,

On The Relationship between Bayesian and Maximum Entropy Inference

Peter Cheeseman * and John Stutz *

MS 269-3 NASA-ARC, Moffet Field, CA

Abstract. We investigate Bayesian and Maximum Entropy methods for doing inference under uncertainty. This investigation is primarily through concrete examples that have been previously investigated in the literature. We find that it is possible to do Bayesian and MaxEnt inference using the same information, despite claims to the contrary, and that they lead to different results. We find that these differences are due to the Bayesian inference not assuming anything beyond the given prior probabilities and the data, whereas MaxEnt implicitly makes strong independence assumptions, and assumes that the given constraints are the only ones operating. We also show that maximum likelihood and maximum a posteriori estimators give different and misleading estimates in our examples compared to posterior mean estimates. We generalize the classic method of maximum entropy inference to allow for uncertainty in the constraint values. This generalized MaxEnt (GME) makes MaxEnt inference applicable to a much wider range of problems, and makes direct comparison between Bayesian and MaxEnt inference possible. Also, we show that MaxEnt is a generalized principle of independence, and this property is what makes it the preferred inference method in many cases.

INTRODUCTION

We refer to the problem of inferring the state of the world given our prior knowledge and observations (data) as *inductive inference*. We refer to the problem of assigning probabilities to future observations, given the current state of world knowledge, as the problem of *predictive inference*. The relationship between these types of inference is shown in Fig. 1. Here, the estimation step is just a probabilistic weighting of some function f of θ. For example, for a mean estimate, $f(\theta) = \theta$. Here we assume that our prior state of world knowledge is represented by a joint probability density function (pdf) over the state space. If we know the world state with certainty, then this joint density is represented by a set of delta functions, and no further observations can change it. More typically, we are less than certain about the world state, and so a pdf is the

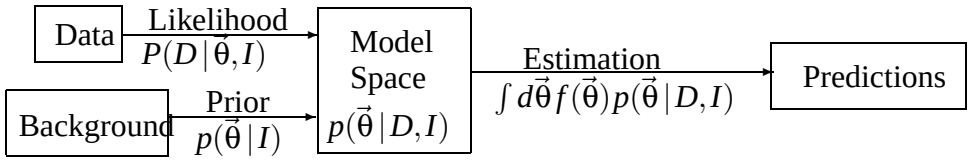

FIGURE 1. The Bayesian inference and prediction processes.

appropriate representation, as it spreads our knowledge over all possible states, with a density concentration where our knowledge is strongest.

In addition to summarizing our world knowledge in a pdf, we often wish to make predictions about future observations, given this pdf. This requires that we assign *point probabilities* over the set of possible outcomes of the observations. There are alternative ways of extracting these point estimates from the state pdf, and one goal of this paper is to show that even in common situations these alternative estimates can be radically different.

To make these observations concrete, we consider a simple dice problem. We show how to do Bayesian inductive inference for variants of this problem, then we show that alternative estimating procedures can produce very different point probabilities. Specifically, we consider Maximum A Posteriori (MAP); Maximum Likelihood (ML); Posterior Mean (PM), and Maximum Entropy (MaxEnt). In summary, we find that MAP and ML estimates are very similar, provided the prior density is relatively "flat", but that they both differ radically from the mean posterior estimate (PM). These estimates represent the answer to different questions, so it is not surprising that they give different results.

We next compare these estimates with those given by maximum entropy (MaxEnt). We find that the "classic" maxent estimation procedure (CME) is only asymptotically correct, because it does not take into account the uncertainty in the value(s) of the contributing constraint(s). We show how to generalize the classic maxent procedure (GME) to take this uncertainty into account, and then show the MaxEnt estimates (classic or generalized) can be radically different from all the other estimates (ML, MAP, PM). This difference is due to the different assumptions embedded in MaxEnt. In particular, MaxEnt makes the strong assumption that the constraints used are *complete*. That is, that the constraints used are the only ones that are operating in the particular domain. If this assumption is incorrect, the MaxEnt estimates can differ significantly from observations. We show that independence assumptions are *essential* for doing practical probabilistic inference, because without a strong prior belief in independence, there is never enough data to estimate all the possible dependencies in real situations. . MaxEnt is a generalized principle of independence, and so is a useful tool for uncovering structure in data. We believe the analysis in this paper solves the mystery of the relationship between Bayesian and MaxEnt inference.

EXAMPLES

We explain our conclusions via a set of examples, where we state as clearly as possible just what information is given and what is not. These examples are variants of those introduced by Jaynes [1, 2, 3]. The goal in each case is to find what the given information tells us about the unknown probabilities.

These examples are:

1. *Face Counts Given* In this case, the number of times each face comes up is given as an integer, n_i for the ith face of the dice. This example sets the stage for the following examples.

2. *Number of Throws and Total Spot Count Given* Only two pieces of information are given: N, the total number of throws, and M the total number of spots observed. Equivalently, we could give N and the mean number of spots, M/N. The goal here is to find what this limited information tells us about the unknown face probabilities.
3. *Contingency Table Data* We consider a contingency table over four discrete attributes having cardinalities (2, 2, 7, 3). We examine the two extreme cases: where all cell counts are given, and where only some of the low order marginals are given.

For each problem we assume that no other information is available.

THE BAYESIAN SOLUTION OF EXAMPLES

We now develop a Bayesian solution to the above examples. In general, Bayes rule is:

$$\begin{aligned} P(\theta \,|\, data, I) &\propto P(data \,|\, \theta, I) \times P(\theta \,|\, I) \\ posterior &\propto likelihood \times prior \end{aligned}$$

where θ is a set of model parameters[1], I the background information, and the proportionality constant is found from the requirement that the sum (or integral) over all allowed combinations of model parameters equals one. We now illustrate this approach on the first example.

Bayesian Solution of Example 1

We have a six-sided die believed to be biased in some way. We assume a fixed probabilistic outcome, representing the unknown model probabilities of the faces by the 6-vector $\vec{\theta}$ with indices i, $1 \leq i \leq 6$. Since one of these faces must appear in any throw, we have the normalization requirement:

$$\forall i,\ 0 \leq \theta_i \leq 1, \qquad \sum_{i=1}^{6} \theta_i = 1, \tag{1}$$

so there are only 5 degrees of freedom in θ space. Our data consists of a 6-vector, $\vec{n}$, where n_i is the number of occasions that face i was observed, with a total of N throws. Our goal is to find the posterior probability of $\vec{\theta}$ given the data–i.e. $P(\vec{\theta} \,|\, \vec{n}, I)$ using Bayes theorem. To do this we first find the likelihood $P(\vec{n} \,|\, \vec{\theta}, N, I)$–i.e. the probability of observing a particular $\vec{n}$ given that we know the intrinsic face probabilities $\vec{\theta}$ and the number of throws N. The I denotes any background information. Under our assumptions,

[1] In this paper Greek symbols denote model parameters, and Roman symbols denote observable quantities.

this is given by the well known discrete multinomial probability:

$$P(\{n_1 \ldots n_6\} \,|\, \vec{\theta}, N, I) \;=\; \frac{N!}{\prod_{i=1}^{6} n_i!} \prod_{i=1}^{6} \theta_i^{n_i}. \tag{2}$$

$$\forall i,\ n_i \geq 0, \qquad N = \sum_{i=1}^{6} n_i.$$

This likelihood is derived under the assumption of conditional independence of the throws–i.e. we assume that:

$$\begin{aligned} &P(\mathit{face} = i \textit{ on next throw} \,|\, \vec{\theta},\ \mathit{face} = j \textit{ on any previous throw},\ I) \\ &\quad = \quad P(\mathit{face} = i \textit{ on next throw} \,|\, \vec{\theta},\ I) \end{aligned}$$

i.e. that the probability of the next face equaling i depends only on the intrinsic face probabilities $\vec{\theta}$, and not on previous throws. This conditional independence assumption is sometimes expressed as the dice sequences being "exchangeable"–i.e. that the order of the throw outcomes does not matter, only the total number of times the different faces come up. This condition is also sometimes expressed by stating that the n_is form "sufficient statistics", since they contain all the information required to determine the likelihood (2).

Next, we express our prior knowledge of $\vec{\theta}$ as a joint pdf[2], here denoted by a p rather than a P. In this example we know very little about $\vec{\theta}$ and so our prior pdf must be very flat, allowing for all possibilities. A mathematically convenient choice is the Dirichlet distribution:

$$p(\vec{\theta} \,|\, \{w_1 \ldots w_k\}, I) \;=\; \frac{\Gamma(W))}{\prod_{i=1}^{k} \Gamma(w_i)} \prod_{i=1}^{k} (\theta_i^{w_i - 1})\, \delta(\sum_i (\theta_i) - 1), \tag{3}$$

$$W \equiv \sum_{i=1}^{k} w_i, \qquad \forall i,\ w_i > 0, \qquad \delta(x) \equiv \begin{cases} 1; & x = 0 \\ 0; & x \neq 0 \end{cases},$$

where k is the number of possible outcomes, 6 for a die, and $\Gamma(.)$ is the well known gamma function. This is a conjugate prior, so called because the resulting posterior has the same form as the prior. It is a generalization of the Laplacian uniform prior, which it reverts to when all w_i equal 1. The magnitude of W is effectively the total weight of prior evidence, measured in instances $\vec{w}$. The $\delta(\cdot)$ term enforces constraint (1).

We can now combine the likelihood (2) and prior (3) to get the desired joint posterior probability density:

$$p(\vec{n}, \vec{\theta} \,|\, \vec{w}, I) = \frac{N!}{\prod_i n_i!} \frac{\Gamma(W)}{\prod_i \Gamma(w_i)} \prod_i (\theta_i^{n_i + w_i - 1})\, \delta(\sum_i (\theta_i) - 1). \tag{4}$$

[2] pdf = probability density function.

Integrating over $\vec{\theta}$, subject to (1), gives $P(\vec{n}\,|\,N,\vec{w},I)$, the data probability and normalization constant for the specified model:

$$P(\vec{n}\,|\,\vec{w},I) = \frac{N!}{\prod_i n_i!}\,\frac{\Gamma(W)}{\prod_i \Gamma(w_i)}\,\frac{\prod_i \Gamma(n_i+w_i)}{\Gamma(N+W)}. \tag{5}$$

Division of (4) by (5) then gives the posterior probability density distribution for $\vec{\theta}$:

$$p(\vec{\theta}\,|\,\vec{n},\vec{w},I) = \frac{\Gamma(N+W)}{\prod_i \Gamma(n_i+w_i)}\prod_i(\theta_i^{n_i+w_i-1})\,\delta(\sum_i(\theta_i)-1). \tag{6}$$

Note that equations (3) and (6) are of the same form. Equation (6) is the general solution to the Bayesian inductive inference problem represented by example 1, using conjugate priors.

Both the joint (4) and posterior (6) are pfds because the given information (prior pdf and $\vec{n}$) is not sufficient to pin down the face probabilities to point values. However, as N gets large, the posterior density becomes increasingly concentrated around the true value and in the limit becomes a delta function at the true $\vec{\theta}$ value. These joint and posterior pdfs on $\vec{\theta}$ neither under nor over state our knowledge given the information that has been used, and contain all the information necessary to answer many other questions. For example, we might be interested in the marginal posterior pdf on some particular θ_i. This is found by marginalizing over all the other θ_j.

$$\begin{aligned} p(\theta_i\,|\,\vec{n},\vec{w},I) &= \iint_0^1 \prod_{\substack{j\\ j\neq i}}(\mathrm{d}\theta_j)\,P(\vec{\theta}\,|\,N,\vec{n},\vec{w},I) \\ &= \frac{\Gamma(N+W)\;\;\theta_i^{n_i+w_i-1}\,(1-\theta_i)^{N+W-n_i-w_i-1}}{\Gamma(n_i+w_i)\,\Gamma(N+W-n_i-w_i)} \end{aligned} \tag{7}$$

Note that this marginal pdf for θ_i could have been derived directly from the binomial distribution given only the information $\vec{n}$ and $\vec{w}$.

Moments are easily estimated from (7), allowing estimation of the central moments or any other function of the moments. These moments are:

$$\begin{aligned} \langle\theta_i^j\,|\,\vec{n},\vec{w},I\rangle &= \int_0^1 \mathrm{d}\theta_i\,\theta_i^j\,p(\theta_i\,|\,N,\vec{n},\vec{w},I) \\ &= \frac{\Gamma(N+W)}{\Gamma(n_i+w_i)}\,\frac{\Gamma(n_i+w_i+j)}{\Gamma(N+W+j)}; \\ &= \prod_{l=0}^{j-1}\frac{n_i+w_i+l}{N+W+l}; \qquad j>0, \quad j\in \mathrm{Int} \end{aligned} \tag{8}$$

In particular, with $j=1$, the first moments are:

$$\langle\theta_i\rangle = \frac{n_i+w_i}{N+W}. \tag{9}$$

If we denote the number of *future* occurrences of the ith face by r_i, and the total number of future throws by $R = \sum_i r_i$, then the predictive probability[3] of observing a particular $\vec{r}$ given $R, \vec{n}, \vec{w}, I$ is given by:

$$\begin{aligned} P(\vec{r}\,|\,R,\vec{n},\vec{w},I) &= \int_0^1 \mathrm{d}\vec{\theta}\, P(\vec{r}\,|\,\vec{\theta},R,I)\; p(\vec{\theta}\,|\,N,\vec{n},\vec{w},I) \\ &= \frac{\Gamma(R+1)}{\prod_i \Gamma(r_i+1)}\,\frac{\Gamma(N+W)}{\prod_i \Gamma(n_i+w_i)}\,\frac{\prod_i \Gamma(r_i+n_i+w_i)}{\Gamma(R+N+W)}; \end{aligned} \tag{10}$$

In particular, the predictive probability that in the next throw (i.e., $R = 1$) the ith face will come up (i.e., $r_i = 1$ and $r_j = 0$ for $j \neq i$) is obtained from equation (10), giving:

$$P(r_i = 1\,|\,\vec{n},\vec{w},R = 1,I) = \frac{n_i + w_i}{N+W}. \tag{11}$$

Note that (9) is the same as (11), showing that the probabilities of the next outcome are identical to the mean of the corresponding posterior densities, without any reference to utility/loss functions. Calculation of a (probabilistic) state of *knowledge* given prior knowledge and data, such as shown in equations (7) and (10), should not make any reference to how we *value* that state of knowledge, as expressed in utility/loss functions. For this reason, we have never understood the popular approach to Bayesian estimation that minimizes some loss function given the prior knowledge and data. In this case, use of a loss function is not necessary, and conceptually suspect.

The above results, going back to Laplace and Dirichlet, are used in the following sections.

The Bayesian Analysis of Example 2

As in the previous example, we seek the posterior pdfs for $\{\theta_i\}$. However we are given only N, the total number of throws, and either M, the total number of spots observed, or m the observed mean number of spots per throw, where $m = M/N$. Clearly, in this example less information has been given than in the previous case: if there are any, there are usually *many* face count vectors $\vec{n}$ such that:

$$N = \sum_{i=1}^{6} n_i; \qquad\qquad M = \sum_{i=1}^{6} i\, n_i. \tag{12}$$

We denote by $\vec{n}_{NM}$, the set of *all* $\vec{n}$ that satisfy conditions (12).

The Bayesian analysis of this example proceeds as in the previous section. The likelihood of observing any one of the $\vec{n}_{NM}$ is given by the multinomial equation (2). The likelihood of observing a total of M counts in N trials, given $\vec{\theta}$, is then the sum of

[3] See Fig. 1 for a schematic description of predictive probabilities

likelihoods for all the $\vec{n}$ of $\vec{n}_{NM}$:

$$\begin{aligned} P(M\,|\,\vec{\theta},N,I) &= \sum_{\vec{n}_{NM}} P(\vec{n}\,|\,\vec{\theta},N,I) \\ &= \sum_{\vec{n}_{NM}} \left(\frac{N!}{\prod_{i=1}^{6} n_i!} \prod_{i=1}^{6} \theta_i^{n_i} \right). \end{aligned} \tag{13}$$

For small N, this sum can be done exactly [4]. Continuing as in example 1, assuming the Dirichlet density (3) for the $\vec{\theta}$ prior, gives the following joint density and marginals:

$$\begin{aligned} p(M,\vec{\theta}\,|\,N,\vec{w},I) &= \sum_{\vec{n}_{NM}} p(\vec{n}\,|\,\vec{\theta},N,I)\, P(\vec{\theta}\,|\,\vec{w},I), \\ &= \sum_{\vec{n}_{NM}} \left(\frac{N!}{\prod_{i=1}^{6} n_i!} \frac{\Gamma(W)}{\prod_{i=1}^{6}\Gamma(w_i)} \prod_{i=1}^{6} \theta_i^{n_i+w_i-1} \right) \end{aligned} \tag{14}$$

$$\begin{aligned} P(M\,|\,N,\vec{w},I) &= \iint_0^1 \prod_i (\mathrm{d}\theta_i)\; p(M,\vec{\theta}\,|\,N,\vec{w},I) \\ &= \sum_{\vec{n}_{NM}} P(\vec{n}\,|\,N,\vec{w},I), \\ &= \sum_{\vec{n}_{NM}} \left(\frac{N!}{\prod_{i=1}^{6} n_i!} \frac{\Gamma(W)}{\prod_{i=1}^{6}\Gamma(w_i)} \frac{\prod_{i=1}^{6}\Gamma(n_i+w_i)}{\Gamma(N+W)} \right) \end{aligned} \tag{15}$$

$$\begin{aligned} p(M,\theta_i\,|\,N,\vec{w},I) &= \iint_0^1 \prod_{\substack{j \\ j\neq i}} (\mathrm{d}\theta_j)\; p(M,\vec{\theta}\,|\,N,\vec{w},I) \\ &= \sum_{\vec{n}_{NM}} \left(P(\vec{n}\,|\,N,\vec{w},I)\, p(\theta_i\,|\,\vec{n},\vec{w},I) \right), \\ &= \sum_{\vec{n}_{NM}} \left(\frac{N!}{\prod_i n_i!} \frac{\Gamma(W)}{\prod_i \Gamma(w_i)} \frac{\prod_i \Gamma(n_i+w_i)\,\theta_i^{n_i+w_i-1}\,(1-\theta_i)^{N+W-n_i-w_i-1}}{\Gamma(N+W-n_i-w_i)\,\Gamma(n_i+w_i)} \right) \end{aligned} \tag{16}$$

where the forms on the right hand side are defined in equations (2, 3, 5, 6).

The posterior marginal pdf for θ_i, $p(\theta_i\,|\,N,M,\vec{w},I)$, is then (16) over (15). The result is an evidence weighted and normalized sum over the distributions induced by the individual count vectors of $\vec{n}_{NM}$. Expectations have exactly the same form:

$$\langle f(\vec{\theta})\,|\,N,M,\vec{w},I\rangle = \frac{\sum_{\vec{n}_{NM}} \left(P(\vec{n}\,|\,N,\vec{w},I)\,\langle f(\vec{\theta})\,|\,\vec{n},\vec{w},I\rangle \right)}{P(M\,|\,N,\vec{w},I)}. \tag{17}$$

While these forms are conceptually simple, the sums generally prevent reduction to computationally simple functions like (8). However the Laplacian version of the Dirichlet

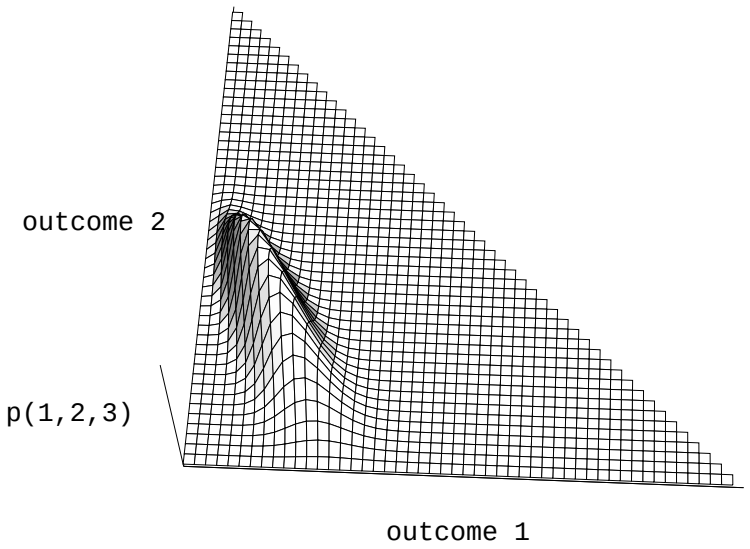

FIGURE 2. The posterior density for the 3-faces die example with a mean spot count of 2.5, N = 60, and prior weights of (1,1,1). Because of the normalization constraint, the third variable (not shown) is given by $\theta_3 = 1 - \theta_1 - \theta_2$.

prior, $(w_i = 1)$, does allow reduction, and yields marginals and estimates that are simple averages over those found for each $\vec{n}$ in $\vec{n}_{N,M}$, see [4] for details.

To illustrate the kind of posterior density generated by this type of problem, we consider the simpler case of a 3 sided die with the faces having 1,2 and 3 spots respectively, and with N = 60 and M (total number of spots) = 150. This data gives a sample mean spot number of 2.5 per throw. In the limit of $N \to \infty$, the following two constraints apply:

$$\theta_1 + \theta_2 + \theta_3 = 1, \qquad \theta_1 + 2\theta_2 + 3\theta_3 = 2.5, \tag{18}$$

leaving only one free variable (e.g. θ_3). For finite N, the normalization constraint implies that there are two degrees of freedom–e.g. (θ_1, θ_2)–and the resulting posterior density for $N = 60$ and prior weights = 1 is shown in Fig. 2 for these variables. Note that this density forms a ridge along the line that is the solution to the constraint equations (18). The end points of this line are: (0,1/2,<1/2>) and (1/4,0,<3/4>), where the density goes to zero.

This example illustrates a general property of Bayesian inference applied to examples of this kind–the posterior density is concentrated around the subspace defined by the constraints, and becomes increasingly concentrated in that subspace as $N \to \infty$. However, within this subspace, the only information comes from the priors on the parameters–i.e., as $N \to \infty$, the data increasingly exclude parts of the total $\vec{\theta}$ space, but say nothing about the distribution within the allowed space.

Fig. 2 also illustrates another interesting property of this type of problem–that the standard estimators give misleading answers! For example, the maximum likelihood (ML) point is the end point (0,1/2,<1/2>). This particular set of θ values is the single

most likely set to give the result $M/N = 2.5$, yet the posterior density is minimal at this point! This apparently contradictory result comes about because other allowed points in the neighborhood of the ML point have much lower relative probability of giving the observed M/N, so the probability *density* of the neighborhood is small. Also, for this example, the maximum a posteriori (MAP) point may seem a better estimator, because the posterior density in Fig. 2 has a clear peak. However, the MAP estimate is not as good as it seems. This is because even though there is a peak along the constraint line, this does not take into account the width of the posterior density *orthogonal* to the constraint line. In fact, the width is minimal where the the peak is maximal, making the total volume of the density (integrated over the orthogonal dimension(s)) relatively flat. In the limit $N \to \infty$ this volumetric measure does not become concentrated around any particular point, and is only informed by the prior density, not the constraints (in the allowed subspace). The posterior means estimate, however, gives the best estimate for the probability of the next throw, as explained above.

The Bayesian Analysis of Example 3

In this example, there is a database of kangaroos, where there are four attributes for each kangaroo (handedness, beer-drinking, state-of-origin and color) which take on one of (2, 2, 7, 3) possible values respectively–e.g. a particular kangaroo might be right-handed, non-beer-drinking, Tasmanian and grey. This example is a generalization of that discussed in [5]. This data can be tabulated in a 4-dimensional array called a contingency table. In this example, we have a 4-dimensional contingency table with a total of 2x2x7x3 = 84 cells, where each cell contains the number of cases with the corresponding set of attribute values. It is always possible to sum rows or columns in a contingency table to produce marginal values (so called because these totals are written in the margins of the table). An example of a first order marginal is the total number of grey kangaroos, regardless of the other attributes (handedness etc.). This marginal can be found by summing over all the other attributes while holding the color attribute fixed at "grey". A second order marginal, for example, is the number of non-beer-drinking, left-handed kangaroos, found by summing over the other attributes. The zeroth order marginal is just the total number of observed kangaroos, N.

If we are given counts for *all* the n-dimensional cells in the contingency table, our goal is to use Bayes' theorem to find the posterior distribution on the full joint probability of all cells. These represent the true (but unknown) probability that a random case would fall into the corresponding cell. These cell probabilities are represented by $\theta_{i,j,k,l}$, where i, j, k, l is the value of the 1st., 2nd., 3rd., and 4th. attribute respectively. For example, $\theta_{1,2,7,3}$ is the probability that a kangaroo is right-handed, beer-drinking, from the state of Victoria and is brown. The cell data consists of all $n_{i,j,k,l}$s–i.e. the observed number of kangaroos with attribute values i, j, k, l for all possible combinations of values.

The situation in this example is the same as for example 1 above, where for each face probability, θ_i, we had the observed number of occurrences of the ith face (n_i)–i.e. the dice example is just a one dimensional contingency table! Because of this correspondence, the Bayesian analysis of example 3 is essentially the same as example

1. In particular, the likelihood is just the multinomial distribution, given by:

$$P(\vec{n}\,|\,\vec{\theta},I) = \frac{N!}{\prod_{i,j,k,l} n_{i,j,k,l}!}\prod_{i,j,k,l}\theta_{i,j,k.l}^{n_{i,j,k,l}} \tag{19}$$

$$N=\sum_{i,j,k,l} n_{i,j,k,l}, \qquad \forall\{i,j,k,l\},\ n_{i,j,k.l}\geq 0$$

If we again assume a Dirichlet conjugate prior probability distribution on the θ_is, we get the joint posterior distribution given by:

$$p(\vec{n},\vec{\theta}\,|\,\vec{w},I) \tag{20}$$

$$= \frac{N!}{\prod_{i,j,k,l} n_{i,j,k,l}!}\frac{\Gamma(W))}{\prod_{i,j,k,l}\Gamma(w_{i,j,k,l})}\prod_{i,j,k,l}\theta_{i,j,k,l}^{n_{i,j,k,l}+w_{i,j,k,l}-1}$$

$$\times\,\delta(\sum_{i,j,k,l}\theta_{i,j,k,l}-1), \tag{21}$$

$$W\equiv\sum_{i,j,k,l=1}^{k} w_i, \qquad \forall i,j,k,l,\ w_i>0$$

Examination of the Bayesian solution shows that for N large enough, so that there is a significant number of observations in every cell, the posterior density for each cell is strongly peaked around the empirical mean value $n_{i,j,k,l}/N$, and asymptotically approaches a delta function as N goes to ∞. In other words, the $n_{i,j,k,l}$s, are sufficient statistics for the $\theta_{i,j,k,l}$s, so that with increasing N, the estimated mean probabilities approach the true cell probabilities.

This example reveals a fundamental problem in statistics. Even for just the four attributes in this example, there are 84 cells in the full joint space, requiring many times that number of samples to get reasonable estimates for each cell. Since the total number of cells is the product of the sizes of all the attributes, and this number thus grows exponentially with the number of attributes, an extraordinary number of samples is required to estimate any particular probability with reasonable accuracy. One way to avoid this combinatorial dilemma is to make additional "independence" assumptions about attribute interactions, and this is exactly what MaxEnt or graphical models do, as explained below.

A more complex situation arises if instead of giving all the cell totals, only some of the marginals (and N) are given. Generally, the given marginal information is not sufficient to determine all the $n_{i,j.k,l}$s, so we have an underdetermined situation. The Bayesian analysis of this example is essentially identical to that of example 2. Each given marginal is a linear constraint on the possible $n_{i,j.k,l}$s. For example, if some $n_{i,j}$ is given, then the possible $n_{i,j.k,l}$ must obey the linear equation: $n_{i,j}=\sum_{k,l} n_{i,j.k,l}$. The given marginals are usually not sufficient to uniquely constrain the $n_{i,j.k,l}$s, so there are many possible sets of $n_{i,j.k,l}$s that satisfy these marginal constraints. The likelihood of each such set is given by the multinomial equation (19). The likelihood of the set of given marginals is then found by summing the likelihoods (19) of all sets of $n_{i,j.k,l}$s satisfying the marginals, in analogy to equation (13). Bayes rule can then be applied to this summed likelihood and

a prior over $\theta_{i,j,k,l}$s (e.g. the Dirichlet prior) to give a posterior over the $\theta_{i,j,k,l}$s, given the marginals. These calculations can be done exactly for small N, but otherwise are very messy.

As for example 2, the result of this analysis is that for $N \to \infty$, the posterior joint pdf on the $\theta_{i,j,k,l}$s is just the prior pdf in the subspace defined by the marginal constraints. That is, the constraints restrict the posterior pdf to the allowed subspace, but do not give any information on the distribution within this subspace. Only the prior density gives information in the subspace. For finite N, any $\vec{\theta}$ consistent with the normalizing constraint $\sum_{i,j,k,l} \theta_{i,j,k,l} = 1$ is possible, but the posterior density is concentrated around the subspace defined by the given empirical marginal constraints, such as: $\sum_{k,l} \theta_{i,j,k,l} = n_{i,j}/N$.

Thus we conclude that when only some marginal information is given, the Bayesian approach to contingency table analysis concentrates the joint posterior pdfs for $\vec{\theta}$ around the subspace defined by the empirical marginal constraints, but the prior on $\vec{\theta}$ is the only source of information within this subspace. This is not surprising, since a Bayesian analysis only reflects the information put into it, in this case, the given marginal information and the priors on $\vec{\theta}$.

THE MAXIMUM ENTROPY SOLUTION

The Classic MaxEnt Solution (CME)

The principle of Maximum Entropy (MaxEnt) is a method for using constraint information to find a set of point probability values, $\vec{P}$, that assumes the least (Shannon) information consistent with the given constraints. When the given constraints are insufficient to uniquely constrain $\vec{P}$ to particular point values, MaxEnt picks out the unique distribution that satisfies all the constraints and also maximizes the entropy.

In the case of a finite set of I mutually exclusive and exhaustive events, described by discrete probabilities P_i, the entropy is defined as:

$$\mathrm{H}(\vec{P}) = -\sum_{i=1}^{i=I} P_i \times Log P_i. \tag{22}$$

Given a set of J independent linear constraints, including the normalization (1), each of the form $\vec{A}_j \cdot \vec{P} = D_j$, with $J < I$, the maximum entropy distribution may be found by the following algorithm [3]: define the partition function:

$$\mathrm{Z}(\vec{\lambda}) \equiv \sum_{i}^{I} \exp(-\sum_{j}^{J} \lambda_j A_{ji}), \tag{23}$$

with the Lagrange multipliers $\vec{\lambda}$ determined by the set of J simultaneous equations:

$$\frac{\partial}{\partial \lambda_j} \log(\mathrm{Z}(\vec{\lambda})) + D_j = 0. \tag{24}$$

Then

$$\mathrm{H}_{\max} = \log(\mathrm{Z}(\vec{\lambda})) + \vec{\lambda} \cdot \vec{D}, \tag{25}$$

and the corresponding probability distribution is:

$$P_i = \mathrm{Z}(\vec{\lambda})^{-1} \exp(-\sum_j^J \lambda_j A_{ji}) = \mathrm{Z}(\vec{\lambda})^{-1} \prod_j \exp(-\lambda_j A_{ji}). \tag{26}$$

Explicit solutions for the dice case are given in [3]. These CME values are different from the MAP, ML and Posterior Mean estimates, reflecting the different assumptions built-in to each estimator.

CME for Case 3

In this example, we are given marginal constraints on a four dimensional contingency table, and we wish to find the CME distribution, $P_{i,j,k,l}$, subject to the constraints. Note that the CME goal is to find a set of point probabilities, not a joint pdf over the parameters, $\vec{P}$. An example of a first order marginal constraint is:

$$\sum_{j,k,l} P_{i,j,k,l} = P_i = n_i/N \tag{27}$$

Note that here we have used the observed marginal frequency n_i/N as the exact value for the marginal constraint. Clearly this is an approximation, referred to as the “constraint rule" in [4, 6]. We remove this approximation in our generalized MaxEnt approach (GME) below, where we translate uncertainty about the true value of the constraints into uncertainty over the MaxEnt values.

Also, note that the exact zeroth order marginal constraint always applies, i.e.

$$\sum_{i,j,k,l} P_{i,j,k,l} = 1. \tag{28}$$

If, for example, we are given the first order marginals on each of the attributes, the CME solution is:

$$P_{i,j,k,l} = P_i P_j P_k P_l; \tag{29}$$

i.e., the CME solution is just the marginal independence solution. This CME solution says that if we are told that a kangaroo is left-handed, for example, this tells us nothing about whether it is also a beer-drinker or from Tasmania, in accord with intuition.

If we are given all the first order marginals and the following second order marginals: $P_{i,j}, P_{j,k}, P_{k,l}$; then the CME solution is:

$$P_{i,j,k,l} = \frac{P_{i,j} P_{j,k} P_{k,l}}{P_j P_k}, \tag{30}$$

which can be rearranged into:

$$P_{i,j,k,l} = P_i P(j|i) P(k|j) P(l|k), \tag{31}$$

where $P(x|y)$ is the conditional probability of x given y. This is an example of conditional independence, where the kth attribute is independent of the ith attribute given j. In other words, the kth attribute only depends on the ith attribute indirectly through attribute j. Similarly, the lth attribute only depends on the ith attribute indirectly through attributes j and k.

If we are given $P_{i,l}$ in addition to the above second order marginal (so that the constraints form a cycle), there is no closed form CME solution or corresponding independence model! However, it is always possible to find a numeric CME solution for any set of consistent constraints. An efficient method for computing CME values for sets of constraints is given in [7]. Generally, CME calculations are efficient if the constraints form a tree-like (hyper)graph, allowing a factorization of the joint probability.

The above examples show that CME gives independence or conditional independence results when the given constraints have a tree-like dependence, but CME is a *general principle of independence* as it always returns a result for any consistent set of constraints regardless of whether a factorized product form exists or not. The connection between CME and independence (product) forms, such as (29) and (30) is clear from the product form of CME shown in equation (26). It is this generalized independence property that makes MaxEnt inference a powerful tool of statistical inference, as show below.

Generalized MaxEnt

Equations (22)–(26) show that the maximum entropy point probabilities P_i are a *function* of the constraint values D_j. If these D_js are estimated from a sample, then Bayes implies that our knowledge of their values is expressed as pdfs (typically Gaussians), whose width deceases with increasing sample size. Using the Jacobian of the function relating the P_is to the D_js, the joint posterior pdf on the D_js can in principle be transformed into a joint pdf on the P_is. We call this mapping of uncertainty in the constraint values into uncertainty in the MaxEnt values (expressed as pdfs) the Generalized principle of Maximum Entropy (GME). We have not previously seen this generalization in the literature, but it is a direct result of applying probability theory to the situation. Note that the normalization constraint (1) is true by definition, so there is no uncertainty associated with its value. Also, in the limit of $N \to \infty$, the constraint values are given exactly, so GME becomes the same as CME in this case.

Because GME generates a pdf over the θ_is, it can be compared to the results of a Bayesian inductive inference, such as equation (6). Jaynes, [2], has claimed that Bayesian and MaxEnt analyses cannot be compared because they operate on different spaces. We have shown above that because GME returns a joint pdf over $\vec{\theta}$, this output *can* be directly compared to a joint Bayesian posterior pdf over $\vec{\theta}$. Also, Bayesian predictive inference–e.g., equation (11)–can be compared directly with CME values (point probabilities). For the dice example 2, the GME pdfs are peaked around the CME values, and generally are very different from the Bayesian posterior pdfs, which depend only on the prior pdfs in the constrained subspace. The difference between the Bayesian and GME pdfs is not surprising, because these analyses make different assumptions. The

Bayesian analysis does not make any independence or completeness assumptions, while GME does.

MaxEnt analysis (both CME and GME) only uses the constraint information (priors are ignored), and makes a strong *completeness* assumption. That is, MaxEnt assumes that the given constraints are the *only* ones that apply. Also, MaxEnt implicitly assumes independence of components, unless dependencies are explicitly given as constraints. If the current MaxEnt predictions do not agree with observations, this is a signal that the current constraints are incomplete, and the differences can suggest missing constraints. We call searching for additional constraints by testing the MaxEnt predictions of the current constraint set against observations the method of Maximum Entropy Inference (MEI). MEI expands the constraint set if significant deviations are detected[4].

COMPARISON OF BAYESIAN AND MAXENT INFERENCE (MEI)

MEI is a method of inductive inference comparable to Bayesian inductive inference. Because MEI makes stronger assumptions than Bayes, at any stage it's predictions can be wrong. In Bayesian inference, on the other hand, the posterior only differs from the prior is the evidence in the data warrants the change (no backtracking). Also, MEI is not incremental. In MEI, discovery of a new constraint invalidates previous inferences–i.e. previous MEI results are in error, and cannot be used as priors for a new calculation, as done in Bayesian inference. Also, if there is reason for believing that the current constraint set is incomplete, there is no good reason for believing the current MaxEnt predictions. In particular, in the dice case 2, MaxEnt predictions based on the mean value and normalizing constraint are likely to be wrong, because these constraints are unlikely to be complete. In view of these differences, Bayesian inductive inference generally seems superior to MEI, because it does not assume any information not explicitly given. For small problems, we believe this to be true, but for larger problems, independence assumptions are *essential*, as explained below. Since MaxEnt is a general principle of independence, it is a convenient method for implicitly introducing independence consistent with the given constraints.

For contingency tables with a large number of attributes, the Bayesian analysis developed above is essentially useless, because there will rarely be enough cases to give good statistics for every cell. This is a direct consequence of our Bayesian analysis not using any more information than the prior on $\vec{\theta}$ and the cell counts. MEI, by comparison, implicitly makes strong independence assumptions, and so is able to provide strong predictions for the cell probabilities, even when the given marginals are far too few to uniquely constrain them. Among all possible methods for assigning probabilities to the cells, MaxEnt assumes the least additional information, and this corresponds to making generalized independence assumptions. However, MaxEnt, even GME, is still assuming far more information than given in the statement of the problem. In particular, MaxEnt

[4] In principle, the correct approach is weight a constraint according to the probability that it differs significantly from the current MaxEnt expectation, rather than an all or nothing inclusion. However, in practice this refinement has little effect on the numerical predictions

amounts to a very strong prior belief in independence, given the known dependencies (constraints). MaxEnt also assumes that the given constraints (dependencies) are the only ones that are operating–i.e. the completeness assumption. The use of the “Null Hypothesis" as the reference standard in classical statistics is an example of implicitly using a very strong prior on independence!

It is common practice in Bayesian analysis of contingency table data to assume a prior independence model, known in the literature as graphical models (see [8] for details). These graphical models amount to an *explicit* assertion of what dependencies (and independencies) exist in the domain through the form of the likelihood function. In some cases, a MaxEnt joint distribution, such as equation 30), has the same form as an assumed graphical model, and sometimes not. In particular, the causal interpretation of some graphical models differ from MaxEnt with the same constraints. This is because causality assumes more information than independence. Also, the number of marginals needed to parameterize a graphical model is usually fixed by the assumed model structure, whereas MaxEnt produces a joint probability using whatever marginals are given, and so is less demanding of information. Even when the graphical models are learned from the data itself, there is no guarantee that the learned structure is complete, as for MEI also. In summary, inductive inference in large domains requires independence assumptions. These assumptions are built into MaxEnt, but must be inserted explicitly in Bayesian inference using dependency models, such as graphical models. If there is strong structure in a domain, then both MEI and Bayesian graphical models will find it. However, generally the structure learned by the two methods can differ, although the predictions from the different methods tend to be very close in practice. The differences arise primarily because MaxEnt does not use prior information, whereas Bayesian inference does.

DISCUSSION

A major question arising out of the above is: “Why does making such strong independence assumptions work in practice?" Our answer to this question is mostly speculative. Firstly, the world we live in *is* largely independent! What is happening in this room is nearly independent of what is happening in the next room. In principle, all observable attributes can depend on all others, but in practice, most observables depend only on a few causes (observed or not). MEI applied to contingency tables works because a small subset of the possible marginals *do* typically capture the real information in data sets we typically investigate. Any initially missing marginal constraints are discovered during MEI or similar graphical model search. More importantly, any statistical or scientific analysis implicitly assumes that any attributes not explicitly considered are irrelevant–i.e. independent of the ones considered. In other words, independence assumptions must be invoked merely to frame a problem! At the other extreme, in a world in which everything strongly depends on everything else, inference is impossible. This is partly because probabilistic inference becomes computationally intractable, but also because the observer is part of the system, and so the act of doing inference changes the system, thus creating a kind of Heisenberg uncertainty.

The second reason independence assumptions work well for contingency table analysis is more subtle. Marginal constraints, such as equation (27), are constraints on a subset of the cell parameters $P_{i,j,k,l}$, but not *arbitrary* subsets. Such marginal constraints represent probabilities of the corresponding attribute(s) value(s), not some meaningless mixture of values drawn from many attributes. In other words, out of all the possible linear constraints we could have imposed, the ones we actually test correspond to meaningful observable combinations. This selective testing of possible constraints is important prior knowledge related to our causal understanding of the world. This is part of the reason that the relatively few constraints imposed in a MaxEnt contingency table analysis seem to capture all or nearly all of the relevant information. More generally, it is our causal world models that provide extremely strong priors on what dependencies exist, with everything else assumed (conditionally) independent. MEI is a good tool for finding new dependencies (constraints), and generating predictions based on the known dependencies, while assuming independence for the rest. However, in using MEI, the possible incompleteness of the current constraints and the corresponding inaccuracy of the predictions should be kept in mind.

A more fundamental question is whether the world is really largely independent, or do we just assume it to be so? Our current understanding of physics indicates that causal propagation of effects tend to rapidly die out because of dilution–both light and sound spread out, free energy degrades to heat and the universe expands. We speculate that this dilution effect is the basic reason for the quasi-independence/causal structure of the world we live in, and so is the reason we can make useful predictions.

Other authors have examined the relation between MaxEnt and Bayes, and have come to different conclusions than ours. In particular, Jaynes [9], makes the claim that these two methods of inference are complementary, and in particular, MaxEnt can be used to generate priors for a subsequent Bayesian inference. We believe otherwise, because we have shown that the two methods of inference make fundamentally different assumptions. More particularly, the pdfs produced by GME incorporate the strong prior assumptions of generalized independence and constraint completeness. These assumptions produce GME pdfs that are concentrated around CME values (based on empirical constraints), and become more concentrated with larger N (see [3] for details). As a result, it would take an extraordinary amount of data to shift away from these pdfs in a Bayesian update if the completeness assumption assumed in GME is incorrect. Because CME produces point probabilities, not pdfs, they are useless as Bayesian priors, because no amount of data in a Bayesian update can change them.

SUMMARY

We have shown that it is possible to directly compare the results of MaxEnt and Bayesian inference using the same information, despite previous claims to the contrary, and that the results are different, reflecting the different assumptions built into the different methods. To correctly apply MaxEnt to situations where the problem constraint values are estimated from data, rather than being given exactly, we generalize the standard MaxEnt framework (GME). GME maps the uncertainty in the constraint values (represented as

pdfs) into a pdf over the model parameter space. Classic MaxEnt (CME) only produces point probabilities in this space, which cannot be compared to posterior probability pdfs.

Both GME and CME make strong assumptions about independence through maximization of the entropy, subject to the given constraints. Both also implicitly assume that the given constraints are the *only* constraints that apply (the completeness assumption). Bayesian inductive inference, on the other hand, does not assume any more information than the priors and given data within the model space defined by the likelihood function. Independence assumptions can be inserted into a Bayesian analysis by an appropriately factorized likelihood, so that in some cases the Bayesian and GME results are very similar. Since GME does not use a prior distribution, they cannot be the same. We also show by examples that common probability estimators give different results, reflecting their different assumptions.

REFERENCES

1. Jaynes, E. T., "Brandis Lectures: Information Theory and Statistical Mechanics", in *E. T. Jaynes: Papers on Probability, Statistics, and Statistical Physics*, edited by R. D. Rosenkrantz, D. Reidel, Dordrecht, 1983, chap. 4, pp. 39–76, http://bayes.wustl.edu/.
2. Jaynes, E. T., "Where do We Stand on Maximum Entropy", in *The Maximum Entropy Formalism*, edited by R. D. Levine and M. Tribus, MIT Press, Cambridge, MA, USA, 1978, pp. 15–118, http://bayes.wustl.edu/.
3. Jaynes, E. T., "Concentration of Distributions at Entropy Maxima", in *E. T. Jaynes: Papers on Probability, Statistics, and Statistical Physics*, edited by R. D. Rosenkrantz, D. Reidel, Dordrecht, 1983, chap. 11, pp. 315–336, http://bayes.wustl.edu/.
4. Uffink, J., *Studies in History and Philosophy of modern Physics*, **27**, 47–79 (1996), http://www.phys.uu.nl/ wwwgrnsl/jos/.
5. Jaynes, E. T., "Monkeys, Kangaroos and N", in *Maximum-Entropy and Bayesian Methods in Applied Statistics*, edited by J. H. Justice, Cambridge Univ. Press, Cambridge, UK, 1986, pp. 26–, updated (1996) version at http://bayes.wustl.edu/.
6. Uffink, J., *Studies in History and Philosophy of modern Physics*, **26b**, 223–26 (1995), http://www.phys.uu.nl/ wwwgrnsl/jos/.
7. Cheeseman, P., "A Method of Computing Generalized Bayesian Probability Values for Expert Systems", in *Proceedings of the Eighth International Joint Conference on Artificial Intelligence*, edited by A. Bundy, International Joint Conference on Artificial Intelligence Inc., Morgan Kaufmann, Los Altos, CA, USA, 1983, pp. 198–202.
8. Jordan, M. I., *Learning in Graphical Models*, MIT Press, Cambridge, MA, USA, 1998.
9. Jaynes, E. T., *Probability Theory: the Logic of Science*, Cambridge University Press, Cambridge, UK, 2003, http://bayes.wustl.edu/.

addition of constraints in MEI