left[1+\hat{\alpha}\nu\left\{(\psi(\theta)-\psi(\hat{\theta}))-(\theta-\hat{\theta})\cdot\hat{\eta}\right\}\right]^{-\frac{1}{\nu}} \tag{43}$$

and in the $(1-\delta)$*-affine coordinates* η*,*

$$\pi^{\dagger}(\eta|\hat{\eta},\hat{\alpha}) = \left[1+\hat{\alpha}\nu\left\{(\phi(\eta)-\phi(\hat{\eta}))-\hat{\theta}\cdot(\eta-\hat{\eta})\right\}\right]^{-\frac{1}{\nu}} \tag{44}$$

where $\hat{\theta} = \theta(\hat{\eta})$ *and* $\hat{\eta} = \eta(\hat{\theta})$ *are the coordinates of* $\hat{q} \in M$*.*

Proof Let $p = p_\theta$ and use (13) to write,

$$\begin{aligned} I_\delta(p:\hat{q}) &= I_\delta(\theta:\hat{\eta}) \\ &= \psi(\theta)+\phi(\hat{\eta})-\hat{\eta}\cdot\theta \\ &= (\psi(\theta)-\psi(\hat{\theta}))-\hat{\eta}\cdot(\theta-\hat{\theta}) \end{aligned}$$

the last equality follows by noticing that $\hat{\eta}$ and $\hat{\theta}$ are the coordinates of the same point $\hat{q}$ and thus $I_\delta(\hat{\theta}:\hat{\eta}) = 0$. Replacing into the ignorance prior given by theorem (5) we obtain (43). To obtain (44) just notice that $I_\delta(\theta:\hat{\eta}) = I_{1-\delta}(\hat{\theta}:\eta)$
Q.E.D.
A simple, but useful, corollary of theorem (6) is,

Theorem 7 *Under the conditions of theorem (6) we have,*

$$\pi(\theta|\hat{\theta},\hat{\alpha}) = \left[1+\hat{\alpha}\nu\left\|\theta-\hat{\theta}\right\|^2_{\hat{\theta}}\right]^{-\frac{1}{\nu}}(1-\varepsilon) \tag{45}$$

where $\varepsilon = o(\left\|\theta-\hat{\theta}\right\|^2_{\hat{\theta}})$ *and* $\|\cdot\|_{\hat{\theta}}$ *denotes the norm induced by the riemannian metric at* $\hat{\theta}$*.*

Proof The result follows at once by expanding $\psi(\theta)$ in a Taylor series about $\hat{\theta}$, using (13) and (11) and replacing into (43).
Q.E.D.
The previous theorem shows that the *scalar densities* of ignorant priors on δ-flat models are approximated around $\hat{\theta}$ by multivariate Student-t distributions centered at $\hat{\theta}$ on the manifold M. Recall that the (unnormalized) density of a Student-t with d degrees of freedom is,

$$T_d(x) = \left[1+\frac{x^2}{d}\right]^{-\frac{(d+1)}{2}} \tag{46}$$

and matching d to ν we obtain,

$$d = \frac{2}{\nu}-1 \tag{47}$$

The approximation (45) becomes,

$$\left[1+\hat{\alpha}\nu\left\|\theta-\hat{\theta}\right\|^2_{\hat{\theta}}\right]^{-\frac{1}{\nu}} = \left[1+\frac{\hat{\alpha}(2-\nu)\left\|\theta-\hat{\theta}\right\|^2_{\hat{\theta}}}{d}\right]^{-\frac{(d+1)}{2}} \tag{48}$$

The extra factor $\hat{n} = \hat{\alpha}(2-\nu)$ has a simple interpretation. It is the equivalent number of virtual observations supporting the prior since it can be thought as a factor of the metric, and the metric (Fisher information) is additive over independent observations. Using (40) we obtain,

$$\hat{n} = \frac{\alpha(2-\nu)}{1+\alpha\nu I_\delta(\hat{q}:t)} \tag{49}$$

I find equation (49) intrinsically cool in the way it relates geometric, statistical and information concepts. It increases as ν decreases from 1 to 0 so that,

$$\frac{\alpha}{1+\alpha I_\delta(\hat{q}:t)} \leq \hat{n} \leq 2\alpha \tag{50}$$

In general, when $t \in M$, $\hat{n} = \alpha(2-\nu)$ and $\hat{n} = \alpha$ only when $\nu = (1+\alpha I_\delta(\hat{q}:t))^{-1}$. In particular, $\hat{n} = \alpha$ when $t \in M$ and $\nu = 1$ as we shall re-discover below.

As a final comment we mention that one should expect theorems (5, 6, 7) in this section to remain useful (as first order approximations) even when the models are not exactly δ-flat. A complete perturbative analysis for a general curved model M involves higher order covariant derivatives and the modifications will surely contain δ-curvature terms. I hope to be able to continue work on this problem in the near future.

IGNORANCE ON EXPONENTIAL FAMILY MODELS

Exponential family models are intrinsically 1-flat and therefore also 0-flat. However their 0-coordinates do not form a flat manifold. The situation is just like the canonical example of a cylinder imbedded in euclidean flat space. The cylinder curves but it can be smoothly unrolled so its intrinsic geometry is not different from the geometry of the plane.

The following theorem is well known. Given its practical importance we provide a complete proof. It also makes a nice concrete example of flat models and the use of the formulas introduced in the IG overview.

Theorem 8 *Exponential family models are intrinsically* 1*-flat. The natural parameter is* 0*-affine and the expectation parameter is* 1*-affine. Their scalar potentials are the negative free energy and the negative entropy respectively.*

Proof In the natural parameter θ, the normalized likelihood of an exponential family model with vector of sufficient statistics $c(x)$ is given by,

$$p_\theta(x) = e^{\theta \cdot c(x) - \psi(\theta)} \tag{51}$$

where $\psi(\theta)$ is the negative free energy,

$$\psi(\theta) = \log \int e^{c(x)\cdot\theta}\, dx \tag{52}$$

This is in fact the scalar potential satisfying (10) since,

$$g_{ij}(\theta) = -E_\theta(\partial_i\partial_j l) = \partial_i\partial_j\psi(\theta) \tag{53}$$

as it can be readily checked by taking derivatives of the 0-coordinates, i.e. the log likelihood,

$$\begin{aligned} l &= \theta\cdot c(x)\theta - \psi(\theta) && (54)\\ \partial_i l &= c_i(x) - \partial_i\psi(\theta) && (55)\\ \partial_i\partial_j l &= -\partial_i\partial_j\psi(\theta) && (56) \end{aligned}$$

Taking expectations on both sides of (55) we obtain the dual 1-affine coordinates as the expectation parameter,

$$\eta_i = \partial_i\psi(\theta) = E_\theta(c_i(x)) \tag{57}$$

The Legendre transform of (52) gives the other potential as,

$$\begin{aligned} \phi(\eta) &= \theta\cdot\eta - \psi(\theta) && (58)\\ &= \int [c(x)\cdot\theta - \psi(\theta)]e^{c(x)\cdot\theta-\psi(\theta)}\, dx \\ \phi(\eta) &= \int p_\theta(x)\log p_\theta(x)\, dx. && (59) \end{aligned}$$

We notice also that,

$$\delta_{ik} = \frac{\partial \eta_i}{\partial \eta_k} = \frac{\partial \eta_i}{\partial \theta^j}\frac{\partial \theta^j}{\partial \eta_k} = g_{ij}\frac{\partial \theta^j}{\partial \eta_k}$$

where we used (57) and (53) for the last equality. This shows $\partial \theta^j/\partial \eta_k = g^{jk}$ to be the entries of the inverse. Thus, if $\partial^i = \frac{\partial}{\partial \eta_i}$ is the tangent vector in the η-coordinates, we have

$$< \partial^i, \partial_j > = < g^{ik}\partial_k, \partial_j > = g^{ik} g_{kj} = \delta^i_j \tag{60}$$

which is a constant (either 0 or 1) and we have,

$$0 = \partial_k < \partial^i, \partial_j > = < \overset{1}{\nabla}_k \partial^i, \partial_j > + < \partial^i, \overset{1}{\nabla}_k \partial_j > = g^{im}\left(\overset{1}{\Gamma}_{kmj} + \overset{1}{\Gamma}_{kjm}\right)$$

from where we deduce that,

$$\overset{1}{\Gamma}_{kij}(\theta) + \overset{1}{\Gamma}_{kji}(\theta) = 0. \tag{61}$$

Now we have all the ingredients to show that exponential family models are indeed intrinsically 1-flat by showing that in the θ-coordinates all the 1-connection coefficients are zero. We have,

$$\overset{1}{\Gamma}_{ijk} = E_\theta(\partial_i \partial_j l \partial_k l) + E_\theta(\partial_i l \partial_j l \partial_k l) \tag{62}$$

The first term on the right hand side of (62) is shown to be zero by just using (56). The second term is also zero as we now show,

$$\begin{aligned}
E_\theta(\partial_i l \partial_j l \partial_k l) &= \int \partial_i p_\theta \partial_j l \partial_k l \\
&= \int \partial_i p_\theta [\frac{\partial_i \partial_k p_\theta}{p_\theta} - \partial_j \partial_k l] \\
&= \int \partial_i l \partial_j \partial_k p_\theta = \int [c_i(x) - \partial_i \psi(\theta)] \partial_j \partial_k p_\theta \\
&= \int c_i(x) \partial_j \partial_k p_\theta = \partial_j \partial_k \int c_i p_\theta = \partial_i \partial_j \partial_k \psi(\theta) \\
&= \partial_i g_{jk}(\theta) = \partial_i < \partial_j, \partial_k > \\
&= \overset{1}{\Gamma}_{ijk}(\theta) + \overset{1}{\Gamma}_{ikj}(\theta) \\
&= 0.
\end{aligned}$$

Where we have used (61) for the last equality.
Q.E.D.

$(0,\nu)$-ignorant prior on exponential family models

As we have seen, exponential family models are $(0,1)$-flat and therefore everything that was shown above, for general convex and flat models, holds in particular for members of the exponential family. Theorem (6) gives the ignorant prior for $\delta \in \{0,1\}$ and arbitrary $\nu \in [0,1]$ in terms of the two potentials: the negative free energy, and the negative entropy. We have

Theorem 9 *The $(0,\nu)$-ignorant prior at level α is given, for models in the exponential family, by*

$$\pi(\theta|\hat{\eta}, \hat{\alpha}) = [1 + \hat{\alpha}\nu(\psi(\theta) + \phi(\hat{\eta}) - \theta \cdot \hat{\eta})]^{-\frac{1}{\nu}} \tag{63}$$

the scalar potential ψ is the negative free energy (52), the dual scalar potential ϕ is the negative entropy (59), and $\hat{\eta}$ are the 1-coordinates of $\hat{p}$.

Proof (trivial).

When data x is available the natural estimate for $\hat{\eta}$ to be used in (63) is the MLE, i.e. the θ that makes the observed x most likely. If follows from (55) that the MLE is $\hat{\eta} = c(x)$. We have,

Theorem 10 *When M is in the exponential family, the scalar posterior density based on data x and the $(0,\nu)$-prior with $\hat{\eta} = c(x)$ can be written as,*

$$\pi(\theta|x, c(x)) = \frac{e^{-s}}{[1+\hat{\alpha}\nu s]^{1/\nu}} \tag{64}$$

where $s = s(\theta,x) \geq 0$ defines the surface of equiprobability given by,

$$\psi(\theta) + \phi(c(x)) - \theta \cdot c(x) = s \tag{65}$$

Proof Use bayes, theorem (9) and assumption (1).
Q.E.D.
Theorem (10) shows a remarkable property of the $(0,\nu)$-ignorant prior for exponential family models. The prior and posterior equiprobability surfaces are the same. It also suggests a simple algorithm for sampling the posterior: Pick a surface s with probability proportional to the right hand side of (64) then sample uniformly from the surface s defined by (65).

Conjugacy

Until around 1990, conjugate priors for models in the exponential family were essentially the only ones being used in multidimensional problems. The well known PCMCMC revolution changed all that.

With conjugate priors, computation of the posterior distribution reduces to plug-in formulas for the posterior parameters that involve only the sufficient statistics of the observed data. Conjugacy was invented to avoid multidimensional integration over the parameter space. Essentially to keep it simple. As we show here, these priors turn out to be ignorant provided you think of them as scalar densities on the manifold M, i.e. as densities with respect to dV_M (see (6)).

Theorem 11 *The $(0,0)$-ignorant prior over a model M in the exponential family of distributions is conjugate.*

Proof Exponential family models are $(0,1)$-flat. In the standard dual coordinates for flat models the $(0,0)$-ignorant prior is given by (43) when $\nu \to 0$ as,

$$\pi(\theta|\hat{\theta}, \alpha) = e^{(\alpha\hat{\theta}\cdot\theta - \alpha\psi(\theta))} \tag{66}$$

where we have used the fact that $\hat{\alpha} = \alpha$ since $\nu = 0$ (see (40)). On the other hand the normalized likelihood is,

$$p_\theta(x) = e^{(c(x)\cdot\theta - \psi(\theta))}. \tag{67}$$

Where, $c(x)$ is the vector of sufficient statistics and,

$$\psi(\theta) = \log \int e^{\theta\cdot c(x)} dx \tag{68}$$

Thus, the posterior is,

$$\pi(\theta|x, \hat{\theta}, \alpha) = e^{((c(x)+\alpha\hat{\theta})\cdot\theta - (\alpha+1)\psi(\theta))} \tag{69}$$

this is in the same family as the prior and therefore it is conjugate.
Q.E.D.

δ-IGNORANCE

As we have seen, ignorance involves a trade off between uniformity and independence. In this section we show that independence by itself is enough to define a simplified notion of ignorance over a model M. Moreover, the previous notion of (δ,ν)-ignorance at level α coincides with this new version when $(\delta,\nu) = (1,0)$.

The idea is straight forward. Pick one of the available measures of separation (e.g. I_δ)on the space of joint distributions of (x,θ) and measure the δ-deviation between $p(x,\theta) = p(x|\theta)\pi(\theta) \equiv p\pi$ and an arbitrary independent model $t(x)\eta(\theta) \equiv t\eta$. The quantity $I_\delta(p\pi : t\eta)$ measures closeness from the dependent to the specified independent model. We have,

Theorem 12 *For normalized p and t (i.e., when $\int p = \int t = 1$) we have,*

$$I_\delta(p\pi : t\eta) = I_\delta(\pi : \eta) + \int_M I_\delta(p : t)\pi^\delta \eta^{1-\delta} \tag{70}$$

Proof Just write,

$$\begin{aligned} \delta(1-\delta)I_\delta(p\pi : t\eta) &= \int\int \left[\delta p\pi + (1-\delta)t\eta - p^\delta \pi^\delta t^{1-\delta}\eta^{1-\delta}\right] \\ &= \int \left[\delta\pi + (1-\delta)\eta - (\int p^\delta t^{1-\delta})\pi^\delta \eta^{1-\delta}\right] &(71) \\ &= \int \left[\delta\pi + (1-\delta)\eta - \pi^\delta\eta^{1-\delta} + \delta(1-\delta)I_\delta(p:t)\pi^\delta\eta^{1-\delta}\right] &(72) \\ &= \delta(1-\delta)I_\delta(\pi:\eta) + \delta(1-\delta)\int I_\delta(p:t)\pi^\delta\eta^{1-\delta} &(73) \end{aligned}$$

where we have used the hypothesis of normalization to obtain (71) and (72).
Q.E.D.
If α is a positive integer we denote simply by p^α and t^α the corresponding joint distributions of α independent copies of the data vector x, i.e.,

$$p^\alpha = \prod_{i=1}^{\alpha} p(x_i|\theta) \;\; \text{and} \;\; t^\alpha = \prod_{i=1}^{\alpha} t(x_i) \tag{74}$$

Thus, for normalized p and t we have,

$$I_\delta(p^\alpha : t^\alpha) = \frac{1}{\delta(1-\delta)}\left(1 - \left\{\int p^\delta t^{1-\delta}\right\}^\alpha\right) \tag{75}$$

we use the right hand side of (75) as the definition of $I_\delta(p^\alpha : t^\alpha)$ for any $\alpha > 0$. By taking limits, it follows immediately from (75) that,

$$I_\delta(p^\alpha : t^\alpha) = \alpha\, I_\delta(p:t) \;\; \text{for } \delta \in \{0,1\} \tag{76}$$

Using (70) together with (75) we define,

Definition 3 (relative δ-ignorance) *We say that π^* is δ-ignorant at level $\alpha > 0$ relative to η,M and O iff*

$$\pi^* = \arg\min_{\pi\in O} I_\delta(p^\alpha \pi : t^\alpha \eta) \tag{77}$$

We have,

Theorem 13 *When $O = \{M\}$ the δ-ignorant prior at level α relative to η is,*

$$\frac{\pi^*}{\eta} = [1 - \delta(1-\delta)I_\delta(p^\alpha : t^\alpha)]^{\frac{1}{1-\delta}} \tag{78}$$

Proof Use theorem (12) to write the action as $\int \mathscr{L}$ then $\delta\mathscr{L}/\delta\pi = 0$ gives (78).
Q.E.D.
The following theorem is also immediate,

Theorem 14

$$I_1(p^\alpha \pi : t^\alpha \omega) = \mathscr{A}(\pi; \alpha, 1, 0) \tag{79}$$

Proof Replace $\delta = 1$ and $\nu = 0$ in (16).
Q.E.D.
It therefore follows that,

$$\frac{\pi^*}{\omega} = e^{-\alpha \int p \log \frac{p}{t}} \tag{80}$$

is both a $(1,0)$-prior and a 1-prior at level α. Notice that it is also possible to switch the positions of p^α with t^α in (77) in which case the $(0,1)$-prior coincides with the new 0-prior at level α.

EXAMPLE: $(0,0)$-PRIOR FOR MIXTURES OF GAUSSIANS

The general theory presented above has many applications. To demonstrate the utility of our new understanding of statistical ignorance we work out all the details of a concrete example. In this section we compute the $(0,0)$-prior for the parameters of a mixture of k one dimensional gaussians. We assume k to be a given known positive integer. A simple parameterization is then given by $\theta = (\mu, \sigma, \omega)$, where $\mu = (\mu_1, \ldots, \mu_k) \in R^k$ is a k-dim vector of means, $\sigma = (\sigma_1, \ldots, \sigma_k) \in R^k_+$ is a k-dim vector of standard deviations, and $\omega = (\omega_1, \ldots, \omega_k) \in \Delta^{k-1}$ is the vector of mixing weights in the $(k-1)$-dim simplex Δ^{k-1}. Thus, our hypothesis space M is a $3k-1$ dimensional manifold.

The Complete Likelihood

The elements of our model M are the probability distributions for the data vector (x,z) indexed by the vector of parameters θ. We assume that the complete data (x,z) is generated by first picking a label $z \in \{1,2,\ldots,k\}$ with probability vector ω and after that, choosing x by sampling from a gaussian with mean μ_z and standard deviation σ_z. Thus, for $x \in R$, $z \in \{1,2,\ldots,k\}$ and $\theta = (\mu,\sigma,\omega) \in R^k \times R^k_+ \times \Delta^{k-1} \equiv \Theta$ we have,

$$\begin{aligned} p(x,z|\theta) &= p(z|\theta)\ p(x|z,\theta) \\ &= \omega_z \frac{1}{\sqrt{2\pi}\sigma_z} \exp\left\{ -\frac{(x-\mu_z)^2}{2\sigma_z^2} \right\} \end{aligned} \tag{81}$$

Hence,

$$M = \{p(x,z|\theta) : \theta \in \Theta\} \tag{82}$$

Computation of dV on M

In order to obtain an ignorant prior on M, relative to the standard lebesgue measure, we need to first find the volume element of the manifold M. To use (6) we only need to find the Fisher information matrix $I(\theta)$ and its determinant $|I(\theta)|$. We proceed in the standard way by computing first the log likelihoods $l(\theta)$, then their second derivatives $\partial_i\partial_j l$, and finally expected values to obtain $I = (g_{ij}) = (-E_\theta(\partial_i\partial_j l))$. The log likelihoods are obtained from (81) as,

$$\begin{aligned} l &= l(\theta) = \log p(x,z|\theta) \\ &= \log \omega_z - \frac{(x-\mu_z)^2}{2\sigma_z^2} - \log \sigma_z - \frac{1}{2}\log 2\pi \end{aligned}$$

$$= \sum_{j=1}^{k} \left\{ \log \omega_j - \frac{(x-\mu_j)^2}{2\sigma_j^2} - \log \sigma_z - \frac{1}{2}\log 2\pi \right\} 1(z=j) \tag{83}$$

where we have used the indicator function $1(z=j)$ which takes the value one when $z=j$ and zero otherwise.

μ-derivatives

From (83),

$$\frac{\partial l}{\partial \mu_j} = \frac{(x-\mu_j)}{\sigma_j^2} 1(z=j) \text{ and } \frac{\partial^2 l}{\partial \mu_j^2} = -\frac{1(z=j)}{\sigma_j^2} \implies g_{\mu_j \mu_j} = \frac{\omega_j}{\sigma_j^2} \tag{84}$$

Notice that the $g_{\mu,\omega}$ terms are all zero since the first equation above does not explicitly depend on the ω_j parameters. We show below that the $g_{\mu,\sigma}$ are also zero.

σ-derivatives

Again from (83) we compute,

$$\frac{\partial l}{\partial \sigma_j} = \left\{ \frac{(x-\mu_j)^2}{\sigma_j^3} - \frac{1}{\sigma_j} \right\} 1(z=j) \implies \frac{\partial^2 l}{\partial \sigma_j^2} = \left\{ \frac{-3(x-\mu_j)^2}{\sigma_j^4} + \frac{1}{\sigma_j^2} \right\} 1(z=j) \tag{85}$$

$$\implies g_{\sigma_j \sigma_j} = \left(\frac{3}{\sigma_j^2} - \frac{1}{\sigma_j^2} \right) \omega_j = \frac{2\omega_j}{\sigma_j^2} \tag{86}$$

The first equation in the line (85) does not explicitly depend on any of the ω parameters nor does it depend on any other μ_i except for μ_j. Thus, all of the off-diagonal terms, $g_{\sigma_j,\mu_i} = 0$ for $i \neq j$ and $g_{\sigma,\omega} = 0$. We also have,

$$\begin{aligned} g_{\mu_j,\sigma_j} &= E_\theta \left(\frac{\partial l}{\partial \mu_j} \frac{\partial l}{\partial \sigma_j} \right) \\ &= E_\theta \left(\frac{(x-\mu_j)^3}{\sigma_j^5} - \frac{(x-\mu_j)}{\sigma_j^3} \right) \\ &= 0 \end{aligned} \tag{87}$$

ω-derivatives

Once more we compute using (83) and noticing that,

$$\omega_k = 1 - \sum_{j=1}^{k-1} \omega_j \tag{88}$$

and obtain,

$$\frac{\partial l}{\partial \omega_j} = \frac{1(z=j)}{\omega_j} - \frac{1(z=k)}{\omega_k} \text{ for } j = 1,2,\ldots,k-1. \tag{89}$$

Hence, it follows from (89) that all the mixed entries of type $g_{\omega\mu} = g_{\omega\sigma} = 0$. The only non zero entries are,

$$
\begin{aligned}
g_{\omega_i\omega_j} &= E\left\{\left(\frac{1(z=i)}{\omega_i}-\frac{1(z=k)}{\omega_k}\right)\left(\frac{1(z=j)}{\omega_j}-\frac{1(z=k)}{\omega_k}\right)\right\} \\
&= E\left\{\frac{1(z=i)1(z=j)}{\omega_i\omega_j}+\frac{1(z=k)}{\omega_k^2}\right\} \\
&= \frac{\delta_{ij}}{\omega_j}+\frac{1}{\omega_k} \qquad (90)
\end{aligned}
$$

where $\delta_{ij} = 1(i = j)$ is Kronneker's delta.

Fisher Information Matrix

Collecting the above findings we can see that Fisher Information matrix is block diagonal,

$$
I(\theta) = \begin{bmatrix} I_\mu & 0 & 0 \\ 0 & I_\sigma & 0 \\ 0 & 0 & I_\omega \end{bmatrix} \qquad (91)
$$

where I_μ and I_σ are $k \times k$ diagonal matrices with entries,

$$
I_\mu = \begin{bmatrix} \frac{\omega_1}{\sigma_1^2} & & & 0 \\ & \frac{\omega_2}{\sigma_2^2} & & \\ & & \ddots & \\ 0 & & & \frac{\omega_k}{\sigma_k^2} \end{bmatrix} \quad \text{and} \quad I_\sigma = \begin{bmatrix} \frac{2\omega_1}{\sigma_1^2} & & & 0 \\ & \frac{2\omega_2}{\sigma_2^2} & & \\ & & \ddots & \\ 0 & & & \frac{2\omega_k}{\sigma_k^2} \end{bmatrix} \qquad (92)
$$

and I_ω is $(k-1) \times (k-1)$ with all non-zero entries,

$$
I_\omega = \begin{bmatrix} \frac{1}{\omega_1}+\frac{1}{\omega_k} & \frac{1}{\omega_k} & \frac{1}{\omega_k} & \cdots & \frac{1}{\omega_k} \\ \frac{1}{\omega_k} & \frac{1}{\omega_2}+\frac{1}{\omega_k} & \frac{1}{\omega_k} & \cdots & \frac{1}{\omega_k} \\ \vdots & & \ddots & & \\ \frac{1}{\omega_k} & & \cdots & \frac{1}{\omega_k} & \frac{1}{\omega_{k-1}}+\frac{1}{\omega_k} \end{bmatrix} \qquad (93)
$$

Using (6) and (91) we obtain

$$
dV = \sqrt{|I_\mu|\,|I_\sigma|\,|I_\omega|}\,d\mu\,d\sigma\,d\omega \qquad (94)
$$

From (92) and (93) the determinants $|I_\mu|$ and $|I_\sigma|$ are just the product of their diagonal entries. The only complication that remains is the determinant $|I_\omega|$ that we compute as follows. First subtract the second column of (93) from the first column to obtain a matrix identical to that given by (93) except that now the first column has entries $1/\omega_1$, $-1/\omega_2$ and zeros down from the third to the $(k-1)$st row. Call the determinant of this new $(k-1)\times(k-1)$ matrix $D_{k-1}(\omega_1, \omega_2, \ldots, \omega_k)$. We have,

$$
|I_\omega| = D_{k-1}(\omega_1, \omega_2, \ldots, \omega_k) \qquad (95)
$$

since the two matrices differ only by an elementary column operation that does not change the determinant. Now expanding the determinant D_{k-1} about the first column, that has all zeros except for the first two entries, we obtain the recursion,

$$D_{k-1}(\omega_1,\ldots,\omega_k) = \frac{1}{\omega_1} D_{k-2}(\omega_2,\ldots,\omega_k) + (\omega_2\omega_3\ldots\omega_k)^{-1} \tag{96}$$

where the second term on the right hand side of (96) is obtained by multiplying $1/\omega_2$ times the determinant of the matrix obtained from (93) by erasing the first column and the second row. This co-factor is brought into an almost upper triangular matrix with main diagonal entries given by, $1/\omega_k, 1/\omega_3, 1/\omega_4, \ldots, 1/\omega_{k-1}$ by subtracting the previous to last column from the last column, the second to last from the previous to last, etc, until the first column. The determinant of this almost upper triangular matrix is shown to be the product of the main diagonal by expanding about its first row that contains only zeros except for the first entry $1/\omega_k$.

From the recursion (96) we easily show,

Theorem 15 *For* $k = 2, 3, \ldots$

$$D_{k-1}(\omega_1,\ldots,\omega_k) = \frac{\sum_{j=1}^{k}\omega_j}{\prod_{j=1}^{k}\omega_j} \tag{97}$$

Proof By induction on k. It is true for $k = 2$ since,

$$D_1(\omega_1,\omega_2) = \left(\frac{1}{\omega_1} + \frac{1}{\omega_2}\right) = \frac{\omega_1+\omega_2}{\omega_1\omega_2} \tag{98}$$

If it is true for k then using (96) and the induction hypothesis we have,

$$\begin{aligned} D_k(\omega_1,\ldots,\omega_{k+1}) &= \frac{1}{\omega_1} D_{k-1}(\omega_2,\ldots,\omega_{k+1}) + (\omega_1\cdots\omega_{k+1})^{-1} \\ &= \frac{1}{\omega_1}\frac{\sum_{j=2}^{k+1}\omega_j}{\prod_{j=2}^{k+1}\omega_j} + \frac{1}{\prod_{j=2}^{k+1}\omega_j} \\ &= \frac{\sum_{j=1}^{k+1}\omega_j}{\prod_{j=1}^{k+1}\omega_j} \end{aligned} \tag{99}$$

Q.E.D.

Finally, from (92), (93) and (97) we write,

$$dV = \prod_{j=1}^{k} \frac{\omega_j^{1/2}}{\sigma_j^2}\, d\mu\, d\sigma\, d\omega \tag{100}$$

Entropy

Assuming $t \in M$ the unnormalized density (with respect to lebesgue measure $d\theta$) of the $(0,0)$-prior is,

$$\pi(\theta|\alpha,\hat{\theta})\, d\theta = e^{-\alpha\, I(\hat{\theta}:\theta)}\, dV \tag{101}$$

where,

$$I(\hat{\theta}:\theta) = I_0(\theta:\hat{\theta}) = E_{\hat{\theta}}\left\{\log\frac{p(x,z|\hat{\theta})}{p(x,z|\theta)}\right\} \tag{102}$$

The distribution in M that represents truth has parameter,

$$\hat{\theta} = (m,s,w) = (m_1,\ldots,m_k,s_1,\ldots,s_k,w_1,\ldots,w_k) \tag{103}$$

We compute the entropy (102) by first conditioning on a value of z and then taking expectation over z, i.e.,

$$I(\hat{\theta}:\theta) = E^{z}E^{x|z}\left\{\log\frac{w_z}{\omega_z} + \log\frac{N(x;m_z,s_z)}{N(x;\mu_z,\sigma_z)}\right\} \tag{104}$$

$$= \sum_{j=1}^{k} w_j\left\{\log\frac{w_j}{\omega_j} + \log\frac{\sigma_j}{s_j} + \frac{(\mu_j-m_j)^2}{2\sigma_j^2} + \frac{s_j^2}{2\sigma_j^2} - \frac{1}{2}\right\} \tag{105}$$

where in (104) $N(x;a,b)$ denotes the usual density for a gaussian distribution with mean a and standard deviation b. Combining (105) with the previously obtained formula for the volume element (100) we obtain,

$$\pi(\theta|\alpha,\hat{\theta}) = \prod_{j=1}^{k} \exp\left\{-\frac{\alpha w_j}{2\sigma_j^2}(\mu_j-m_j)^2\right\} \cdot\sigma_j^{-\alpha w_j-2}\exp\left\{-\frac{\alpha w_j}{2\sigma_j^2}s_j^2\right\} \cdot\omega_j^{\alpha w_j+\frac{1}{2}} \tag{106}$$

Equation (106) shows, I believe correctly for the first time, the $(0,0)$-prior for mixtures of univariate gaussians. It is clearly integrable and remarkably simple. The mixing weights ω follow a Dirichlet, the vector of standard deviations σ has independent components that follow inverse-Gamma distributions, and the vector of means μ, conditionally on σ, also has independent components that follow gaussian distributions. Independent samples from this distribution can be obtained exactly, without the need of asymptotic Gibbs sampling, with very efficient procedures in the public domain.

Posterior

Model (82) becomes very useful when assuming that the actual observed data is $D = x^n = (x_1,\ldots,x_n)$ such that $(x_1,z_1),\ldots,(x_n,z_n)$ are independent samples from a distribution in (82) but the vector of labels $z^n = (z_1,\ldots,z_n)$ is not available, it is missing. The accurate identification of the vector of missing labels is often of great practical importance with applications scattered all over the spectrum of modern science and technology. Estimation of the missing labels, as always, is done by simply wiggling the bayesian wand, i.e., applying bayes theorem. Bayes theorem gives the chances of what we don't know given what we do know. There is a caveat: A model (i.e. a hypothesis space) and a prior on the model are needed. In the case of mixtures of univariate gaussians we now have all the ingredients.

The Complete Posterior

If we knew the vector of labels z^n then the unnormalized complete posterior would be,

$$\pi(\theta|x^n,z^n,\alpha,\hat{\theta}) = \pi(\theta|\alpha,\hat{\theta})\prod_{i=1}^{n}\frac{\omega_{z_i}}{\sigma_{z_i}}\exp\left\{-\frac{(x_i-\mu_{z_i})^2}{2\sigma_{z_i}^2}\right\} \tag{107}$$

Let k_j be the total number of labels that are equal to j, i.e.,

$$k_j = \sum_{i=1}^{n} 1(z_i = j) \tag{108}$$

Grouping common labels together we can simplify (107) as,

$$\pi(\theta|x^n,z^n,\alpha,\hat{\theta}) = \pi(\theta|\alpha,\hat{\theta})\prod_{j=1}^{k}\frac{\omega_j^{k_j}}{\sigma_j^{k_j}}\exp\left\{-\frac{1}{2\sigma_j^2}\sum_{z_i=j}(x_i-\mu_j)^2\right\} \tag{109}$$

Bringing in the prior (106) we obtain the complete posterior as,

$$\begin{aligned}\pi(\theta|x^n,z^n,\alpha,\hat{\theta}) = \prod_{j=1}^{k} \quad & \exp\left\{-\frac{1}{2\sigma_j^2}[\alpha w_j(\mu_j-m_j)^2+\sum_{z_i=j}(x_i-\mu_j)^2]\right\} \\ & \cdot\sigma_j^{-\alpha w_j-2-k_j}\exp\left\{-\frac{\alpha w_j s_j^2}{2\sigma_j^2}\right\} \\ & \cdot\omega_j^{\alpha w_j+k_j+\frac{1}{2}}\end{aligned} \tag{110}$$

Again, independent samples of θ with distribution (110) can be obtained exactly, without the need of Gibbs iterations. We factorize the complete posterior distribution of θ as,

$$(\theta) = (\mu|\sigma)\ (\sigma)\ (\omega) \tag{111}$$

Thus, to sample a vector $\theta = (\mu,\sigma,\omega)$ from (110) we first produce ω with,

$$\omega \sim \mathscr{D}(\alpha w_1+k_1+\frac{3}{2},\alpha w_2+k_2+\frac{3}{2},\ldots,\alpha w_k+k_k+\frac{3}{2}) \tag{112}$$

where $\mathscr{D}$ stands for the Dirichlet distribution. Then we sample σ from its unconditional distribution obtained from (110) by integrating over μ. We collect the resulting distribution in the following theorem,

Theorem 16 *In the complete posterior (110) the distribution of the vector σ is,*

$$\pi(\sigma|x^n,z^n,\alpha,\hat{\theta}) = \prod_{j=1}^{k} \sigma_j^{-\alpha w_j-k_j-1}\ \exp\left\{\frac{-1}{2\sigma_j^2}[\alpha w_j s_j^2+A_j]\right\} \tag{113}$$

where,

$$A_j = k_j\ \left[V_j^2+\frac{\alpha w_j}{\alpha w_j+k_j}(\bar{x}_j-m_j)^2\right] \tag{114}$$

and

$$V_j^2 = \frac{1}{k_j}\sum_{z_i=j}(x_i-\bar{x}_j)^2 = \bar{x_j^2}-(\bar{x}_j)^2 \tag{115}$$

and we are using the notation,

$$\bar{x}_j = \frac{1}{k_j}\sum_{z_i=j}x_i \tag{116}$$

Proof Tedious but straight forward. We have,

$$\pi(\sigma|x^n,z^n,\alpha,\hat{\theta}) = \int \pi(\theta|x^n,z^n,\alpha,\hat{\theta})\ d\mu$$

Looking at (110) we see that we need to evaluate the integrals,

$$J_j = \int \exp\left\{-\frac{1}{2\sigma_j^2}[\alpha w_j(\mu_j-m_j)^2+\sum_{z_i=j}(x_i-\mu_j)^2]\right\}\ d\mu_j$$

Expanding the squares, collecting terms involving μ_j, completing back the square, performing the gaussian integral, and simplifyiiing we obtain,

$$J_j = \frac{\sigma_j}{\sqrt{\alpha w_j+k_j}}\ \exp\left\{\frac{-1}{2\sigma_j^2}A_j\right\} \tag{117}$$

Q.E.D.
Also,

Theorem 17 *With the notation of theorem (16). For $j = 1,2,\ldots,k$ let,*

$$\sigma_j = \sqrt{\frac{\alpha w_j s_j^2 + A_j}{2\tau_j}} \tag{118}$$

If $\tau_1, \tau_2, \ldots, \tau_k$ are chosen independently with distributions,

$$\tau_j \sim \mathscr{G}(\frac{\alpha w_j + k_j}{2}) \tag{119}$$

where $\mathscr{G}$ denotes a Gamma distribution. Then the vector $\sigma = (\sigma_1, \ldots, \sigma_k)$ follows the distribution (113).

Proof Just change the variables. **Q.E.D.**

Finally, once the vector σ is given, we sample μ from the conditional distribution,

$$\pi(\mu|\sigma, x^n, z^n, \alpha, \hat{\theta}) = \prod_{j=1}^{k} N(\mu_j; \tilde{\mu}_j, \tilde{\sigma}_j) \tag{120}$$

where,

$$\tilde{\mu}_j = \frac{\alpha w_j m_j + k_j \bar{x}_j}{\alpha w_j + k_j} \quad \text{and} \quad \tilde{\sigma}_j = \frac{\sigma_j^2}{\alpha w_j + k_j} \tag{121}$$

This is the standard computation of the posterior distribution of a gaussian mean when the likelihood and the prior are both gaussian.

Gibbs Sampler

To sample the actual posterior $\pi(\theta|x^n, \alpha, \hat{\theta})$ we need to sample from the joint distribution of (θ, z^n) and discard the labels z^n. We use a Gibbs sampler to eventually sample vectors, (θ, z^n) by iterating samples from the two conditional distributions: The complete posterior $\pi(\theta|x^n, z^n, \alpha, \hat{\theta})$, as indicated above, and the posterior distribution for the labels, given by

$$\begin{aligned} p(z^n|\theta, x^n, \alpha, \hat{\theta}) &\propto p(x^n, z^n|\theta, \alpha, \hat{\theta}) = \prod_{i=1}^{n} p(x_i, z_i|\theta) \\ &\propto \prod_{i=1}^{n} \frac{\omega_{z_i}}{\sigma_{z_i}} \exp\left\{ -\frac{(x_i - \mu_{z_i})^2}{2\sigma_{z_i}} \right\} \end{aligned} \tag{122}$$

Equation (122) shows that conditionally on (θ, x^n), the labels $z_1, z_2, \ldots, z_n$ are independent with probabilities, for $j = 1,2,\ldots,k$ given by

$$P[z_i = j|\theta, x^n] = \frac{1}{c} \frac{\omega_j}{\sigma_j} \exp\left\{ -\frac{(x_i - \mu_j)^2}{2\sigma_j^2} \right\} \tag{123}$$

where $c > 0$ is the normalizing constant. Notice also, that estimation of the labels can be done by using the same joint samples of (θ, z^n) but now discarding θ instead of z^n. The observed z^n will be samples from $p(z^n|x^n, \alpha, \hat{\theta})$.

MODEL SELECTION

Having been presented with two alternative models M_0 and M, possibly of different dimensions, for the same vector $x^n = (x_1, \ldots, x_n)$ of observations, when should we choose M over M_0?. This is a decision problem. Once we choose a loss function L and prior probabilities for the two alternative hypotheses M_0 and M the optimal decision is the one given by bayes rule: *choose the model that minimizes the expected loss conditional on whatever is known.*

The optimal values of the invariant actions for ignorance that we have found in this paper provide the decision maker with a natural class of loss functions. For fix values of α, δ and ν we can quantify the loss of deciding to use model M when the true distribution is t by,

$$L(t, M; \alpha, \delta, \nu) = \mathscr{A}_M^* = \mathscr{A}_M(\pi^*; \alpha, \delta, \nu, t, O) \tag{124}$$

Thus, we have a collection of rankings for models. In the absence of any observations, when the true distribution is t we choose model M over M_0 when $\mathscr{A}_M^* < \mathscr{A}_{M_0}^*$. We have possibly different rankings for different choices of the parameters α, δ, and ν. But these are only theoretical, ideal rankings, for they depend on the unknown true distribution t. When a vector x^n of observations is available it seems reasonable to estimate the unknown t by $\hat{t}(x^n)$, e.g. with the M.L.E., in each model. The obvious question is: How good is that?. We show below that the case $\delta = \nu = 0$ is asymptotically equivalent to the plain bayesian ranking, i.e. the one that chooses the model with the highest posterior probability.

In fact, given two alternative models M_0 and M, observed data x^n, prior probabilities $\mathbf{P}(M_0)$ and $\mathbf{P}(M)$ and prior distributions on the models, π_M and π_{M_0} the plain bayesian ranking chooses M over M_0 iff

$$\mathbf{P}(M|x^n) > \mathbf{P}(M_0|x^n) \tag{125}$$

By bayes theorem the posterior probabilities are given by,

$$\mathbf{P}(M|x^n) = \frac{1}{Z} p(x^n|M)\mathbf{P}(M) \tag{126}$$

where, Z is a normalizing constant independent of the model M. The marginal likelihood can be expanded as,

$$\begin{aligned}
p(x^n|M) &= \int p(x^n, \theta|M)\, dV(\theta) \\
&= \int p(\theta|M) p(x^n|\theta, M)\, dV(\theta) \\
&= \int_M \pi_M(p) \prod_{i=1}^n p(x_i)\, dV(p) && (127) \\
&= \int_M \pi_M(p) e^{nL_n(p)}\, dV && (128)
\end{aligned}$$

and by the standard central limit theorem,

$$\begin{aligned}
L_n(p) &= \frac{1}{n} \sum_{i=1}^n \log p(x_i) \\
&= \int t(x) \log p(x)\, dx + \frac{\sigma_t}{\sqrt{n}} N(0,1) + o(n^{-1/2}) \\
&= -I_0(p:t) + \int t(x) \log t(x)\, dx + O(n^{-1/2}) && (129)
\end{aligned}$$

Thus, when $\mathbf{P}(M_0) = \mathbf{P}(M)$ replacing (129) into (128) we obtain that the plain bayesian ranking will choose among the alternative models the M that minimizes,

$$-\log \mathbf{P}(M|x^n) = -\log \int_M \pi_M(p) e^{-n(I_0(p:t) + O(n^{-1/2}))}\, dV + (\text{constant}) \tag{130}$$

This shows that for large n, neglecting lower order terms and the irrelevant additive constant, the model M with the highest posterior probability is the one that minimizes,

$$\mathscr{L}_M = -\log \int_M \pi_M(p) e^{-nI_0(p:t)}\, dV \tag{131}$$

When M has finite volume $V < \infty$ the most ignorant prior on M is the uniform (relative to dV) given by $\pi_M(p) = 1/V$ and we have,

$$\mathscr{L}_M = \mathscr{A}_M^*|_{\alpha=n,\delta=\nu=0} \tag{132}$$

When M has infinite volume, we replace Jeffreys prior with an ignorant prior $\pi_M = [1+\alpha\nu I_\delta]^{-1/\nu}$ with $\alpha \to \alpha_{\min}$. The resulting objective function $\mathscr{L}_M$ can still be thought as an $\mathscr{A}_M^*$ if we replace in (14) the uniform ω with its approximation π_M.

Ignorant Razors

The previous discussion suggests interpreting the optimal values of the actions for ignorance $\mathscr{A}_M^*$ as measures of model accuracy and complexity. A general family of Occam's razors is given by,

$$\mathscr{R}(M) = \frac{\int_M e^{-\bar{\alpha} I_{\bar{\delta}}(p:t)} [1+\alpha\nu I_\delta(p:t)]^{-1/\nu}\, dV}{\int_M [1+\alpha\nu I_\delta(p:t)]^{-1/\nu}\, dV} \tag{133}$$

The case $\nu = 0$ has the simplest form,

$$\mathscr{R}(M) = \frac{\int_M e^{-(\alpha+\beta) I_\delta(p:t)}\, dV}{\int_M e^{-\alpha I_\delta(p:t)}\, dV} \tag{134}$$

obtained when $\bar{\alpha} = \beta$ and $\bar{\delta} = \delta$. In this case,

$$\mathscr{A}_M^* = -\log \frac{Z_M(\alpha+\beta)}{Z_M(\alpha)} \tag{135}$$

where,

$$Z_M(\alpha) = \int_M e^{-\alpha I_\delta(p:t)}\, dV \tag{136}$$

By imitating Balasubramanian's paper we produce a perturbative expansion of the normalizing constants of type (136) as $\alpha \to \infty$. We should expect to see the effects of the δ-geometry explicitly in this expansion since the δ-connection coefficients and its derivatives are obtainable from the Eguchi relations introduced in the IG overview.

With the help of a parameterization we write,

$$Z_M(\alpha) = \int e^{-\alpha E(\theta)}\, d\theta \tag{137}$$

where,

$$\begin{aligned} E(\theta) &= I_\delta(\theta:t) - \alpha^{-1} \log \sqrt{\det I(\theta)} && (138)\\ &= J(\theta) - \alpha^{-1} F(\theta) && (139) \end{aligned}$$

Let us now define,

$$\overset{\delta}{\theta} = \arg\min_\theta I_\delta(\theta:t) \tag{140}$$

Now recall that α is interpreted as a continuous number of apriori independent observations from the true distribution t (see (79)). Hence, the central limit theorem provides a justification for an α-dependent scaling of the variable of integration. Let $r = \alpha^{-1/2}$ and define the new variable of integration,

$$\phi = \sqrt{\alpha}(\theta - \overset{\delta}{\theta}) \iff \theta = \overset{\delta}{\theta} + r\phi \tag{141}$$

so that,

$$Z_M(\alpha) = \alpha^{-\kappa/2} \int e^{-\alpha E(\overset{\delta}{\theta}+r\phi)}\, d\phi \tag{142}$$

where κ is the dimension of the manifold model M. Considering the exponent as a function of $r > 0$ and expanding the functions J and F in a Taylor series about $r = 0$ we get,

$$\begin{aligned} E(\overset{\delta}{\theta} + r\phi) &= E(\overset{\delta}{\theta}) + \sum_{j=2}^{\infty} \frac{r^j}{j!} J_{\mu_1\cdots\mu_j} \phi^{\mu_1} \cdots \phi^{\mu_j} - r^2 \sum_{i=1}^{\infty} \frac{r^i}{i!} F_{\mu_1\cdots\mu_i} \phi^{\mu_1} \cdots \phi^{\mu_i} \\ &= E(\overset{\delta}{\theta}) + \frac{\alpha^{-1}}{2!} J_{\mu_1\mu_2} \phi^{\mu_1} \phi^{\mu_2} + \alpha^{-1} G(\phi) \end{aligned} \tag{143}$$

where,

$$J_{\mu_1\cdots\mu_j} = \partial_{\mu_1} \cdots \partial_{\mu_j} I_\delta(\theta : t)\Big|_{\theta=\overset{\delta}{\theta}} \tag{144}$$

$$F_{\mu_1\cdots\mu_i} = \partial_{\mu_1} \cdots \partial_{\mu_i} F(\theta)\Big|_{\theta=\overset{\delta}{\theta}} \tag{145}$$

$$G(\phi) = \sum_{i=1}^{\infty} \alpha^{-i/2} \left[\frac{1}{(i+2)!} J_{\mu_1\cdots\mu_{i+2}} \phi^{\mu_1} \cdots \phi^{\mu_{i+2}} - \frac{1}{i!} F_{\mu_1\cdots\mu_i} \phi^{\mu_1} \cdots \phi^{\mu_i} \right] \tag{146}$$

and we have used the fact that when $j = 1$ the rhs of (144) is zero. It is easy to verify that equations (144) and (145) are directly connected to the δ-geometry. In fact, we have

$$J_{ij}\Big|_{t=\overset{\delta}{\theta}} = g_{ij}(\overset{\delta}{\theta}) \tag{147}$$

$$J_{ijk}\Big|_{t=\overset{\delta}{\theta}} = \overset{\delta}{\Gamma}_{ijk}(\overset{\delta}{\theta}) + \Gamma_{jki}(\overset{\delta}{\theta}) + \Gamma_{kij}(\overset{\delta}{\theta}) \tag{148}$$

$$F_j = \Gamma^i_{ji}(\overset{\delta}{\theta}) \tag{149}$$

$$F_{jk} = \Gamma^i_{ji,k}(\overset{\delta}{\theta}) = \Gamma^i_{ki,j}(\overset{\delta}{\theta}) \tag{150}$$

Equations (147) and (148) are an alternative form of the Eguchi relations that can be checked by straight forward computation of the derivatives. Connection coefficients appearing without a δ on top are the metric (Levi-Civita) $\delta = 1/2$ case, and partial derivatives in (150) are indicated by the sub index following a comma. Obviously (150) follows from (149) but (149) is not obvious. It is however well known and often left as an exercise in general relativity text books. For the sake of completeness we provide a proof in the following theorem.

Theorem 18 *For any metric g_{ij} with associated connection coefficients Γ^k_{ij} we have,*

$$\partial_k \log \sqrt{\det(g_{ij})} = \Gamma^l_{kl} \tag{151}$$

Proof The metric $G = (g_{ij})$ is symmetric and positive definite. Thus, there exists an orthogonal matrix P, i.e. $P^T P = Id$, with $G = P^T DP$ and D a diagonal matrix with positive entries λ_i. Hence,

$$\begin{aligned} \partial_k \log \sqrt{\det(g_{ij})} &= \frac{1}{2} \partial_k \log\, \det G = \frac{1}{2} \partial_k \left(\sum_i \log \lambda_i \right) \\ &= \frac{1}{2} \sum_i \frac{1}{\lambda_i} \partial_k \lambda_i = \frac{1}{2} \mathbf{tr}\, D^{-1} \partial_k D \\ &= \frac{1}{2} \mathbf{tr} \left(P^T D^{-1} P P^T \partial_k D P \right) = \frac{1}{2} \mathbf{tr}\, G^{-1} \partial_k G \end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{2} g^{im} g_{mi,k} \\
&= g^{im}\frac{1}{2}(g_{im,k}+g_{mk,i}-g_{ki,m}) + \frac{1}{2}\left\{g^{im}(g_{ki,m}-g_{mk,i})\right\} \\
&= \Gamma^i_{ki} \qquad (152)
\end{aligned}
$$

where we have used the cyclic property of the trace, the definition of the connection coefficients $\Gamma_{ijk} = \frac{1}{2}(g_{jk,i} + g_{ki,j} - g_{ij,k})$ and the fact that the expression within curly brackets is zero. This last fact follows from differentiating the constant identity $\partial_j(g^{im}g_{mk}) = 0$ so that,

$$
g^{im}(g_{ki,m}-g_{mk,i}) = (-g^{l_1l_2}_{,l_1}g_{kl_2}) - (-g^{l_1l_2}_{,l_1}g_{kl_2}) = 0
$$

Q.E.D.

Now from the fact that,

$$
\partial_{h_{\mu_1}}\cdots\partial_{h_{\mu_k}} e^{h_i\phi^i} = e^{h_i\phi^i}\phi^{\mu_1}\cdots\phi^{\mu_k} \qquad (153)
$$

it follows, by writing the exponential as a power series, that

$$
e^{-G(\phi)} = e^{-G(\nabla_h)} e^{h_i\phi^i}\Big|_{h=0} \qquad (154)
$$

where $\nabla_h \equiv (\partial_{h_1},\ldots,\partial_{h_\kappa})$, and $G(\phi)$ is any multivariate series like (146). Thus, replacing (143) into (142) using (154) and moving the derivatives outside the integral, we obtain, [4]

$$
\begin{aligned}
Z_M(\alpha) &= \alpha^{-\kappa/2} e^{-\alpha I_\delta(\overset{\delta}{\theta}:t)} \sqrt{\det I(\overset{\delta}{\theta})}\; e^{-G(\nabla_h)} \int e^{-\frac{1}{2}J_{ij}\phi^i\phi^j + h_k\phi^k}\, d\phi\Big|_{h=0} \qquad (155) \\
&= \alpha^{-\kappa/2} e^{-\alpha I_\delta(\overset{\delta}{\theta}:t)} \sqrt{\det I(\overset{\delta}{\theta})}\; e^{-G(\nabla_h)} \left[\left(\frac{(2\pi)^\kappa}{\det J}\right)^{1/2} \exp(\frac{1}{2}h_iJ^{ij}h_j)\right]\Bigg|_{h=0} \qquad (156)
\end{aligned}
$$

where the gaussian integral (155) was computed by noticing that if we let $m^i = h_kJ^{ki}$ then we can write the integrand as,

$$
\begin{aligned}
\exp\left\{-\frac{1}{2}J_{ij}\phi^i\phi^j + h_k\phi^k\right\} &= \exp\left\{\frac{1}{2}J_{ij}m^im^j\right\}\exp\left\{-\frac{1}{2}(\phi^i-m^i)J_{ij}(\phi^j-m^j)\right\} \\
&= \exp\left\{\frac{1}{2}J^{ij}h_ih_j\right\}\exp\left\{-\frac{1}{2}(\phi^i-m^i)J_{ij}(\phi^j-m^j)\right\} \qquad (157)
\end{aligned}
$$

Let us re-write (146) as,

$$
G(\phi) = -\sum_{i=1}^{\infty} \alpha^{-i/2}A_i(\phi) \qquad (158)
$$

Using (158) and (156) we obtain an expansion of $Z_M(\alpha)$ to all orders. For example, if we neglect terms of order α^{-2} and smaller we have,

$$
\begin{aligned}
e^{-G(\phi)} &= 1 - G(\phi) + \frac{G^2(\phi)}{2!} - \frac{G^3(\phi)}{3!} + \cdots \\
&= 1 + \left[\alpha^{-1/2}A_1(\phi) + \alpha^{-1}A_2(\phi) + \alpha^{-3/2}A_3(\phi)\right] +
\end{aligned}
$$

[4] This is a very powerful trick that I learned from Balasubramanian's paper. Notice however that equation (41) in that paper is missing a factor $-1/2$ in the exponent.

$$\frac{1}{2}\left\{\alpha^{-1}A_1^2(\phi)+2\alpha^{-3/2}A_1(\phi)A_2(\phi)\right\}+$$
$$\frac{1}{6}\alpha^{-3/2}A_1^3(\phi)+O(\alpha^{-2}) \tag{159}$$

the terms within square brackets correspond to the terms in $-G(\phi)$ of order lower than α^{-2}, those within curly brackets are from $G^2(\phi)$, and the last term is, the only other contribution of order lower than α^{-2}, from $G^3(\phi)/3!$. It can be shown that the terms with an odd number of phi's, like the terms of orders $\alpha^{-1/2}$ and $\alpha^{-3/2}$ will end up not contributing anything when (159) is transformed into a differential operator and used in (156). Thus, for model selection, the choice of δ-geometry will appear in the terms of order α^{-1}. Let us use the following convenient notation,

$$f = \exp\left(\frac{1}{2}J^{\mu\nu}h_\mu h_\nu\right) \tag{160}$$
$$m^j = J^{\mu j}h_\mu \tag{161}$$

Thus, using (159) we obtain,

$$\begin{aligned} e^{-G(\nabla_h)}f\Big|_{h=0} &= 1+\alpha^{-1}\left\{A_2(\nabla_h)+\frac{1}{2}A_1^2(\nabla_h)\right\}f\Big|_{h=0}+O(\alpha^{-2}) \\ &= 1+\alpha^{-1}\left\{[\frac{1}{2}\Gamma^\mu_{i\mu,j}T^{ij}-\frac{1}{4!}J_{ijkl}T^{ijkl}]+\right. \\ &\left.\frac{1}{2}[\Gamma^\mu_{i\mu}\Gamma^\nu_{j\nu}T^{ij}-\frac{2}{3!}\Gamma^\mu_{i\mu}J_{jkl}T^{ijkl}+\left(\frac{1}{3!}\right)^2 J_{ijk}J_{lmn}T^{ijklmn}]\right\}+O(\alpha^{-2}) \end{aligned} \tag{162}$$

where the first term within square brackets is from A_2 and the second is from A_1^2 and,

$$T^{\mu_1\ldots\mu_i} = \partial_{h_1}\cdots\partial_{h_i}f\big|_{h=0} \tag{163}$$

Applying the formulas $\partial_j f = fm^j$ and $\partial_i m^j = J^{ij}$ recursively we obtain, after a long but straight forward computation, that

$$T^{ij} = J^{ij} \tag{164}$$
$$T^{ijkl} = J^{ij}J^{kl}+J^{ik}J^{jl}+J^{il}J^{jk} \tag{165}$$
$$T^{ijklmn} = J^{ij}J^{kl}J^{mn}+(..11\text{ similar terms}) \tag{166}$$

Substituting into (162) and simplifying we find,

$$e^{-G(\nabla_h)}f\Big|_{h=0} = 1+\frac{\alpha^{-1}}{24}J^{ij}S_{ij}+O(\alpha^{-2}) \tag{167}$$

where,

$$S_{ij} = 12\Gamma^\mu_{i\mu,j}-3J_{ijkl}J^{kl}-3\Gamma^\mu_{i\mu}\Gamma^\nu_{j\nu}-12\Gamma^\mu_{i\mu}J_{jkl}J^{kl}+4J_{ijk}J_{lmn}J^{kl}J^{mn} \tag{168}$$

The above tensor is closely related to the Ricci tensor. As it can be readily checked (see MTW p.222, (8.51a,b)), the Ricci tensor can be written as,

$$R_{ij} = -\Gamma^\mu_{i\mu,j}+\Gamma^\mu_{ij,\mu}+\Gamma^\nu_{\mu\nu}\Gamma^\mu_{ij}-\Gamma^\mu_{\nu i}\Gamma^\nu_{j\mu} \tag{169}$$

We can therefore write (168) in terms of the Ricci tensor as,

$$S_{ij} = -12(R_{ij}+T_{ij}) \tag{170}$$

where the new tensor T_{ij} collects all the left-over terms. We can regard the above equation as the defining equation for T_{ij}. We expect T_{ij} to have zero trace when the model contains the true distribution. More detailed computations will be the subject of an upcoming paper. With this notation we can write once again (167) in the following more useful form,

$$e^{-G(\nabla_h)} f\Big|_{h=0} = 1 - \frac{\alpha^{-1}}{2} J^{ij}[R_{ij}+T_{ij}] + O(\alpha^{-2}) \tag{171}$$

From (171), (159) and (156) we get,

$$\begin{aligned}
-\log Z_M(\alpha) \quad &= \quad \alpha\, I_\delta(\overset{\delta}{\theta} : t) && (172)\\
&+\frac{\kappa}{2}\log\frac{\alpha}{2\pi} && (173)\\
&+\frac{1}{2}\log\left(\frac{\det J(\overset{\delta}{\theta})}{\det I(\overset{\delta}{\theta})}\right) && (174)\\
&-\log\left(1-\frac{\alpha^{-1}}{2}J^{ij}[R_{ij}+T_{ij}]+O(\alpha^{-2})\right) && (175)
\end{aligned}$$

This formula provides a new measure of complexity for models whose volumes have been normalized. If the models to be compared occupy different volumes in the space of distributions then we need to add $\log V_M$ as a fifth term to the above formula. The first three terms for the case $\delta = 0$ have been previously obtained in connection to the so called Minimum Description Length (MDL) principle, attributed to Rissanen. The general δ case, and specially the fourth term involving the scalar built out of the curvature connection coefficients, were previously unknown.

Each of the four terms of the above expansion has a clear interpretation. The leading term (172) measures the accuracy of the model M in terms of the δ-distance from the model to the true distribution and increases linearly with the amount of prior information (data) α. This first term ensures asymptotic correctness (consistency) as $\alpha \to \infty$, i.e. as more and more prior information is available. The second term (173) penalizes models of high dimension and encodes parsimony relative to the amount of available prior information. It is linear in the dimension κ but logarithmic in α. This term was first discovered by Schwartz in connection with the so called Bayesian Information Criterion (BIC). This term does not depend on δ in any way. The third term, constant in α, encodes what Balasubramanian calls the *naturalness* or robustness of the model. It is a penalty on models that have few indistinguishable points that are close to the true one, e.g. when the model looks highly curved from the point of view of the true distribution. The fourth term (175), let us call it S_4, is the most remarkable for it has not previously appeared in the literature. Expanding the logarithm and neglecting the lower order terms we have,

$$S_4 = \frac{\alpha^{-1}}{2} J^{ij}\,[R_{ij}+T_{ij}] \tag{176}$$

We remind the reader that the above quantities are evaluated at a special point in M, namely the closest point to t in the δ-deviation distance. Also, J depends on the true distribution t as well.

When $M \ni t$

When the model M is sufficiently large or informative, so that we can assume that it contains, or at least closely approximates, the true distribution the action for model selection takes a simpler form.

In fact when $t = \overset{\delta}{\theta}$ we have from (147) that $J^{ij} = g^{ij}$ and the terms (172) and (174) become zero. Hence, as $\alpha \to \infty$

$$-\log Z_M(\alpha) = \frac{\kappa}{2}\log\frac{\alpha}{2\pi} - \log\left(1-\frac{\alpha^{-1}}{2}g^{ij}[R_{ij}+T_{ij}]\Big|_{\theta=\overset{\delta}{\theta}}\right) + O(\alpha^{-2}) \tag{177}$$

Now consider π to be any prior scalar density that peaks at $\overset{\delta}{\theta}$ as $\alpha \to \infty$. For example a natural choice is given by the ignorant $(\delta, 0)$-prior,

$$\pi(\theta) = e^{-\alpha \, I_{\delta}(\theta : \overset{\delta}{\theta})} \tag{178}$$

For any such prior π the action for model selection (177) can be thought as an approximation to the global action given mod α^{-2} by:

$$-\log Z_M(\alpha) = \frac{\kappa}{2}\log\left(\frac{\alpha}{2\pi}\right) \; - \; \log\left(1 - \frac{\alpha^{-1}}{2}\int_M (R+T)\pi \, dV\right) \tag{179}$$

where R is the Ricci scalar and T is the trace of T_{ij} plus any term, which may depend on π, that becomes negligible around the peak. Notice that when comparing models with the same number of parameters κ, the first term in (179) becomes an irrelevant constant and the action becomes equivalent to the integral, i.e., the prior expectation of $(R+T)$.

The practical utility of the new ranking procedures introduced in (179), (177) or the original (172-175) need to be tested on real and simulated data. Preliminary computer experiments of the effect of the geometry for model selection, just at the level of correct volume normalization in the MDL principle, are currently being investigated by Eitel Lauria and myself. These experiments show that adding (to the MDL score) a simple approximation to the logarithm of the total volume occupied by a binary bayesian network of a given structure never hurts. In fact the modified score chooses the correct model in (on the average) about 30% more cases than the standard MDL score across a wide range of structures and sample sizes. These results are encouraging but nothing is known, at the moment of writing, about the actual performance of the new curvature term introduced in this paper.

AUTOMATIC MODEL PRODUCTION

In this section we show how the actions for model selection, introduced in the previous section, can be transformed into actions for the automatic production of statistical models. The resulting theory seems to be completely unexplored and extremely rich. What follows is just a very first attempt at describing what seems to me as an ocean of new statistical questions.

Simply put. Take the exp of (179) and elevate the obtained score to the status of *universal ranking* for all the unknowns, i.e., the dimension κ, the information metric g_{ij} and the unnormalized prior π, or equivalently its log likelihood f, where $\pi = e^{-f}$. When $T = |\nabla f|^2$ and $\tau = (2\alpha)^{-1}$ the obtained score is equivalent mod α^{-2} to minimizing,

$$\mathscr{F} = \int_M \left[\tau(R + |\nabla f|^2) - \frac{\kappa}{2}\right](4\pi\tau)^{-\kappa/2} e^{-f} \, dV \tag{180}$$

over smooth functions f satisfying the normalization condition,

$$\int_M (4\pi\tau)^{-\kappa/2} e^{-f} \, dV = 1 \tag{181}$$

Perelman's Action

In a recent paper[5] Grisha Perelman shocked the topology world with what the experts say may be a final proof of the so called geometrization conjecture for three-manifolds. This implies the famous Poincaré conjecture. It is much more famous now that there is a \$1M prize for its proof [6]. What shocks this pedestrian statistician is that (180) is the action that produces the flow of metrics $g_{ij}(\tau)$ that is able to smooth out all manifolds and prove that topologically, spheres are not doughnuts in any dimensions. I believe that the entropy formula for Ricci flows, that solves the presumably esoteric *ball $\neq$ doughnut* problem has an statistical interpretation with the potential for unifying not just heaven and earth but the mind with the moon. In the mean τime,

[5] see `arXiv:math.DG/0211159`
[6] see `www.claymath.org`

Truth, time & τemperature

time is cold Truth
and
Truth is hot time.
Red hot is a true first second
of love.
But all Truth is temporal.
Ergo, F$$$.
that's true
capitalism.
T is t and τ
τ & t is T & T
T is t
and τ is T
god save TT&T
god save τT&t
god save Tτ&t
god save the t.

ACKNOWLEDGMENTS

REFERENCES

1. Brown, M. P., and Austin, K., *The New Physique*, Publisher Name, Publisher City, 2000, pp. 212–213.
2. Brown, M. P., and Austin, K., *Appl. Phys. Letters*, **85**, 2503–2504 (2000).
3. Mittelbach, F., and Schöpf, R., *TUGboat*, **11**, 297–305 (1990), URL `http://www.latex-project.org`.
4. Wang, R., "Title of Chapter," in *Classic Physiques*, edited by R. B. Hamil, Publisher Name, Publisher City, 2000, pp. 212–213.
5. van Herwijnen, E., Future Office Systems Requirements, Tech. rep., CERN DD Internal Note (1988).
6. Liang, F. M., *Word Hy-phen-a-tion by Com-pu-ter*, Ph.D. thesis, Stanford University, Stanford, CA 94305 (1983), also available as Stanford University, Department of Computer Science Report No. STAN-CS-83-977.
7. Smith, C. D., and Jones, E. F., "Load-Cycling in Cubic Press," in *Shock Compression of Condensed Matter-1999*, edited by M. D. F. et al., AIP Conference Proceedings 505, American Institute of Physics, New York, 1999, pp. 651–654.
8. Knuth, D. E., The WEB System of Structured Documentation, Tech. Rep. STAN-CS-83-980, Department of Computer Science, Stanford University, Stanford, CA 94305 (1983).