MDL, Bayesian Inference and the Geometry of the Space of Probability Distributions

Vijay Balasubramanian[*]
Department of Physics and Astronomy
David Rittenhouse Laboratory, University of Pennsylvania
Philadelphia, PA 19104, U.S.A.

Abstract

The Minimum Description Length (MDL) approach to parametric model selection chooses a model that provides the shortest codelength for data, while the Bayesian approach selects the model that yields the highest likelihood for the data. In this article I describe how the Bayesian approach yields essentially the same model selection criterion as MDL provided one chooses a Jeffreys prior for the parameters. Both MDL and Bayesian methods penalize complex models until a sufficient amount of data has justified their selection. I show how these complexity penalties can be understood in terms of the geometry of parametric model families seen as surfaces embedded in the space of distributions. I arrive at this understanding by asking how many different, or distinguishable, distributions are contained in a parametric model family. By answering this question, I find that the Jeffreys prior of Bayesian methods measures the density of distinguishable distributions contained in a parametric model family in a reparametrization independent way. This leads to a picture where the complexity of a model family is related to the fraction of its volume in the space of distributions that lies close to the truth.

1 Introduction

Occam's Razor, the principle of economy of thought invented by the Scholastic philosopher, William of Ockham (see, e.g., [1]), remains a fundamental heuristic guiding the thought of modern scientists. As a rule of thumb it states that simple explanations of a given phenomenon are to be preferred over complex ones. But why are simple explanations better? Simple explanations are certainly easier for us to understand, but is there any fundamental sense in which simple explanations are actually better at describing phenomena? Clearly, the answer to this question hinges on what the meaning of simplicity is in this context. It also has a bearing on what the physicist and mathematician Eugene Wigner called the "inexplicable effectiveness of mathematics in the natural sciences" [2]. Namely, mathematical models derived to fit a small amount of restricted data often correctly describe surprisingly general classes of phenomena.

[*]vijay@endive.hep.upenn.edu

In the modern context, Occam's Razor has found a technical statement in the Minimum Description Length (MDL) principle, which states that the best model of a collection of data is the one that permits the shortest description of it. In the context of statistical inference of parameteric families of models, one collects N data points and uses a statistical model to encode them as compactly as possible. Theorems from information theory then bound the length of the encoding in bits to be at least

$$SC = -\ln(E|\hat{\Theta}) + \frac{d}{2}\ln N + O(1) \tag{1}$$

where E is the data, N is the number of data points, $\hat{\Theta}$ are the maximum likelihood parameters and d is the number of parameters of the model. Rissanen has called this quantity the *stochastic complexity* of a parametric family of models [3, 4]. The first term turns out to be $O(N)$ term as we will discuss later, and penalizes models which assign the data low likelihood and the $O(\ln N)$ term penalizes models with many parameters. A model with lower stochastic complexity must therefore be both accurate and parsimonious. The MDL principle asserts that the best guide to the "truth" from which the data are drawn is given by the model which minimizes the stochastic complexity for describing the N available data points. This principle is *consistent* – if the truth lies in one of the model families under consideration the $O(N)$ term in the stochastic complexity guarantees that it will eventually be selected as giving the best description of the data (see, e.g., the classic papers [3, 4, 5, 6]).

However, at least intuitively, complexity of a parametric statistical model should involve more than just the number of parameters. For example, a good model should be robust in that it should not depend too sensitively on the choice of parameters. The purpose of this paper is to approach model selection through a more intuitive route than coding theory. Given a collection of data drawn from some unknown distribution we can compare the quality of two parametric models by simply asking which one is more likely to have produced the data. While carrying out this procedure in Sec. 2, the essential step is the use of Bayes' formula to find the likelihood of a model family given the data from the likelihood of the data given the model. We then need to know the *a priori* likelihood that the truth is given by a model a particular set of parameters. One might think that an unbiased choice of prior likelihood is to declare all parameter choices to be equally likely. However, we will see that this choice depends on the choice of paramterization and is therefore not suitable [7, 8].

In Sec. 3 I will argue that we can arrive at an unbiased (or reparametrization-invariant) choice of prior likelihood by demanding that all *distributions* rather than parameters are equally likely a priori. We will find such a prior distribution by devising a method to essentially count the different distributions are indexed by the parameters of a model family and by weighting all of these equally. The resulting prior distribution on parameters will be the famous Jeffreys prior of Bayesian inference [7]. We will see how this prior is the reparametrization-invariant measure associated to a natural metric (the Fisher Information matrix) on the space of probability distributions.

In Sec. 4 will will employ the Jeffreys' prior in a Bayesian formulation of parametric model selection. When the number of data points is large we will be able to use the techniques of "low temperature expansions" in statistical physics (see, e.g., [11]) to evaluate the likelihood of a model given the data. Indeed, there will be several attractive analogies between quantities appearing in the inference problem and quantities like energy and temperature in physical systems, leading to useful intuitions. We will see that probability theory advises us to select models that minimize the quantity

$$\chi = -\ln \Pr(E|\hat{\Theta}) + \frac{d}{2} \ln \frac{N}{2\pi} + \ln \int d\theta \sqrt{\det J(\Theta)} + \frac{1}{2} \ln \left(\frac{\det I(\hat{\Theta})}{\det J(\hat{\Theta})} \right) + O(1/N) \quad (2)$$

where E is the data, N is the number of data points, d is number of parameters, $\hat{\theta}$ are the maximum likelihood parameters, and J and I are analogues of the Fisher information matrix that will be explained further in the text. Notice that the $O(N)$ and $O(\ln N)$ terms coincide with stochastic complexity (1). The second and third terms, are completely independent of the data and have been called the "geometric complexity" of the model in [12]. We will see that the third and fourth terms, both of $O(1)$, together essentially measure the fraction of the volume of a model family, as measured in the Fisher Information metric, that lies close to the truth. Thus models that are "unnatural" or lack "robustness" in the sense of mostly describing hypotheses far from the truth are penalized. In this way, the Bayesian approach provides an intuitive understanding of the origin of complexity of a model in terms of the geometry of the space of distributions.

The first of the $O(1)$ terms in (2) has appeared in Rissanen's refinement of the Minimum Description Length principle [13] based on a more accurate form of stochastic complexity. As we will, the second terms is relevant when the true model does not lie within the model family under consideration. An important purpose of this article is to provide some intuitions for the origin of the Minimum Description Length principle in the geometry of the space of distributions. As such I will not strive for mathematical rigor, taking rather the approach of a physicist that the various approximations that I use will be valid under suitably general (but often unspecified!) circumstances. I will be extensively using material that appears in [14].

2 The Bayesian Approach to Parametric Inference

Suppose we are given a collection of outcomes $E = \{e_1 \dots e_N\}$, $e_i \in X$ drawn independently from a density t. Suppose also that we are given two parametric families of distributions A and B and we wish to pick one of them as the model family that we will use. The Bayesian approach to this problem consists of computing the posterior conditional probabilities $\Pr(A|E)$ and $\Pr(B|E)$ and picking the family with the higher probability. Let A be parametrized by a set of real parameters $\Theta = \{\theta_1, \ldots \theta_d\}$. Then Bayes Rule tells us that:

$$\Pr(A|E) = \frac{\Pr(A)}{\Pr(E)} \int d^d\Theta \; w(\Theta) \Pr(E|\Theta) \tag{3}$$

In this expression $\Pr(A)$ is the prior probability of the model family, $w(\Theta)$ is a prior density on the parameter space and $\Pr(E)$ is a prior density on the N outcome sample space. I denote the measure induced by the parametrization of the d dimensional parameter manifold as $d^d\Theta$ in a notation familiar to physicists. (For example, if x and y are real parameters, this integration measure is just $dx\,dy$.) Since we are interested in comparing $\Pr(A|E)$ with $\Pr(B|E)$, the prior $\Pr(E)$ is a common factor that we may omit, and for lack of any better choice we take the prior probabilities of A and B to be equal and omit them. For the present we will assume that the model families of interest to us have compact parameter spaces so that integral over Θ occurs over a bounded domain. In applications the parameter space is often unbounded and understanding how to deal with this situation is a very important practical issue. We will return to this in Sec. 5. As the parameters range over their different values, a given model family sweeps out a a surface, or manifold, in the space of probability distributions. This is illustrated in Fig. 1 which illustrates the space of distributions, with two model families embedded in it, one with one parameter, and the other with two. We will refer to the *parameter manifold* for, say, model family A, by the notation $\mathcal{M}_A$.

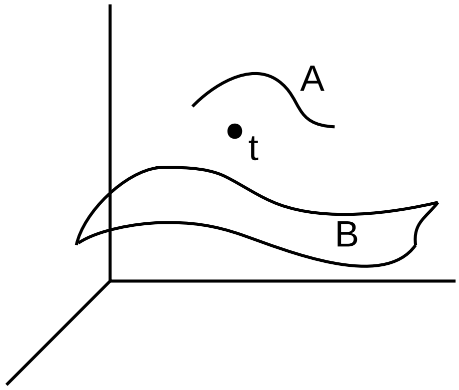

Figure 1: The space of probability distributions with the true data-generating distribution labelled as "t". A and B label two parametric model families seen as surfaces embedded in the space of distributions. A is a 1 parameter (one dimensional) family, while B has two parameters.

2.1 The importance of reparametrization invariance

For the present, after dropping $\Pr(A)$ and $\Pr(E)$ our goal is to evaluate the posterior likelihood of a model:

$$P_{A|E} = \int_{\mathcal{M}_A} d^d\Theta \; w(\Theta) \Pr(E|\Theta) \tag{4}$$

To evaluate this we must determine the prior probability of the parameters Θ of the model family, or, equivalently, determine an appropriate measure $d\mu(\Theta) = d^d\Theta \, w(\Theta)$ for integration over the parameter space. What is the correct choice of $w(\Theta)$ in the absence of adidtional prior information? Since $d\mu$ must be a probability measure, it must be that the integral over the parameter space is equal to one: $\int_{\mathcal{M}_A} d\mu(\Theta) = \int_{\mathcal{M}_A} d^d\Theta \, w(\Theta) = 1$. If we wish to be unbiased in our inference we should now pick a $w(\Theta)$ that does not favor any part of the model space over another. Frequently, it is supposed that the correct way to this is to pick a constant $w(\Theta)$ so that all the parameters are given equal weight *a priori*. The requirement that the integral over the parameter space is one then gives

$$w(\Theta) = \frac{d^d\Theta}{\int_{\mathcal{M}_A} d^d\Theta} \tag{5}$$

The denominator is the volume of the parameter space as measured by the Lebesgue measure on the parameter manifold $\mathcal{M}_A$.

Although this choice of a uniform prior seems natural, it is in fact a biased choice in the sense that uniform priors relative to different arbitrary parametrizations can assign different probability masses to the same subset of parameters. To illustrate this, suppose that a model has two parameters x and y. Then (5) becomes

$$d^d\Theta \, w(\Theta) = \frac{dx \, dy}{\int_{\mathcal{M}_A} dx \, dy} \, . \tag{6}$$

We could have chosen to parametrize the same model in terms of $r = \sqrt{x^2 + y^2}$ and $\phi = \arctan(y/x)$. In that case, given the pair (r, ϕ) the measure (5) which weights all parameter choices equally gives

$$d^d\Theta \, w(\Theta) = \frac{dr \, d\phi}{\int_{\mathcal{M}_A} dr \, d\phi} \, . \tag{7}$$

By contrast, if we change coordinates in the measure (6) from (x, y) to (r, ϕ), and include the Jacobian of the transformation, the measure becomes

$$d^d\Theta \, w(\Theta) = \frac{r \, dr \, d\phi}{\int_{\mathcal{M}_A} dr \, d\phi} \, . \tag{8}$$

Notice that (8) and (7) are not the same thing. In other words, the prescription (5) for giving equal weights to all parameters is itself parameter-dependent and thus an undesirable method of selecting a prior distribution.

Of course, once we have picked a particular prior distribution $w(\Theta)$ Bayesian inference is reparameterization invariant provided we remember to include the Jacobian of coordinate transformations in the integration measure as we are instructed to do in elementary calculus classes. The point here is that the apparently unbiased measure (5) that gives equal weight to all parameters is not *reparametrization invariant* and is therefore unacceptable; if $w(\Theta)$ was uniform in the parameters, the probability of a model family given the observed data would depend on the arbitrary parametrization. We need some other way of determining an unbiased distribution of the parameter space of a model. In the next section we will propose that a good method is to give equal prior weight to all the *distributions* contained in a model family as opposed to all the parameters, which are only an arbitrary scheme for indexing these distributions.

3 Counting probability distributions

We would like to determine a prior probability density of the parameters of a model that that implement the reasonable requirement that all *distributions* rather than all *parameters* are equally likely. The basic obstacle to doing this is that the parameters of model can cover the space of distributions unevenly; some regions of parameter space might index probability distributions more "densely" than others. If this happens, the "denser" regions should be given more weight since they contain more distinguishable probability distributions. So let us ask the question, "How do we count the number of distinct distributions in the neighbourhood of a point on a parameter manifold?" Essentially, this is a question about the embedding of the parameter manifold within the space of distributions. Distinguishable choices of parameters might be indexing indistinguishable distributions (in some suitable sense) and we need to account for this to give equal weight to different distributions rather than different parameters.

To answer the question, let Θ_p and Θ_q index two distributions in a parametric family and let $E = \{e_1 \cdots e_N\}$ be drawn independently from one of Θ_p or Θ_q. In the context of model estimation, a suitable measure of distinguishability can be derived by asking how well we can guess which of Θ_p or Θ_q produced E. (See Fig. 2.) Let α_N be the probability that Θ_q is mistaken for Θ_p and let β_N be the probability that Θ_p is mistaken for Θ_q. Let β_N^ϵ be the smallest possible β_N given that $\alpha_N < \epsilon$. Then Stein's Lemma tells us that $\lim_{N\to\infty}(-1/N)\ln\beta_N^\epsilon = D(\Theta_p\|\Theta_q)$ where

$$D(p\|q) = \int dx\, p(x)\ln(p(x)/q(x)) \tag{9}$$

is the relative entropy between the densities p and q ([15]). As shown in the Appendix of [14], the proof of Stein's Lemma shows that the minimum error β_N^ϵ exceeds a fixed β^* in the region where

$$\kappa/N \geq D(\Theta_p\|\Theta_q) \quad ; \quad \kappa \equiv -\ln\beta^* + \ln(1-\epsilon). \tag{10}$$

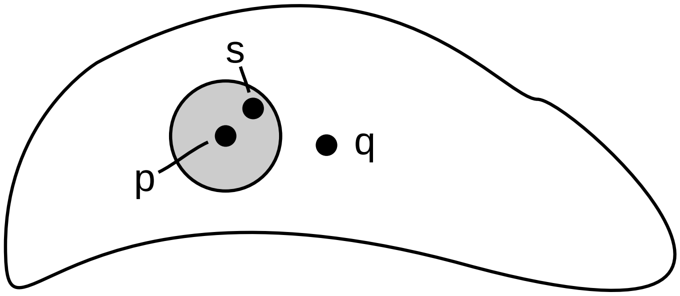

Figure 2: Three distributions p, q and s are labelled in this picture of a parametric model family. The grey region indicates the neighbourhood of p which contains distributions that are sufficiently similar that it will be difficult to guess which one of then produced a given sample of N data points. Thus, p and s will be *indistinguishable* and given only N data points they should not be counted as different distributions for the purposes of statistical inference. By contrast, p and q should be treated as *distinguishable* distributions.

(This assertion is not strictly true, but will do for our purposes. See the Appendix of [14] for more details.) By taking β^* close to 1 we can identify the region around Θ_p where the distributions are not very distinguishable from the one indexed by Θ_p. As N grows large for fixed κ, any Θ_q in this region is necessarily close to Θ_p since $D(\Theta_p\|\Theta_q)$ attains a minimum of zero when $\Theta_p = \Theta_q$. Therefore, setting $\Delta\Theta = \Theta_q - \Theta_p$, Taylor expansion gives

$$D(\Theta_p\|\Theta_q) \approx \frac{1}{2}\sum_{ij} J_{ij}(\Theta_p)\Delta\Theta^i\Delta\Theta^j + O(\Delta\Theta^3) \tag{11}$$

where

$$J_{ij} = \nabla_{\phi_i}\nabla_{\phi_j} D(\Theta_p\|\Theta_p + \Phi)|_{\Phi=0} \tag{12}$$

is the Fisher Information.[1]

Summary: The upshot of all of this is simple. For any given number of data points N, there is a region around Θ_p in which the distributions are not very distinguishable from the one indexed by Θ_p itself, in the sense that we would not be able to reliably guess which of these distributions the N data points really came from. As the number of data points grows, this region of indistinguishability is described by the following ellipsoid in the parameter space:

$$\frac{\kappa}{N} \geq D(\Theta_p||\Theta_q) \approx \frac{1}{2}\sum_{ij} J_{ij}(\Theta_p)\Delta\Theta^i\Delta\Theta^j + O(\Delta\Theta^3) \tag{13}$$

[1] We have assumed that the derivatives with respect to Θ commute with expectations taken in the distribution Θ_p to identify the Fisher Information with the matrix of second derivatives of the relative entropy.

Here κ is given in terms of the probability of error in guessing the data generating distribution as in (10). (See Fig. 2.)

3.1 A uniform prior on distributions

We will now devise a measure that gives equal weight to the distributions indexed by a model family as opposed to the parameters. The basic strategy is to begin by giving equal weight to every ellipsoid of the form (13) containing essentially indistinguishable distributions given N data points. By taking the limit $N \to \infty$ we will arrive at a measure on the parameter manifold that effectively gives equal weight to all those distributions that can be told apart or distinguished in a statistical experiment. (See Fig. 3.)

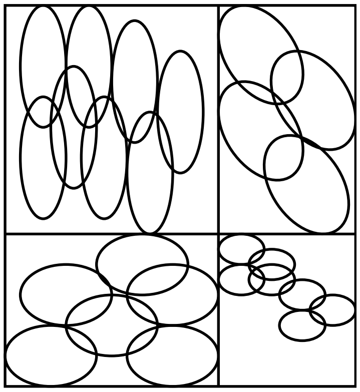

Figure 3: The figure a parameter manifold divided into four regions in each of which the Fisher Information matrix is constant leading to a fixed shape volume of indistinguishability at any given value of N. To derive a measure of the parameter space, we partition the parameter manifold into volumes of indistinguishability as shown. (These will necessarily overlap a little.) Given only N data points we might as well consider each of these volumes as containing only a single distribution since the distributions within them cannot be told apart reliably. In effect, this converts the continuous parameter manifold into a lattice. We can then derive a prior probability density on the parameters by considering each of the discrete number of distributions representing the volumes of indistinguishability as being equally likely. As $N \to \infty$ the volumes of indistinguishability shrink, and in the continuum limit we recover the Jeffreys prior as the measure on the parameter manifold that gives equal weight all equally distinguishable distributions.

To this end, define the *volume of indistinguishability* at levels ϵ, β^*, and N to be the volume of the region around Θ_p where $\kappa/N \geq D(\Theta_p \| \Theta_q)$ so that the probability of error in distinguishing Θ_q from Θ_p is high. We find to leading order:

$$V_{\epsilon,\beta^*,N} = \left(\frac{2\pi\kappa}{N}\right)^{d/2} \frac{1}{\Gamma(d/2+1)} \frac{1}{\sqrt{\det J_{ij}(\Theta_p)}} \tag{14}$$

If β^* is very close to one, the distributions inside $V_{\epsilon,\beta^*,N}$ are not very distinguishable and the Bayesian prior should not treat them as separate distributions. We wish to construct a measure on the parameter manifold that reflects this indistinguishability. We also assume a principle of "translation invariance" by supposing that volumes of indistinguishability at given values of N, β^* and ϵ should have the same measure regardless of where in the space of distributions they are centered. This amounts to an assumption that all distinguishable probability distributions are *a priori* on an equal footing.

An integration measure reflecting these principles of indistinguishability and translation invariance can be defined at each level β^*, ϵ, and N by covering the parameter manifold economically with volumes of indistinguishability and placing a delta function in the center of each element of the cover. This definition reflects indistinguishability by ignoring variations on a scale smaller than the covering volumes and reflects translation invariance by giving each covering volume equal weight in integrals over the parameter manifold. The measure can be normalized by an integral over the entire parameter manifold to give a prior distribution. The continuum limit of this discretized measure is obtained by taking the limits $\beta^* \to 1$, $\epsilon \to 0$ and $N \to \infty$. In this limit the measure counts distributions that are completely indistinguishable ($\beta^* = 1$) even in the presence of an infinite amount of data ($N = \infty$).[2] (See Fig. 3.)

To see the effect of the above procedure, imagine a parameter manifold which can be partitioned into k regions in each of which the Fisher Information is constant. Let J_i, U_i and V_i be the Fisher Information, parametric volume and volume of indistintguishability in the ith region. Then the prior assigned to the ith volume by the above procedure will be

$$P_i = \frac{(U_i/V_i)}{\sum_{j=1}^{k}(U_j/V_j)} = \frac{U_i\sqrt{\det J_i}}{\sum_{j=1}^{k} U_j\sqrt{\det J_j}}\,. \tag{15}$$

Since all the β^*, ϵ and N dependences cancel we are now free to take the continuum limit of P_i. This suggests that the prior density induced by the prescription described in the previous paragraph is:

$$w(\Theta) = \frac{\sqrt{\det J(\Theta)}}{\int d^d\Theta\,\sqrt{\det J(\Theta)}} \tag{16}$$

By paying careful attention to technical difficulties involving sets of measure zero and certain sphere packing problems, it can be rigorously shown that the normalized continuum measure on a parameter manifold that reflects indistinguishability

[2] The α and β errors can be treated more symmetrically using the Chernoff bound instead of Stein's lemma, but we will not do that here.

and translation invariance is $w(\Theta)$ or Jeffreys' prior [16]. In essence, the heuristic argument above and the derivation in [16] show how to "divide out" the volume of indistinguishable distributions on a parameter manifold and hence give equal weight to equally distinguishable volumes of distributions. In this sense, Jeffreys' prior is seen to be a uniform prior on the *distributions* indexed by a parametric family. The density $w(\Theta)$ is the answer to following question: *What is the fraction of the total number of distributions indexed by a model family that is contained with infinitesimal neighbourhood of a parameter value Θ?*

Summary: We are seeking a measure on the parameter space, or a prior probability density, which gives equal weight to equally distinguishable probability distributions. We did this by determining a statistical notion of distinguishability that told us that it was difficult, given only N data points, to tell apart the distributions indexed by the parameters lying within the ellipsoids (13). Our strategy was therefore to discretize the parameter manifold into a grid of such (minimally overlapping) ellipsoids, and then to give equal prior probability to each of the distinguishable distributions located at the grid points. As the number of data points N grow large the grid points approach each other since it becomes easier to tell distributions apart when more data is available. In the limit we recover the continuum measure (16) on the parameter space that has effectively "divided out" redundant descriptions of the same distributions by parameters infinitesimally close to a given one. (See Fig. 3.) Once we have committed to this prior distribution, we can work with any choice of parameters. Although (16) might look different given different choices, we can be sure that (16) always gives a fixed region of parameter space the same probability.

The geometry of the space of probability distributions: In the discussion above we derived the well-known Jeffreys prior (16) as a probablity density giving equal weight to all distinguishable distributions indexed by a parameter manifold. However, the form of the measure $w(\Theta) = \sqrt{\det J(\Theta)}$ suggests another useful interpretation. In Riemannian geometry, the infinitesimal distance Δs between two points separated by a small coordinate difference Δx^i is given by an equation

$$\Delta s^2 = \sum_{ij} g_{ij}(x)\, \Delta x^i\, \Delta x^j \,. \tag{17}$$

This is the generalization to curved manifolds of the Pythagorean theorem and g_{ij} is called the *metric* and is for our purposes simply a matrix that varies over the surface. The corresponding measure for integrating over the curved surface is $d^d x \sqrt{\det g}$. Comparing this with (16) suggests strongly that in the context of statistical inference, a parameter manifold is a curved surface endowed naturally with a metric given by the Fisher information, i.e., $g_{ij} = J_{ij}$ on parameter surface. From this perspective the Jeffreys prior in (16) has the following simple interpretation. First

$$V(A) = \int d^d\Theta \sqrt{\det J(\Theta)} \tag{18}$$

measures the volume of the parameter manifold in the distinguished Fisher Information metric J_{ij}. We then get a uniform prior distribution on the parameters by measuring the volume of of small region of parameters as $d^d\Theta\sqrt{\det J}$ and dividing by the total volume of the parameters so that the distribution integrates to one.

We will not have occasion to exploit this tempting additional structure since all our results will on depend on the measure $\sqrt{\det J}$. However, it is a very interesting question to consider whether the apparatus of classical differential geometry such as measures of curvature, geodesics and other quantities play a role in statistical inference. The reader may wish to consult the works of Amari and others on Information Geometry (see, e.g., [9, 10] in which Fisher information is taken seriously as a metric describing the curved geometry of parameter spaces seen as surfaces embedded in the space of all probability distributions.

4 Occam's Razor, MDL and Bayesian Methods

Putting everything together we get the following expression for the Bayesian posterior probability of a parametric family in the absence of any prior knowledge about the relative likelihood of the distributions indexed by the family.

$$\begin{aligned} P_{A|E} &= \frac{\int d^d\Theta \, \sqrt{\det J} \; \Pr(E|\Theta)}{\int d^d\Theta \, \sqrt{\det J}} && (19) \\ &= \frac{\int d^d\Theta \, \sqrt{\det J} \; \exp\left[-N\left(\frac{-\ln\Pr(E|\Theta)}{N}\right)\right]}{\int d^d\Theta \, \sqrt{\det J}} && (20) \end{aligned}$$

The second form of the expression is useful since the strong law of large numbers says that $(-1/N)\ln\Pr(E|\Theta) = (-1/N)\sum_{i=1}^{N}\ln\Pr(e_i|\Theta)$ converges in the almost sure sense to a finite quantity:

$$E_t\left[\frac{-\ln\Pr(e_i|\Theta)}{N}\right] = \int dx\, t(x)\ln\left(\frac{t(x)}{\Pr(x|\Theta)}\right) - \int dx\, t(x)\ln\left(t(x)\right) = D(t\|\Theta) + h(t) \tag{21}$$

where $h(t)$ is the entropy of the true distribution which generates the data and $D(t|\Theta)$ is the relative entropy (9) between the true distrubution and the one indexed by Θ. This means that as N grows large the integrand in (20) will be dominated by the parameter value that comes closest to the truth. Readers familiar with statistical physics will recognize the structure of these equations. The basic quantity of interest in statistical physics is the partition function

$$Z = \frac{\int d^dx\, \mu(x) e^{-\beta\, E(x)}}{\int d^dx\, \mu(x)} \tag{22}$$

where x labels the space of configurations of a physical system, $\mu(x)$ is a measure on the configuration space, $\beta \equiv 1/T$ is the inverse temperature of the system and $E(x)$ is the energy of the configuration x [11]. The analogy with the Bayesian posterior probability (20) is now clear – for example, inference with a large number N of data points is in analogy to statistical physics at a low temperature T. There are classic techniques in statistical physics to compute Z in various limits that might be useful in Bayesian statistical inference. In this paper we will be interested in studying inference where N is large. We can then borrow the well-known method of low-temperature expansions in statistical physics [11] and apply it to the problem of evaluating (20).

4.1 Asymptotic expansion and MDL

We will now approximately evaluate (20) when the number of data points is large. The method we use applies when: (a) the maximum likelihood parameter $\hat{\Theta}$ which globally maximizes $\Pr(E|\Theta)$ lies in the interior of the parameter space, (b) locally maxima of $\Pr(E|\Theta)$ are bounded away from the global maximum, and (c) both the Fisher information $J_{ij}(\Theta)$ and $\Pr(E|\Theta)$ are sufficiently smooth functions of Θ in a neighbourhood of $\hat{\Theta}$. In this case, for sufficiently large N, the integral in (20) is dominated by a neighbourhood of the maximum likelihood parameter $\hat{\Theta}$. We can then approximate the integrand in the neighbourhood of $\hat{\Theta}$ as follows.

First collect the measure $\sqrt{\det J}$ into the exponent as

$$P_{A|E} = \frac{\int d^d\Theta \, \exp\left[-N\left(\frac{-\ln \Pr(E|\Theta)}{N}\right) + (1/2)\,\mathrm{Tr}\ln J(\Theta)\right]}{\int d^d\Theta \, \sqrt{\det J}} \tag{23}$$

Next we Taylor expand the exponent around the maximum likelihood parameter which satisfies $\nabla_{\theta_\mu} \ln \Pr(E|\Theta) = 0$. So the Taylor expansion of the first term in the exponent begins with ∇_{θ_μ}. It is convenient to define a kind of empirical Fisher information as $I_{\mu\nu} = (-1/N)\nabla_{\theta_\mu}\nabla_{\theta_\nu} \ln \Pr(E|\Theta)|_{\hat{\Theta}}$ so that $I_{\mu\nu}$ approaches a finite limit as $N \to \infty$.

We can then evaluate (23) as follows. First define a shifted integration variable $\Phi = (\Theta - \hat{\Theta})$. Then, we can write

$$\begin{aligned} P_{A|E} &= \frac{e^{-\left[\ln \Pr(E|\hat{\Theta}) - \frac{1}{2} Tr \ln J(\hat{\Theta})\right]} \int d^d\Phi e^{-((N/2)\sum_{\mu\nu} I_{\mu\nu}\phi^\mu\phi^\nu + G(\Phi))}}{\int d^d\Theta \sqrt{\det J_{ij}}} && (24) \\ &= \frac{e^{-\left[\ln \Pr(E|\hat{\Theta})\right]} \sqrt{\det J(\hat{\Theta})} \int d^d\Phi e^{-((N/2)\sum_{\mu\nu} I_{\mu\nu}\phi^\mu\phi^\nu + G(\Phi))}}{\int d^d\Theta \sqrt{\det J_{ij}}} && (25) \end{aligned}$$

where $G(\Phi)$ collects the cubic and higher order terms and all the in the Taylor expansion of $\ln \Pr(E|\Theta)$ and all terms in the Taylor expansion of $\mathrm{Tr}\ln J$ around $\hat{\Theta}$. As the number of data points N gets large the integrand is very sharply peaked around

$\hat{\Theta}$ and the terms collected in $G(\Phi)$ will only make subleading contributions to the integral. Indeed, we can approxime the integral as a multivariate Gaussian with a covariance matrix $N\, I_{\mu\nu}(\hat{\theta})$. (Some technical conditions are required as discussed in; see, e.g, [5].)

When N is large, the Gaussian is very narrow and therefore the integral can be performed [5, 14] to give

$$-\ln P_{A|E} \equiv \chi_E(A) = -\ln \Pr(E|\hat{\theta}) + \ln\left(\frac{V(A)}{V_c(A)}\right) + O(1/N) \tag{26}$$

We have defined [12]

$$V_c(A) = \left(\frac{2\pi}{N}\right)^{d/2} \sqrt{\frac{\det J(\hat{\theta})}{\det I(\hat{\theta})}}\,. \tag{27}$$

$V_c(A)$ is essentially the volume of a small ellipsoid around $\hat{\theta}$ within which the probability of the data $\Pr(E|\Theta)$ is appreciable. Specifically, $V_c(A)$ only differs by a numerical factor from the volume of a region where $\Pr(E|\Theta) \geq \lambda \Pr(E|\hat{\Theta})$ for any λ close to 1. As such, it measures the volume of distinguishable distributions in A that come close to the truth, as measured by predicting the data E with good probability.

The ratio $V_c(A)/V(A)$ penalizes models which occupy a small volume close to the truth *relative to* the total volume of the model. The second term expands to

$$C = \ln\left(\frac{V(A)}{V_c(A)}\right) = \frac{d}{2}\ln\left(\frac{N}{2\pi}\right) + \ln\int d\theta\,\sqrt{\det J(\theta)} + \frac{1}{2}\ln\left(\frac{\det I(\hat{\theta})}{\det J(\hat{\theta})}\right) \tag{28}$$

In Bayesian model selection, C functions as a penalty for complexity.

Assembling everything, when selecting between two model families, A and B, probability theory instructs to compute

$$\chi_E = -\ln \Pr(E|\hat{\theta}) + \frac{d}{2}\ln\frac{N}{2\pi} + \ln\int d\theta\sqrt{\det J(\theta)} + \frac{1}{2}\ln\left(\frac{\det I(\hat{\theta})}{\det J(\hat{\theta})}\right) + O(1/N) \tag{29}$$

for each family given the data E. The observed data are more likely to have come from the model family with a smaller χ. We will interpret this as a manifestation of Occam's Razor and the Minimum Description Length principle and will explain how the various terms in (29) arise from the geometry of the parameter space of the model family.

4.1.1 Interpretation: Occam's Razor

The first term in (29) is maximum log likelihood of the data given a model and therefore measures how accurately the model is able to describe the data. This term is $O(N)$ since, as we discussed $(1/N)\ln(\Pr(E|\hat{\Theta})$ approaches a finite limit. This happens because probabilities multiply and as this causes the probability of any given sample to decrease exponentially with the sample size. In any case, for sufficiently large N this $O(N)$ term always dominates and therefore with enough data the most accurate model family is chosen by Bayesian methods.

As we described the remaining three terms arise in our analysis essentially as a measurement of the fraction of the volume of a model's parameter space that lies close to truth. The first and second terms in C are independent of the true distribution as well as the data, and therefore represent an intrinsic property of the model family. The term proportional to $d/2$ arises because as the number of data points increases the radius of the region near the maximum likelihood parameter that gives a good description of the data shrinks in proportion to $1/N$ so that the volume of this region shrinks as $(1/N)^{d/2}$ as in (27). As discussed earlier the integral $\int d^d\Theta \sqrt{\det J(\Theta)}$ measures in a sense the volume of distinguishable distributions that the model can describe. Thus the third term in (29) penalizes models that are very unconstrained. Finally, the last term in (29) penalizes models that are not robust in the sense that they depend very sensitively of the choice of parameters. We can see then by observing that the covariance matrix I determines how rapidly the integrand of $P_{A|E}$ falls off around the maximum likelihood parameter. So if I is large we have a situation like in Fig. 4a where the model gives a good description of the data only for very restricted parameters. Conversely if I is small, there is large basin of parameters that describes the data well. The ration $\det I / \det J$ appears because how narrow or wide the the "good" region of a model family is should really be determined with respect to the natural measure on the parameter manifold, which we have argued to be the Fisher information matrix.

The Minimum Description Length model selection criterion [13] chooses the statistical model that minimizes the sum of the first three terms in (29). Note that if the true distribution lies within the considered model family, $J(\hat{\theta})$ approaches $I(\hat{\theta})$ as N grows large, and consequently, $\ln(V(f)/V_c(f))$ becomes equal to the complexity penalty in the MDL selection criterion. This shows that as the sample size grows, the log of the Bayesian posterior probability of a model family $(-\ln \Pr(f|y))$, coincides with MDL when the truth lies in the model family. Therefore, selecting the most probable model is essentially equivalent to choosing the model that gives the Minimum Description Length of the data, and the Bayesian complexity C coincides with Rissanen's modified stochastic complexity [13]. It would be very nice to give an adequate interpretation of the final term in (29) in the context of the coding theory that gives rise directly to the MDL criterion.

To summarize again we, have arrived at an intuitive geometric interpretation of the meaning of complexity in the MDL and Bayesian approaches to model selection: *"complexity" measures the ratio of the volume occupied by distinguishable distributions in a model that come close to the truth relative to the volume of the model as a whole.* The apparent complexity of a models functional form in a particular parametrization and even the number of parameters in a model are simply components of this gen-

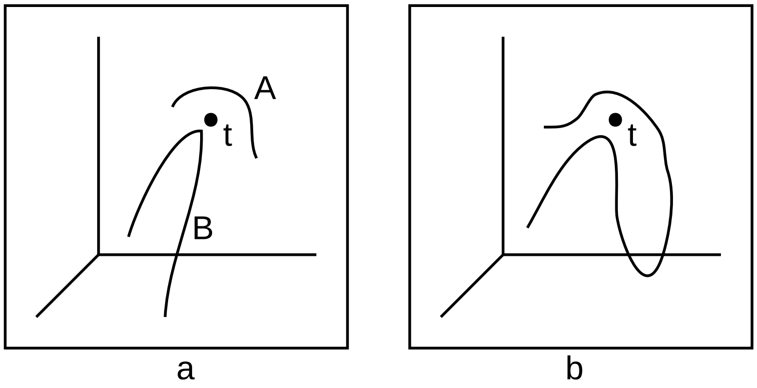

Figure 4: In box (a) model A comes very close to true distribution at one point, but is mostly far away. Model B is close to truth for many choices of its parameters. When the amount of data is small MDL will tend to prefer B because this model is more robust and enough data has not accumulated to identify a specific distributionin A as come close to the truth. As the number of data points increases, however, A will eventually be preferred. Box (b) illustrates a situation where a single model family will have two local maxima in the log likelihood it assigns the data – it come close to the truth in two regions of the parameter space. When the number of data points is small one region (the more robust one) will dominate the Bayesian posterior and as the number of data points increases the other region (the more accurate one) will dominate.

eral understanding of how probability theory incorporates complexity in statistical inference.

5 Some Challenges

This article is appearing in the proceedings of a workshop which described both theoretical developments and practical applications of the MDL techniques (see, e.g., [12]). The principal obstacle to the general application of the results presented here is that we were obliged to assume a bounded parameter space in order to make sense of the Jeffreys prior and consequently of the second term in in the complexity penalty (28) which involved an integral over the parameter space. Actually the problem is not really that the parameter space can be unbounded, but that the integral

$$\int d^d\Theta \sqrt{\det J} \tag{30}$$

can diverge. This can even happen with a bounded parameter space if the Fisher Information J_{ij} becomes infinite sufficiently quickly in some region. In either case the situation is that there are ”too many” candidate hypotheses included in the model family. One simple way to deal with this situation is to bound the domain of parameters in such a way that (30) is finite. In this case we should consider the added variables describing how the parameter space is bounded as parameters of the model themselves and one might imagine doing a "meta-Bayesian analysis" to determine them. Another promising approach is to declare that we are only practically interested in those distributions which assign a probability greater than some small ϵ to the observed data. This will naturally give a bounded domain of parameters describing the data with a reasonable probability. Then we can repeat the entire analysis of this paper for such bounded domains. I hope to report on this approach in a future publication.

Another interesting issue that has been avoided both here and elsewhere in the literature is what happens when there are multiple local likelihood maxima for a given model family. This would arise in a situation such as the one depicted in Fig. 4b where the model family approaches the two distribution in two locations. In such a circumstances $-\ln P_{A|E}$ will be a sum of multiple contributions like the one in (26) each arising from a saddlepoint of the exponent in (23). This sort of situation occurs often in statistical physics and will lead here to an analogue the fascinating phenomenon of *phase transitions* – as the number of data points increases, one saddlepoint or another will suddenly come dominate the log likelihood of the model family leading to potentially very different descriptions of the data.

References

[1] A.A. Maurer, *Medieval Philosophy*, Pontifical Institute of Medieval Studies, 1982.

[2] E. Wigner, "The Unreasonable Effectiveness of Mathematics in the Natural Science", Commun. Pure and App. Math., 13(1), 1960.

[3] J. Rissanen. Universal coding, information, prediction and estimation. *IEEE Transactions on Information Theory*, 30(4):629–636, July 1984.

[4] J. Rissanen. Stochastic complexity and modelling. *The Annals of Statistics*, 14(3):1080–1100, 1986.

[5] B.S. Clarke and A.R. Barron. Information-theoretic asymptotics of bayes methods. *IEEE Transactions on Information Theory*, 36(3):453–471, May 1990.

[6] A.R. Barron and T.M. Cover. Minimum complexity density estimation. *IEEE Transactions on Information Theory*, 37(4):1034–1054, July 1991.

[7] H. Jeffreys. *Theory of Probability.* Oxford University Press, 3rd edition, 1961.

[8] P.M. Lee. *Bayesian Statistics: An Introduction.* Oxford University Press, 1989.

[9] S.I. Amari. *Differential Geometrical Methods in Statistics*. Springer-Verlag, 1985.

[10] S.I. Amari, O.E. Barndorff-Nielsen, R.E. Kass, S.L. Lauritzen, and C.R. Rao. *Differential Geometry in Statistical Inference*, volume 10. Institute of Mathematical Statistics Lecture Note-Monograph Series, 1987.

[11] S.K. Ma. *Statistical Mechanics*. World Scientific, 1985.

[12] I.J. Myung, V. Balasubramanian and M.A. Pitt, "Counting Probability Distributions: Differential Geometry and Model Selection", Proceedings of the National Academy of Science, 97(21) 11170–11175, 2000.

[13] J. Rissanen, IEEE Trans. Inform.Theory, 42:40-47, 1996.

[14] V. Balasubramanian, "Statistical Inference, Occam's Razor and Statistical Mechanics on The Space of Probability Distributions", Neural Computation, 9(2),1997.

[15] T.M. Cover and J.A. Thomas, *Elements of Information Theory*, Wiley, New York, 1991.

[16] V. Balasubramanian, "A geometric formulation of occam's razor for inference of parametric distributions",available as Princeton University Physics Preprint PUPT-1588, January 1996m and preprint number adap-org/9601001 from http://arXiv.org/

[17] I. J. Myung, M. A. Pitt, S. Zhang and V. Balasubramanian. *The use of MDL to select among computational models of cognition*. To appear in Advances in Neural Information Processing Systems 13, MIT Press.