Bayesian Statistical Analysis

Samir Chettri

October 7, 2005

Abstract

There are a great many books on Bayesian statistics available these days[]. This tutorial is being written since many of our papers and future work in gene arrays rely upon repeated use of these methods. Rather than continually have our readers refer to a variety of textbooks, we thought it best to have a reasonably complete introduction to the subject in one place.

Keywords: Bayes Theorem, parameter estimation, model selection, Bayes networks, remote sensing.

Contents

1 Principles of Bayesian Analysis

. . . you cannot do inference without making assumptions.
David MacKay [56].

Need some material here.

1.1 The notion of probability

1.2 Deductive logic and inference

In order to present the specifics of Bayesian methods, we will make passage through Aristotlean logic [27] and inference. We are taught that propositions are statements that may become true or false but not both. For example a proposition A could be "the color of the bowl is blue," and its negation would be $\overline{A}$, "the color of the bowl is not blue." A particular type of statement is the conditional $A \rightarrow B$; this is read as "A implies B," or as "if A then B." Another important statement is that the contrapositive of the conditional statement is logically equivalent to the conditional statement, i.e., $\overline{B} \rightarrow \overline{A}$, read as "if not B, then not A," is the same as $A \rightarrow B$.

A key aspect of propositional logic is the deductive syllogism which has the structure: major premise, minor premise and conclusion. *Modus ponens* (Latin: method of affirmation) can be defined as:

$$\begin{array}{rll} A \rightarrow B & & \texttt{major premise} \\ A & \text{is true} & \texttt{minor premise} \\ \therefore B & \text{is true} & \texttt{conclusion.} \end{array} \tag{1}$$

Similarly another deductive syllogism is *modus tollens* (Latin: method of denying):

$$\begin{array}{r} A \rightarrow B \\ \overline{B} \\ \therefore \overline{A} \end{array} \tag{2}$$

A little bit of Boolean algebra [58] is also necessary in order to understand the next section. We have already introduced A and $\overline{A}$ as a proposition and its negation respectively. Negation is *unary* operator in Boolean algebra since it takes a single proposition. *Binary* operations take two propositions: Two of them are used very often, conjunction and disjunction, written in order as.

$$\begin{aligned} AB &= \text{both } A \text{ and } B \text{ are true.} \\ A+B &= \text{at least one of } A \text{ or } B \text{ are true.} \end{aligned}$$

For conjunction we sometimes write (A, B) or $A \cap B$ while for disjunction we have $A \cup B$. Depending on the context we might use any of the above notations.

There are are several laws satisfied by propositions (or more generally sets) in the Boolean algebra. For example [27] there are the commutative laws, absorption laws etc. The one we will have occasion to use later is the distributive law for sets A, B, C [40]:

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C). \tag{3}$$

A more general form is written below:

$$A \cap \left(\bigcup_i B_i \right) = (A \cap B_1) \cup (A \cap B_2) \cup (A \cap B_3) \cup \ldots, \tag{4}$$

which can be derived by repeated application of equation (3). We will also use the associative law [27]:

$$A \cap B \cap C = (A \cap B) \cap C = A \cap (B \cap C). \tag{5}$$

Finally, we will need some simple theorems: If we are given a proposition D and:

$$\begin{aligned} \overline{B} &= AD \text{ then} \\ A\overline{B} &= \overline{B} \text{ and} \\ B\overline{A} &= \overline{A}. \end{aligned} \tag{6}$$

In inference we are interested in syllogisms that are different from modus ponens or modus tollens. Here are some examples ([46], page 5):

$$\begin{aligned} A \to B & \\ B & \\ \therefore A & \quad \text{is more plausible,} \end{aligned} \tag{7}$$

$$\begin{aligned} A \to B & \\ \overline{A} & \\ \therefore B & \quad \text{is less plausible.} \end{aligned} \tag{8}$$

and

$$\begin{aligned} A \to B & \quad \text{is more plausible} \\ B & \\ \therefore A & \quad \text{is more plausible} \end{aligned} \tag{9}$$

Cox's theorem provides us the means by which we can perform inference of the sort described in equations (7), (8) and (9).

1.3 Cox's Theorem

The unique aspect of Cox's [1] theorem is that starting from two simple desiderata (he called them axioms) we can derive the sum and product rules of probability and from these the weak syllogisms listed in the prevous subsection can be reproduced. In this paper we will work with Jaynes version [46] of Cox's theorem and therefore state three desiderata below. In reading the desiderata, note that we do not use the word probability anywhere, but rather the word plausibility [46] which one can equate with degrees of belief [56]. Probabilities arise naturally from the desiderata [45]. Books and papers that have substantial discussions of this theorem are [46], [88], [66], [78], [31] and [90] and of course Cox's original work [19, 20].

In order to proceed, we introduce some notation originally used in [78, 46]. We call $u(A|B)$ the conditional plausibility that A is true, given that B is true. Also, the symbol I will always represent our prior information which is everything we know about the problem at hand. Thus $u(A|I)$ is the plausibility of A given our prior knowledge I. Similarly, $u(A|BI)$ is the conditional plausibilty of A given that the prior knowledge I and the datum B are both true. Therefore the weak syllogism identified by equation (7) can be rewritten as:

$$\begin{aligned} A &\rightarrow B \\ &B \\ \therefore u(A|BI) &\geq u(A|I) \end{aligned} \tag{10}$$

All we are saying in the above equations is that the plausibility that A is true given both B and our prior I are true, is greater or equal to the plausibility of A given just our prior knowledge I. If I is irrelevant, then the plausibilities should be identical. **Check to see if the previous statement should read, "if B is irrelevant." I believe it is true, since if I is irrelevant, then if I indicates our relationship between A and B, i.e., $A \rightarrow B$, and we are saying that I is irrelevant, then we are saying that we don't know if A implies B. In that case knowledge that B is true, does not affect the plausibility of A and $u(A|B) = u(A)$.**

1.3.1 The desiderata

Given the preceding discussion, the desiderata are:

Desideratum 1 *The measures of plausibilities are real numbers.*

This informs us that if $u(A) > u(B)$ and $u(B) > u(C)$ then $u(A) > u(c)$.

Desideratum 2 *Plausibilities must exhibit qualitative agreement with rationality.*

This desideratum suggests that the plausibility of a logical proposition A and its negation $\bar{A}$ are related [56, 20], and that as we get more and more information on a proposition then the number representing the plausibility will increase continuously and monotonically [78].

Desideratum 3 *All rules relating plausibilities must be consistent.*

The consistency desideratum can be viewed in two ways Jaynes [45]:

[1] Richard Threlkeld Cox was born in 1898 in Portland, Oregon. He obtained his doctorate in Physics in 1924 from John Hopkins University and subsequently took a position at New York University. He returned to JHU in 1943 and retired from there in 1964. While he worked in different technical areas in physics; today physicists who are interested in statistics chiefly remember him for his work in inference as embodied in [19] and the book [20] which is an expansion of the first reference. We have obtained biographical information on Cox from [89].

1. If a result can be arrived at in more than one way, then each way must provide the same result.
2. As we pass to the limit where propositions become true or untrue (as opposed to plausible), our mathematical formulae must reduce to deductive reasoning.

In order to investigate this desideratum we use the tree diagram in Figure 1.

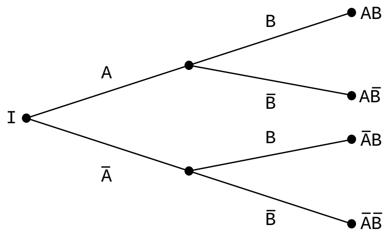

Figure 1: Consistency of plausibilities.

We are interested in the truth of the proposition AB. Since AB can only be reached through A and B, we can state that $u(AB|I)$ will be some function of $u(A|I)$ and $u(B|AI)$ or

$$\begin{aligned} u(AB|I) &= F\left[u(A|I), u(B|AI)\right] \\ u(BA|I) &= F\left[u(B|I), u(A|BI)\right]. \end{aligned} \tag{11}$$

Here, the second of Equations (11) is obtained from the first by interchange of A and B and is necessary by our desideratum of consistency wherein if we can reach a conclusion in more than one way, then all ways must lead to the same result.

1.3.2 The product rule

Using equation (5) i.e., $(A \cap B) \cap C$ and $A \cap (B \cap C)$ and substituting in the first of Equations (11) gives us

$$\begin{aligned} u(ABC|I) &= F\left[u(AB|I), u(C|ABI)\right] \\ &= F\left[u(A|I), u(BC|AI)\right]. \end{aligned} \tag{12}$$

Recursively applying the first of Equations (11) to Equations (12) and making the substitutions $x = u(A|I)$, $y = u(B|AI)$ and $z = u(C|ABI)$ leads us to a functional equation known as the associativity equation: [2]

$$F\left[F(x,y), z\right] = F\left[x, F(y,z)\right]. \tag{13}$$

The solution to equation (13) as given in [1] is $F(x,y) = G^{-1}\left[G(x)G(y)\right]$. After substituting for x, y, z in equations (13) and the first equation in (11), we have

$$G\left[u(AB|I)\right] = G\left[u(A|I)\right] G\left[u(B|AI)\right]. \tag{14}$$

Defining $v(A|I) = G\left[u(A|I)\right]$ and using equations (14) and the second equation in (11) gives us:

$$\begin{aligned} v(AB|I) &= v(A|I)v(B|AI) \\ &= v(B|I)v(A|BI). \end{aligned} \tag{15}$$

Astute readers will note the similarity of equations (15) to the product rule of probability. Note that we are not yet using the term probability but are working with some function of the plausibilities whose scale hasn't been determined yet. We consider the limit where our prior information I determines that A is true. Then

[2]Known to Abel in 1826 and discussed in some detail in [1].

$(AB|I) = (B|I)$ and $(A|BI) = (A|C) = (A|A)$. Substituting in equation (15) gives us $v(B|I) = v(A|A)v(B|I)$ which must hold for all $v(B|I) \neq 0$. Hence we can say that the plausibility of a true proposition must be equal to unity.

In order to further limit v we consider the case when our prior information I tells us that B is false. [3] Then (AB) (or $A \cap B$) has the same truth value as the false proposition and $v(B|AI) = v(\mathbf{F}|AI) = v(\mathbf{F}|I)$, where $\mathbf{F}$ stands for falsehood. Using equation (15) with these results gives us $v(\mathbf{F}|I) = v(A|I)v(\mathbf{F}|I)$. For arbitrary $v(A|I)$ there are three choices for $v(\mathbf{F}|I)$ that satisfy this equation, $v(\mathbf{F}|I) = 0$ or $\pm\infty$. As a convention we choose falsehood to be $v(\mathbf{F}|I) = 0$: Therefore $0 \leq v \leq 1$.

1.3.3 The sum rule

By desideratum 2, the plausibility of a logical proposition A and its negation $\overline{A}$ are related, i.e.:

$$v(\overline{A}|I) = T\left[v(A|I)\right]. \tag{16}$$

Since A and $\overline{A}$ are reciprocally related we must have

$$v(A|I) = T\left[v(\overline{A}|I)\right]. \tag{17}$$

Hence our function $T(x)$ must satisfy $T\left[T(x)\right] = x$, i.e., $T(x)$ is self reciprocal.

Another condition that $T(x)$ must satisfy is derived by applying the product rule, equation (15) and our equation for desideratum 2, equation (16) in succession to get:

$$v(AB|I) = v(B|I)\; T\left[\frac{v(\overline{A}B|I)}{v(B|I)}\right]. \tag{18}$$

Also, since $v(AB|I)$ is symmetric in A, B, we also have

$$v(AB|I) = v(A|I)\; T\left[\frac{v(\overline{B}A|I)}{v(A|I)}\right]. \tag{19}$$

Equating the right–hand sides of equations (18, 19) immediately gives us:

$$v(A|I)\; T\left[\frac{v(\overline{B}A|I)}{v(A|I)}\right] = v(B|I)\; T\left[\frac{v(\overline{A}B|I)}{v(B|I)}\right]. \tag{20}$$

The above must hold for all propositions A, B, I and so must hold for a proposition $\overline{B} = AD$. Therefore by the results of equation (6) and equation (16) we get:

$$v(A\overline{B}|I) = v(\overline{B}|I) = S\left[v(B|I)\right] \tag{21}$$
$$v(B\overline{A}|I) = v(\overline{A}|I) = S\left[v(A|I)\right] \tag{22}$$

Defining $x \equiv v(A|I)$ and $y \equiv v(B|I)$ we see that equation (20) reduces to a functional equation:

$$xT\left[\frac{T(y)}{x}\right] = yT\left[\frac{T(x)}{y}\right], \tag{23}$$

the solution to which was given in [19, 20, 46] and is:

$$\left[T(x)\right]^n + x^n = 1. \tag{24}$$

[3] Jaynes [45] considers the case when I tells us that A is false. We prefer the presentation in [78] – the case under consideration and which is equivalent to *modus tollens*.

Using our definition for x, i.e., $x \equiv v(A|I)$ and equation (16) we get:

$$[v(A|I)]^n + \left[v(\overline{A}|I)\right]^n = 1. \tag{25}$$

This is nothing but the sum rule in *mufti.* Note, that we could well have written the product rule, equation (15) in powers of n and by defining $p(x) \equiv v^m(x)$ we therefore get the product and sum rules of probability:

$$p(AB|I) = p(A|I)p(B|AI) = p(B|I)p(A|BI) \tag{26}$$

$$p(A|I) + p(\overline{A}|I) = 1. \tag{27}$$

1.3.4 Is it consistent?

In order to check whether our sum and product rules work with deductive logic as defined in (1) and (1) (modus ponens and modus tollens respectively), we start by defining $I \equiv A \to B$. Then (1) corresponds to our use of the product rule as

$$p(B|AI) = \frac{p(AB|I)}{p(A|I)}.$$

But, $p(AB|I) = p(A|I)$ (the probability that A is true and B is true given our background knowledge $I \equiv A \to B$, is the same thing as saying A is true with our background knowledge I). Therefore $p(B|AI) = 1$ which is exactly what is expected by modus ponens. One can make similar arguments for modus tollens [46].

In order to complete this subsection let us take a look at the weak syllogism (7) which is the product rule written like

$$p(A|BI) = p(A|I)\frac{p(B|AI)}{p(B|I)}.$$

However, $p(B|AI) = 1$, i.e., this is just modus ponens, equation (1), and by definition $p(B|I) \le 1$, which gives us

$$p(A|BI) \ge p(A|I),$$

exactly what our syllogism implies. Similar reasoning gives us weak syllogisms (8) and (9).

At this point it is important to step back and consider what has been achieved. We quote [31]: "The sum and product rules have been known for centuries to apply to proportions (relative frequencies), but this derivation places them on a different, deeper foundation. Our theory is a generalization of deductive, boolean logic, the new ingredient being the notion of the extent to which truth of one proposition is implied by truth of another, and its quantification on a continuous scale." We have not only shown that our version of probability theory works with deductive logic, but it also works well with weaker syllogisms that were never even considered by Aristotle. This is a conceptual advance because it frees us from the need to treat probabilities only as frequencies and instead to consider them as degrees of belief.

1.4 Bayes Theorem and marginalization

In one sense, Bayes Theorem is a trivial restatement of the product rule, equation (26),

$$p(H|D, I) = \frac{p(H|I)p(D|H, I)}{P(D|I)}. \tag{28}$$

However, its implications are enormous. Based on our discussion of Cox's theorem we now know that it is of great importance in problems of inference. In order to accomodate our future work, we change the notation used in equation (26) in order to accomodate our work in future sections: H now stands for a hypothesis,

D stands for the data, I is our prior knowledge, whose symbol remains unchanged. Thus Bayes theorem is read as

$$p(hypothesis|data, I) \propto p(hypothesis|I) \times p(data|hypothesis, I). \tag{29}$$

Though we have not included $p(data|I)$ above we will have more on this in our discussion of model selection, section 3. In the literature this term is sometimes called the *evidence* [77, 56, 54], but this is certainly not standard as there are just as many references that do not use the term [71, 55, 13]. We will use it throughout this presentation.

The other terms have names that are consistently used in practise: $p(H|D, I)$ is the *posterior probability* of the hypothesis, H, given the data and our prior information and $p(H|I)$ is the *prior probability* of the hypothesis given our background information I. $p(D|H, I)$ requires special mention: It is called the *likelihood* [4] when the data are constant (i.e., there is only one data set) and the hypothesis is varied, i.e., the hypothesis H is a parameter θ that we are attempting to find or characterize, and is called the *sampling distribution* when different sets of data are considered and the hypothesis is held constant [13]. Thus Bayes Theorem reduces to

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}, \tag{30}$$

or in the case where we are only interested in the parameters we have

$$\text{posterior} \propto \text{likelihood} \times \text{prior}. \tag{31}$$

In order to study marginalization we start with a large number of hypotheses $H_1, H_2, \ldots H_n$ that are mutually exclusive and exhaustive. This means that we can write:

$$p(H_1 + H_2 + \ldots |I) = p\left(\sum_i H_i|I\right) = 1. \tag{32}$$

Given a proposition A, H_i, our prior knowledge I and our data D, consider the following proposition: $(A, \sum_i H_i|D, I) = (A \text{ and } \sum_i H_i|D, I) = ((A \text{ and } H_1) \text{ or } (A \text{ and } H_2) \ldots |D, I) = ((A \cap H_1) \cup (A \cap H_2) \ldots |D, I)$, where we have used the general distributive law of equation (4) to expand $(A, \sum_i H_i|D, I)$. The probability of this proposition is

$$p(A, \sum_i H_i|D, I) = p(A|D, I)p(\sum_i H_i|A, D, I) = p(A|D, I). \tag{33}$$

The terms after the first equal sign are obtained from the product theorem. The term $p(\sum_i H_i|A, D, I)$ is unity by virtue of the fact that our hypotheses H_i, are exhaustive giving us our final term in the above equation. Let us continue with $p(A, \sum_i H_i|D, I)$. This can be written as

$$\begin{aligned} p(A, \sum_i H_i|D, I) &= p\left[(A \text{ and } H_1) \text{ or } (A \text{ and } H_2) \ldots |D, I\right] \\ &= p(A \cap H_1|D, I) + p(A \cap H_1|D, I) + \ldots \\ &= \sum_i p(A, H_i|D, I) \end{aligned} \tag{34}$$

If we equate equations (33, 34) we get our final result

$$p(A|D, I) = \sum_i p(A, H_i|D, I). \tag{35}$$

Now we have considered discrete H, i.e., $H_1, H_2 \ldots$. In the case we have continuous H, the sums go to integrals and we have [13]

$$p(A|D, I) = \int p(A, H|D, I)dH. \tag{36}$$

We now give the reader an interpretation of equation (33). Let A and H be the parameters, but we are only interested in the probability distribution of A. In common terminology H is called a nuisance parameter [13] - and this is integrated out. The term $p(A|D, I)$ is the marginal density of A, hence the term marginalization.

[4] Mackay [56] calls this the *likelihood of the parameters* or *the likelihood of the parameters given the data*.

2 Parameter Estimation

In this section we will work through a series of problems in higher and higher dimensions in order to explain how we can use Bayes Theorem in parameter estimation. We first choose a historical problem, first solved by Reverend Thomas Bayes more than two hundred years ago using the binomial distribution. We follow up with a two dimensional problem using the Cauchy density and finally end the section with detecting signals in noisy sinusoidal data when the noise is Gaussian. The reason for this sequence of problems is because the power of Bayesian methods becomes manifest with each iteration.

2.1 The binomial distribution

The biased coin is the unicorn of probability theory – everybody has heard of it, but it has never been spotted in the flesh.

Andrew Gelman & Deborah Nolan [32].

There are good reasons to choose the binomial as our first example. It has an important place in the study of statistics - it is used in the study of coin tossing experiments and is therefore of pedagogical value. It was studied in great detail by Bernoulli and de Moivre and it is possible that the failure of these two illustrious mathematicians in solving the problem of inverse probability is what led Thomas Bayes to a successful attack in his historic paper [5]. We will have more to say about Bayes paper and the binomial in Appendix B.

When we toss a coin many times we anticipate that, on average, we will get as many heads as tails. However, suppose we have reason to believe that the coin is biased and that we have a unicorn on our hands. A natural question that arises is - how fair is the coin?

In order to answer the question let us introduce θ - a continuous variable that represents the bias of the coin, $0 \leq \theta \leq 1$. $\theta = 0$ represents a two tailed coin, i.e., every coin flip produces a tail. Therefore, $\theta = 1$ represents the two headed case and $\theta = 0.5$ is a fair coin. Our goal will be to obtain θ after a number of tosses of this coin.

We start by using Bayes Theorem, equation (28):

$$p(\theta|D, I) = \frac{p(\theta|I)p(D|\theta, I)}{P(D|I)}. \tag{37}$$

The prior probability, $p(\theta|I)$, represents what we know about this coin. If we know nothing about this coin, then we would say that any value of θ is equally likely, i.e., we would choose a uniform distribution. On the other hand if we had strong reason to believe that the coin was fair, then we would choose a prior that was peaked around $\theta = 0.5$ and would assign a very low probability to other cases including the two tails and two heads one. Assuming uniform prior gives us:

$$p(\theta|I) = \begin{cases} 1 & \text{if } 0 \leq \theta \leq 1 \\ 0 & \text{otherwise.} \end{cases} \tag{38}$$

To obtain the likelihood, consider the fact that we have tossed the coin N times and have obtained r heads and $s = n - r$ tails. Assume that each trial (coin toss) is independent, this then gives us:

$$p(D|\theta, I) = \,^{n}C_r \, \theta^r (1 - \theta)^s, \tag{39}$$

where, $^{n}C_r$ (read "n choose r") is a standard combinatorial identity $\frac{n!}{r!(n-r)!}$, and we have used the notion of Bernoulli trials [65] to get the binomial distribution in (39). In order to obtain the posterior density we need to obtain $p(D|I)$ but this is just a normalizing factor in order to make the area under $p(\theta|D, I)$ equal to unity. Hence the posterior is

$$p(\theta|D,I) = \frac{{}^nC_r\,\theta^r(1-\theta)^s}{\int_0^1 \ {}^nC_r\,\theta^r(1-\theta)^s d\theta}. \tag{40}$$

Equation (40) is the incomplete Beta function and we can numerically evaluate it [72], provide asymptotic expansions [83], or even use tables [68] so as to get the posterior.

Figure 2 shows examples of an experiment we ran. We simulated the tosses of a coin assuming with a uniform prior. As the computer experiment proceeded we evaluated the posterior, equation (40) and drew the associated PDF for $n = 1, 2, 4, 8 \ldots 4096$ repetitions. In the figure we see that for upto four tosses we only got tails and so our PDF kept changing to represent to represent more and more certainty in the fact that this is a two tailed coin. The appearance of a single head changes the shape of the PDF to reflect this after eight trials. Eventually, we get sharper and sharper peaked distributions around $\theta = 0.4$ which is how we biased our coin.

Need coin tossing problem with Lindley prior instead of the uniform prior.

2.2 The Cauchy distribution

...by the nineteenth century it had acquired a singular name "the witch of Agnesi," a name that is still commonly found in dictionaries and encyclopaedias.

Stephen Stigler [84]

Consider Figure 3 which represents a lighthouse emitting light flashes in the X–Y plane. These light flashes are emitted randomly and therefore leave the lighthouse at random azimuths θ_i. The radiation hits a detector that runs along the X axis at position x_i. We are not given the random azimuths - the only data we have are the distances x_i, $i = 1, \ldots N$. Gull's lighthouse problem [5] [35] requires us to find x_0 and y_0, the position of the lighthouse.

The relationship between θ_i, x_i, x_0 and y_0 is $y_0 \tan\theta_i = x_i - x_0$. We assume the lighthouse generates data according to the uniform distribution, i.e.,θ_i is uniform random. The question is how does knowledge of the distribution of θ_i provide us the distribution of x_i? The answer is provided by the theory of transformation of random variables as described in [65], [77] or [18]:

$$p(\theta_i|I) = p(x_i|x_0, y_0)\left|\frac{dx_i}{d\theta_i}\right| \tag{41}$$

Knowing that $\frac{d}{dx}\left[\tan^{-1} x\right] = \frac{1}{1+x^2}$ [8], we can write

$$\frac{d\theta_i}{dx_i} = \frac{y_0}{y_0^2 + (x_i - x_0)^2}.$$

Also, given the distribution of θ_i is uniform random, i.e., $p(\theta_i|I) = \frac{1}{\pi}$, we have

$$p(x_i|x_0, y_0) = \frac{1}{\pi}\frac{y_0}{y_0^2 + (x_i - x_0)^2}. \tag{42}$$

Thus we see that uniform random variables θ_i are transformed to *Cauchy* random variables x_i.

Equation (42) is the likelihood for a single point and with the assumption of independence, the combined likelihood is the product of the individual likelihoods:

$$p(\mathbf{x}|x_0, y_0) = \prod_{i=1}^{N} p(x_i|x_0, y_0) = \left(\frac{y_0}{\pi}\right)^N \prod_{i=1}^{N} \frac{1}{y_0^2 + (x_i - x_0)^2}. \tag{43}$$

[5] Steve Gull has been responsible for creating many elementary but fundamental problems in Bayesian statistics, thereby influencing a whole generation of Bayesians. For an amusing description of another "Gull problem," see page 48 of [56].

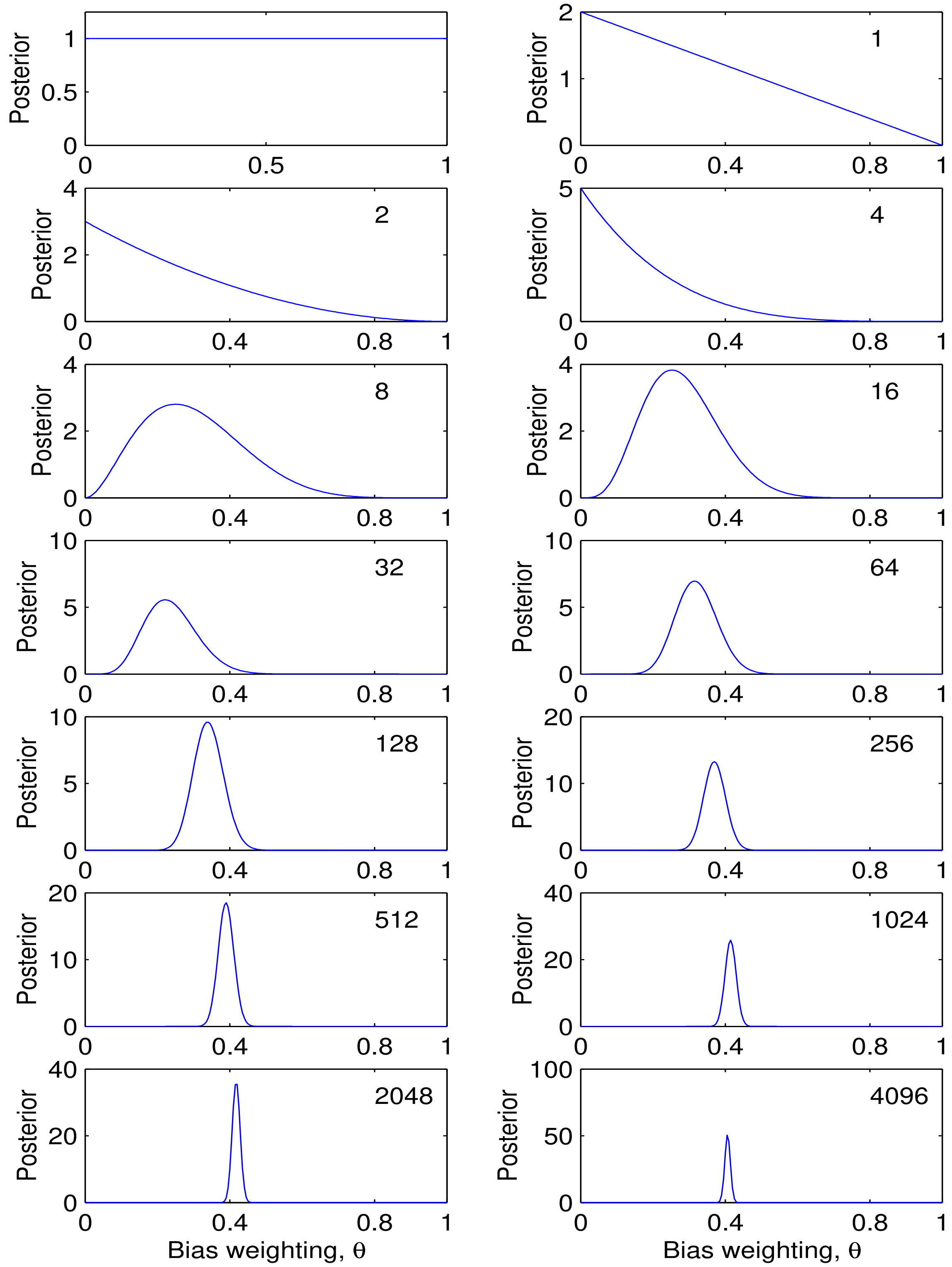

Figure 2: Estimating the bias in a two sided coin using Bayes theorem. The number of data is indicated on the upper right hand corner of each sub–plot except for the first one which represents the *prior*. Note that the area under the curve remains unity as we move from graph to graph. From 512 trials onward, the curves show higher and higher peaks with a corresponding narrowing about the abcissa, indicating greater and greater certainty about the bias in the coin.

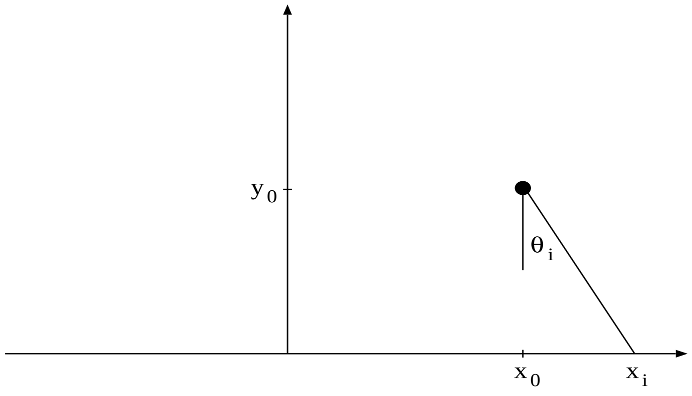

Figure 3: A lighthouse, at (x_0, y_0) m away from (an arbitrary) origin on a perfectly straight coast line, emits flashes of light at random times and hence at random azimuths. These flashes of light are recorded by a detector on the X axis and these are the only data we have. Gull's lighthouse problem requires us to find x_0 and y_0. This figure shows the geometry of the problem - for one particular random emission, the azimuth is θ_i and the corresponding intercept on the coast is x_i. For more mathematical details, see text.

The joint PDF is:

$$p(\mathbf{x}, x_0, y_0) = p(\mathbf{x}|x_0, y_0)p(x_0, y_0) = p(x_0, y_0|\mathbf{x})p(\mathbf{x}).$$

We assume that $p(x_0, y_0)$ has uniform density since it is a location parameter and $p(\mathbf{x})$ is also uniform. This leads to the final posterior distribution for x_0, y_0:

$$p(x_0, y_0|\mathbf{x}) = \prod_{i=1}^{N} \frac{1}{\pi} \frac{y_0}{y_0^2 + (x_i - x_0)^2}. \tag{44}$$

We wish to obtain the maximum of the posterior distribution [6] - this will give us the most likely location of the lighthouse in our coordinate system. The posterior distribution, equation 44 while analytical, is hard to handle analytically, so we resort to a simulation. We first generate the random distribution of points on the X axis where light from the lighthouse has been observed. The posterior is generated as in equation (44) by evaluating for $x \in [-44]$ and $y \in [04]$ for the observations $x_i, i = 1, \ldots, N$. Figure 4 shows the progress of the simulation for $N = 2, 5, 50$ and 400 points. As the number of points collected increases, the accuracy of the estimate increases. In this particular run we had the detector at $X = 1, Y = 1$ and the data in Figure 4 support that. Further discussion of a one–dimensional version of the problem is in [39] along a discussion of a maximum a posteriori estimator.

2.3 The Normal distribution

. . . The theory [of errors] is to be distinguished from the doctrine, the false doctrine, that generally, whenever there is a curve with single apex representing a group of statistics - one axis denoting size, the other frequency - that the curve must be of the "normal" species. The doctrine has been nicknamed "Quetelismus," on the ground that Quetelet exaggerated the prevalence of the normal law.

Francis Ysidro Edgeworth [25]

. . . In the middle 50s the writer heard an after–dinner speech by Professor Willy Feller, in which he roundly denounced the practice of using Gaussian probablity distributions for errors, on the

[6] It is also possible that the posterior has multiple maxima.

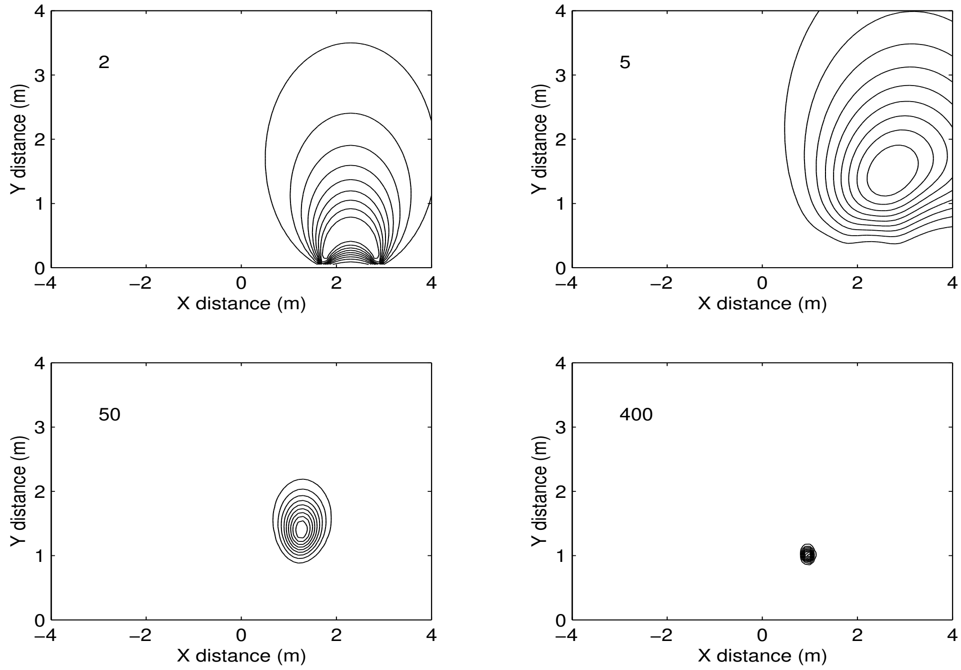

Figure 4: Gull's lighthouse problem II. As we note the position flashes of light on a detector along the X axis we plot the location of the lighthouse along the coast. The upper left panel shows the contours of the posterior distribution after only two flashes were observed. As we collect larger amounts of data (the quantity of data being collected is indicated on the upper left hand corner of each sub–graph) our accuracy in the estimate of (x_0, y_0) increases. After four–hundred light flashes, we are almost certain that the lighthouse is at location $X = 1, Y = 1$. For mathematical details, see text.

grounds that the frequency distributions for real errors are almost never Gaussian. Yet in spite of Feller's dissaproval, we continued to use them, and their ubiquitous success in parameter estimation continued.

E. T. Jaynes [46]

In this sub–section we present the general linear model - one that is used in an enormous number of applications. We usually write this as a matrix equation

$$\mathbf{d} = \mathbf{W}\mathbf{b} + \mathbf{e}. \tag{45}$$

In equation (45) $\mathbf{d}$ is an $N \times 1$ vector of observed data, i.e., there are N observations. $\mathbf{W}$ is an $N \times M$ matrix of basis functions. In the application described below, these are sin or cos but they could be virtually anything; $\mathbf{b}$ is and $M \times 1$ vector of linear coefficients, indeed in the harmonic model, they are amplitudes of the sines and cosines. Finally, $\mathbf{e}$ is an $N \times 1$ error vector that is Gaussian iid, i.e., $\mathbf{e} \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$ where σ is the standard deviation of the noise process.

Rearranging terms in (45) gives us $\mathbf{e} = \mathbf{d} - \mathbf{W}\mathbf{b}$. Now, by assumption,

$$p(\mathbf{e}|I) = \frac{1}{[2\pi\sigma^2]^{N/2}} \exp\left[-\frac{1}{2\sigma^2}\mathbf{e}^T\mathbf{e}\right].$$

Therefore, again using the rule for transformation of random variables [65], [77] or [18], we have

$$p(\mathbf{d}|\mathbf{W}, \mathbf{b}, I) = p(\mathbf{e}|I)\left|\frac{\partial \mathbf{e}}{\partial \mathbf{d}}\right|.$$

However, $\left|\frac{\partial \mathbf{e}}{\partial \mathbf{d}}\right| = 1$ and hence $p(\mathbf{d}|\mathbf{W}, \mathbf{b}, I) = p(\mathbf{e}|I)$ which is the likelihood of the data. Expanding, we get

$$p(\mathbf{d}|\{\omega\}, \mathbf{b}, \sigma, I) = \frac{1}{[2\pi\sigma^2]^{N/2}} \exp\left[-\frac{1}{2\sigma^2}(\mathbf{d} - \mathbf{W}\mathbf{b})^{\mathbf{T}}(\mathbf{d} - \mathbf{W}\mathbf{b})\right] \tag{46}$$

The use of $\{\omega\}$ and the disappearance of $\mathbf{W}$ in equation (46) requires some explanation. Until now we have spoken of $\mathbf{W}$ in most general terms but it is time to give a specific example. Let us consider one row of the matrix equation (45),

$$d_i = b_1 \cos(\omega_1 i) + b_2 \cos(\omega_2 i) + b_3 \cos(\omega_3 i) + b_4 \cos(\omega_4 i) + b_5 \cos(\omega_5 i) + e_i, \tag{47}$$

i.e., the $\mathbf{W}$ matrix is a function of $\omega_i, i = 1, \ldots 5$. In fact, we shall use the above basis functions later on in this section when we give a numerical example.

By Bayes' theorem

$$p(\mathbf{b}, \{\omega\}, \sigma|\mathbf{d}, I) \propto p(\mathbf{d}|\mathbf{b}, \{\omega\}, \sigma, I)\; p(\mathbf{b}, \{\omega\}, \sigma|I).$$

The above is the joint probability density function for $\mathbf{b}, \{\omega\}, \sigma$ conditioned on our prior information I and our data d. Assuming $\mathbf{b}$, $\{\omega\}$ and σ are independent we have

$$p(\mathbf{b}, \{\omega\}, \sigma|I) = p(\mathbf{b}|I)\; p(\{\omega\}|I)\; p(\sigma|I).$$

In the above equation, $p(\mathbf{b}|I)$, $p(\{\omega\}|I)$ and $p(\sigma|I)$ are prior probabilities and some judgement is required in selecting them. As it turns out, we have help from Harold Jeffreys [48] who states: "If the parameter may have any value in a finite range, or from $-\infty$ to $+\infty$, its prior probability should be taken as uniformly distributed. If it arises in such a way that it may conceivably have any value from 0 to ∞, the prior probability of its logarithm should be taken as uniformly distributed." For a detailed discussion on Jeffreys' rules see [97] while those who want a deeper understanding of why Jeffreys' rules should be true via group theory should refer to [46]. Applying this we have $p(\mathbf{b}|I) = \text{const.}$, $p(\{\omega\}|I) = \text{const.}$, and $p(\sigma|I) \propto 1/\sigma$.

Now suppose it is not $p(\mathbf{b}, \{\omega\}, \sigma|\mathbf{d}, I)$ that we are interested in but rather $p(\{\omega\}|\mathbf{d}, I)$. Given independence of $\mathbf{b}$, $\{\omega\}$ and σ, we can apply Bayes' theorem and repeated use of the marginalization rule to eliminate dependence on σ and $\mathbf{b}$. We this below. Using Bayes Theorem and equation (46) we get

$$p(\mathbf{b}, \{\omega\}, \sigma|\mathbf{d}, I) \propto \frac{1}{\sigma}\frac{1}{[2\pi\sigma^2]^{N/2}} \exp\left[-\frac{1}{2\sigma^2}(\mathbf{d}-\mathbf{Wb})^{\mathbf{T}}(\mathbf{d}-\mathbf{Wb})\right] \tag{48}$$

In many cases we are only interested in the parameters $\{\omega\}$ which form the elements of $\mathbf{W}$. In such a case we turn to the marginalization rule (**CITE SECTION HERE**), that is to say, we eliminate the nuisance variables ($\mathbf{b}$ and σ) step by step until we we are left with an expression for $p(\{\omega\}|\mathbf{d}, I)$.

To start with, we expand the quadratic in equation (48)

$$\mathbf{e}^T\mathbf{e} = \mathbf{d}^T\mathbf{d} - 2\mathbf{d}^T\mathbf{Wb} + \mathbf{b}^T\mathbf{W}^T\mathbf{Wb}.$$

Minimizing wrt $\mathbf{b}$ we have

$$\frac{\partial}{\partial\mathbf{b}} = 0 = -2\mathbf{W}^T\mathbf{d} + 2\mathbf{W}^T\mathbf{W}\hat{\mathbf{b}}.$$

Simplifying, we get $\hat{\mathbf{b}} = (\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T\mathbf{d}$, or the least–squares solution. If we denote $\mathbf{f}$ as the projection of the estimated parameters onto the basis space, i.e., $\mathbf{f} = \mathbf{W}\hat{\mathbf{b}}$, then we have [7]

$$(\mathbf{d}-\mathbf{Wb})^T(\mathbf{d}-\mathbf{Wb}) = (\mathbf{b}-\hat{\mathbf{b}})^T(\mathbf{W}^T\mathbf{W})(\mathbf{b}-\hat{\mathbf{b}}) + (\mathbf{d}^T\mathbf{d} - \mathbf{f}^T\mathbf{f}).$$

Substituting in equation (48) we get

$$p(\mathbf{b}, \{\omega\}, \sigma|\mathbf{d}, I) \propto \frac{1}{\sigma}\frac{1}{[2\pi\sigma^2]^{N/2}} \exp\left[-\frac{(\mathbf{b}-\hat{\mathbf{b}})^T(\mathbf{W}^T\mathbf{W})(\mathbf{b}-\hat{\mathbf{b}})}{2\sigma^2}\right] \exp\left[-\frac{\mathbf{d}^T\mathbf{d}-\mathbf{f}^T\mathbf{f}}{2\sigma^2}\right]. \tag{49}$$

Marginalizing gives us

$$p(\{\omega\}, \sigma|\mathbf{d}, I) \propto \frac{1}{\sigma}\frac{1}{[2\pi\sigma^2]^{N/2}} \exp\left[-\frac{\mathbf{d}^T\mathbf{d}-\mathbf{f}^T\mathbf{f}}{2\sigma^2}\right] \int_{R^M} \exp\left[-\frac{(\mathbf{b}-\hat{\mathbf{b}})^T(\mathbf{W}^T\mathbf{W})(\mathbf{b}-\hat{\mathbf{b}})}{2\sigma^2}\right] d\mathbf{b}. \tag{50}$$

The integral in equation (50) is almost in the standard form for an M–dimensional Gaussian [9], i.e.,

$$\int \exp(-\frac{1}{2}\mathbf{x}^T\mathbf{H}\mathbf{x})\, d\mathbf{x} = \frac{(2\pi)^{\frac{M}{2}}}{\sqrt{\det(\mathbf{H})}},$$

where $\mathbf{H}$ is an $M \times M$ symmetric matrix. Hence the integral in equation (50) reduces to $\frac{(2\pi\sigma^2)^{\frac{M}{2}}}{\sqrt{\det(\mathbf{W}^T\mathbf{W})}}$. Therefore we have

$$p(\{\omega\}, \sigma|\mathbf{d}, I) \propto \frac{1}{\sigma}\frac{(2\pi\sigma^2)^{\frac{M-N}{2}}}{\sqrt{\det(\mathbf{W}^T\mathbf{W})}} \exp\left[-\frac{\mathbf{d}^T\mathbf{d}-\mathbf{f}^T\mathbf{f}}{2\sigma^2}\right] = I_1. \tag{51}$$

In order to integrate out σ we need the following integral, i.e., $I_3 = \int_0^\infty \eta^{-(\nu+1)}\exp(-\beta/\eta)\, d\eta = \beta^{-\nu}\Gamma(\nu) = I_2$, by substituting $x = 1/\eta$. [8] In order to see how I_3 can be used we rearrange equation (51) by making the substitutions $\eta = 2\sigma^2$, $\beta = \mathbf{d}^T\mathbf{d} - \mathbf{f}^T\mathbf{f}$ and integrating over σ from 0 to ∞. After simplification we get:

$$I_4 = \frac{\pi^{\frac{M-N}{2}}}{2\sqrt{\det(\mathbf{W}^T\mathbf{W})}} \int_0^\infty \eta^{-\frac{N-M}{2}-1}\exp(-\frac{\beta}{\eta})\, d\eta.$$

[7] This relies on several simple matrix properties, in particular: $\mathbf{W}^T\mathbf{W}$ is symmetric, therefore $\left[(\mathbf{W}^T\mathbf{W})^{-1}\right]^T = \left[(\mathbf{W}^T\mathbf{W})^T\right]^{-1}$ and so $\left[(\mathbf{W}^T\mathbf{W})^T\right]^{-1}\mathbf{W}^T\mathbf{W} = \mathbf{I}$. These properties are discussed in [70].

[8] The Gamma function [11] is defined as $\int_0^\infty x^{\nu-1}\exp(x)^{-x}\, dx = \Gamma(\nu)$, for $\nu > 0$. This can be generalized: Start with $I_2 = \int_0^\infty x^{\nu-1}\exp(-\beta x)$. With the substitution $y = \beta x$ we can show that $I_2 = \beta^{-\nu}\Gamma(\nu)$. Similarly we can get $I_3 = \int_0^\infty \eta^{-(\nu+1)}\exp(-\beta/\eta)\, d\eta = \beta^{-\nu}\Gamma(\nu)$, by substituting $x = 1/\eta$.

Comparing I_3 with I_4 (let $\nu = \frac{N-M}{2}$) gives us

$$I_4 = \frac{\pi^{\frac{M-N}{2}}}{2\sqrt{\det(\mathbf{W}^T\mathbf{W})}}\Gamma(\frac{N-M}{2})\left[\mathbf{d}^T\mathbf{d} - \mathbf{f}^T\mathbf{f}\right]^{-\frac{N-M}{2}}$$

Hence, we have

$$p(\{\omega\}\,|\mathbf{d}, I) \propto \frac{\left[\mathbf{d}^T\mathbf{d} - \mathbf{f}^T\mathbf{f}\right]^{-\frac{N-M}{2}}}{\sqrt{\det(\mathbf{W}^T\mathbf{W})}}. \tag{52}$$

Thus by some lengthy but relatively straightforward manipulation, we have obtained a solution for the pdf $p(\{\omega\}\,|\mathbf{d}, I)$. Many readers will note that we do not have the true density but a quantity that is proportional to the density. For parameter estimation problems, this is sufficient since typically we are interested in those $\{\omega\}$'s that maximize the the posterior pdf, equation (52).

In order to reduce the mathematics to practice, let us study an example referred to in passing earlier in this section, equation (47),

$$d_i = b_1\cos(\omega_1 i) + b_2\cos(\omega_2 i) + b_3\cos(\omega_3 i) + b_4\cos(\omega_4 i) + b_5\cos(\omega_5 i) + e_i.$$

Let us choose $\{\omega\} = [\omega_1\ \omega_2\ \omega_3\ \omega_4\ \omega_5] = [0.1\ 0.15\ 0.3\ 0.31\ 1]$. This particular choice of $\{\omega\}$'s was taken from [13] because of the two closely spaced frequencies in noise - it being a well known fact that the Schuster periodogram does not resolve the spectrum well under such conditions. Nevertheless, the periodogram does contain information that is very relevant to the problem as may be seen in Figure (5). The top panel shows noisy data generated using equation (47) with $[\omega_1\ \omega_2\ \omega_3\ \omega_4\ \omega_5] = [0.1\ 0.15\ 0.3\ 0.31\ 1]$ along with normally distributed additive noise, $\mathcal{N}(0,1)$. Looking at the data, we can barely tell whether it has been generated using cosines, let alone that there are five frequencies in the data. The Schuster peroiodogram [44] is shown in the lower panel is certainly of more help since we can see at least four well defined frequences. However we are unable to tell whether the frequency around 0.31 is real or not.

For now let us assume that there are five frequencies - i.e., we were told by the sender of the signal that there would be five of them. Ascertaining whether the data supports four or five or even ten frequencies is a hard problem - but Bayesian methods are up to the task. This is left to our section entitled, "Model Selection." In order to determine the frequencies that lead to a maximum for equation (52) we use the Nelder–Mead simplex algorithm [72], [63] for derivative–free maxima finding. A well known flaw of most optimization algorithms is the fact that it only converges to a local maximum but this can be mitigated by good initial guesses. Luckily the Schuster periodogram provides that. By zooming into the local peaks we selected initial frequencies of [0.098 0.1472 0.294 0.3125 0.99]. Upon completion of the Nelder–Mead algorithm we obtained [0.1 0.15 0.3 0.3102 1.0] which is very close to the original frequencies. In fact one may see that we have gotten an order of magnitude better accuracy after maximization as compared to our initial guess from the Schuster Periodogram.

3 Model selection

Quando propositio verificatur pro rebus, si tres res vul duae sufficiunt ad veritatem illius propositionis, quarta res superfluit.

William of *Occam.* [57]

In this section we are going to study how Bayes Theorem can be used to decide whether a particular hypothesis (or theory) is true. In some ways, Bayes Theorem embodies a quantitative description of *Occam's Razor*, [9] which posits that a plurality should not be posited except of necessity. Einstein's statement, "everything should be as simple as possible, but not simpler," is another representation of the same.

[9] William of Ockham was born in 1285 in the town of Ockham, SW of London. He joined the oblates of St. Francis and was ordained a priest sometime in the early months of 1318. He studied at Oxford University eventually occupying the Franciscan

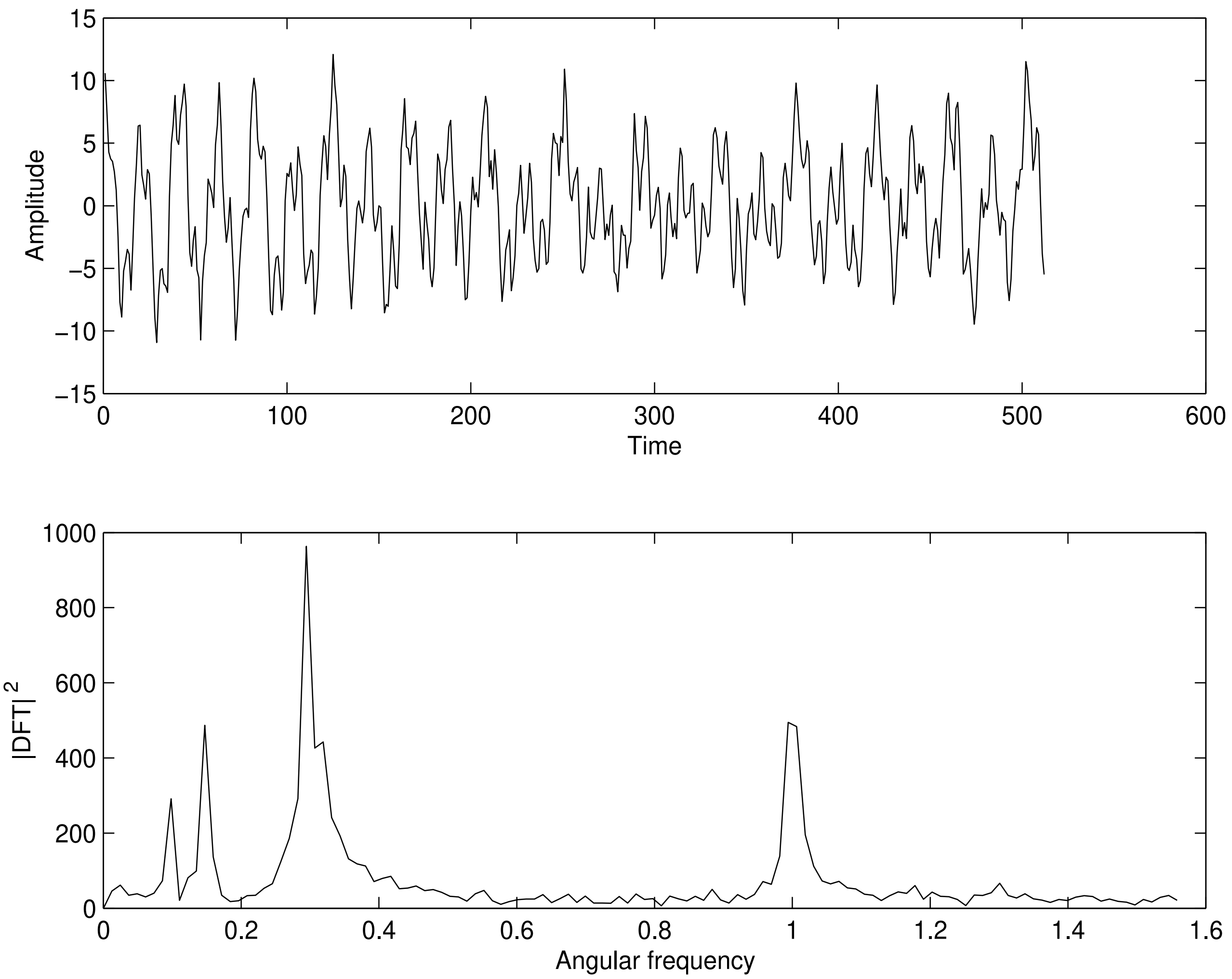

Figure 5: Noisy cosine data (top panel) and the Schuster periodogram (lower panel). The noisy data was generated using equation (47) with $[\omega_1\ \omega_2\ \omega_3\ \omega_4\ \omega_5] = [0.1\ 0.15\ 0.3\ 0.31\ 1]$ along with normally distributed additive noise, $\mathcal{N}(0, 1)$. The Schuster periodogram shows at least four frequencies with it being unclear whether the smaller peak near 0.3 is real.

$\log_{10} B$	B	Evidence for M_1
0 to $\frac{1}{2}$	1 to 3.2	Barely worth a mention
$\frac{1}{2}$ to 1	3.2 to 10	More likely
1 to 2	10 to 100	Strong
> 2	> 100	Impossible to ignore

Table 1: Deciding upon a model based on the Bayes Factor.

The methods to be presented were first studied by Harold Jeffreys who published his results in [47]. In the literature this is sometimes called hypothesis testing though the original phrase used by Jeffreys was, "tests of Significance," probably from reading R. A. Fisher (see [28] for example). This term has fallen out of use, with the term "model selection," being used instead. Text books that discuss model selection in some detail are [56], [46] and [77].

3.1 Essentials

Bayes theorem for a model is written as:

$$P(M|D,I) = \frac{P(D|M,I)\,P(M|I)}{P(D|I)}, \tag{53}$$

with $M \equiv$ Model, $D \equiv$ given data and $I \equiv$ background information. The data D could have come from one of a set of mutually exclusive models M_i, $i = 1, 2, \ldots, N$, $\sum_{i=1}^{N} P(M_i) = 1$, and we wish to see which model fits the data best. In the two model case:

$$P(M_i|D,I) = \frac{P(D|M_i,I)\,P(M_i|I)}{P(D|M_1,I)\,P(M_1|I) + P(D|M_2,I)\,P(M_2|I)},\ i = 1, 2. \tag{54}$$

The denominator is common to both models and is eliminated when we take the ratio, so that

$$\frac{P(M_1|D,I)}{P(M_2|D,I)} = \frac{P(D|M_1,I)\,P(M_1|I)}{P(D|M_2,I)\,P(M_2|I)}. \tag{55}$$

In words, equation (55) read from left to right is [49]

$$\text{posterior odds} = \text{Bayes Factor} \times \text{prior odds},$$

where odds $= p/(1-p)$ and p is a probability. In what follows, the Bayes Factor is written as BF. BF can be used to choose between models using a simple numerical scheme originally outlined in Appendix B of Jeffreys [48] (also in [49]) and reproduced in Table 1.

3.2 Tumbling Dice

Consider a dice that is rolled $n = 20,000$ times so that no face is systematically preferred over the other. At the conclusion of the experiment, the five or six face was observed $r = 7567$ times. Hence the one or two or three or four face showed up $s = 12373$ times. The astute reader will note that the problem though involving dice tumbling, is actually more like the original ball–and–table problem of Bayes. Based on the

Chair. He was accused of heresy and was summoned to Avignon, France by Pope John XXII. Ockham believed the same of the Pope and escaped to Italy. Eventually he came to Munich, spending the rest of his life there, dying in 1347. Other versions of Occam's Razor are, *"essentia non sunt multiplicanda, praeter necessitatem."* *Occam* is Latin for Ockham. Biographical information and latin quotes are from from [57].

data, two engineers have different ideas as to the models that generated the observations cited above. This situation is formalized as:

Let $M_1 \equiv$ "The data were generated by a binomial model, with unknown parameter θ. Similarly we can define, $M_2 \equiv$ "The data were generated by a fair dice - each face has the same chance of being observed", and finally, $D \equiv$ "the observed data, n, r and s". As always, I represents all the other information we have about the problem.

In equation (55), assume the prior odds is unity, i.e., we are unable to decide which of the models is more probable. So we are left with the task of calculating the quantities in the Bayes Factor. With the marginalization rule for any M_i we get

$$P(D|M_i, I) = \int P(D|\theta, M_i, I) P(\theta|M_i, I) d\theta. \tag{56}$$

For M_1, $P(D|\theta, M_1, I) = {}^nC_r \theta^r (1-\theta)^s$. Assuming a uniform prior on $0 \le \theta \le 1$ as in the Bayes–table example, the integral in equation (56) becomes:

$$P(D|M_1, I) = \int_0^1 {}^nC_r \theta^r (1-\theta)^s d\theta = {}^nC_r \frac{r!s!}{(n+1)!} \tag{57}$$

where, ${}^nC_r = \frac{n!}{r!(n-r)!}$. Equation (57) is the well known Beta integral. From our description of M_2 we can assume that the probability of getting a 5 or a 6 face up is $\theta_f = 1/3$. Using equation (56) with $M_i = M_1$ and $P(D|\theta, M_1, I) = \delta(\theta - \theta_f)$ gives us

$$P(D|M_s, I) = {}^nC_r \theta_f^r (1-\theta_f)^s. \tag{58}$$

Substituting equations (58) and (57) in (55) gives:

$$BF = \frac{P(M_2|D, I)}{P(M_1|D, I)} = \frac{(n+1)!}{r!s!} \theta_f^r (1-\theta_f)^s \tag{59}$$

In our dice problem $n = 20,000$ and r and s are also large, hence evaluating equation (59) is problematic. However, there is an approximation to the binomial, originally due to DeMoivre and Laplace [37] that simplifies calculation which we derive below.

For notational simplicity let $\theta_f = p$ and $1 - \theta_f = q$ and let us start with the Stirling approximation [37], i.e.,

$$r! \sim (2\pi r)^{1/2} r^r e^{-r}, \tag{60}$$

and substitute it into equation (59) to get

$$BF = (n+1) \left[\frac{n}{2\pi r(n-r)} \right]^{1/2} \left(\frac{np}{r} \right)^r \left(\frac{nq}{n-r} \right)^{n-r}. \tag{61}$$

Let $\delta = r - np$ and hence $n - r = nq - \delta$. Equation (61) is now modified to

$$BF = (n+1) \left[\frac{n}{2\pi(np+\delta)(nq-\delta)} \right]^{1/2} \frac{1}{\left(1+\frac{\delta}{np}\right)^{np+\delta} \left(1-\frac{\delta}{nq}\right)^{nq-\delta}}. \tag{62}$$

The Taylor series for $\log(1+x)$ is $x - x^2/2 + x^3/3 - x^4/4 \ldots$ for $-1 < x < 1$. Using this expansion the denominator of the last term in (62) can be written as

$$\log D = \left[\delta + \frac{1}{2}\frac{\delta^2}{np} - \frac{1}{6}\frac{\delta^3}{n^2p^2} \ldots \right] - \left[\delta - \frac{1}{2}\frac{\delta^2}{nq} + \frac{1}{6}\frac{\delta^3}{n^2q^2} \ldots \right]. \tag{63}$$

In order for the expansion to be valid we need ensure that $|\delta/np| < 1$ and $|\delta/nq| < 1$, which is easily verified for our given values of n, p, q and the calculated value of δ.

Keeping terms upto second order in δ reduces equation (63) to

$$\log D = \frac{\delta^2}{2n}\left(\frac{1}{p} + \frac{1}{q}\right) = \frac{\delta^2}{2npq}.$$

Hence,

$$D = \exp\left(\frac{\delta^2}{2npq}\right).$$

Modify the second term in (62) by dividing numerator and denominator by n^2 and assume $\delta/n \to 0$ as $n \to \infty$. Then

$$E = \left[\frac{1}{2\pi n(p+\delta/n)(q-\delta/n)}\right]^{\frac{1}{2}} \to \frac{1}{(2\pi npq)^{1/2}}.$$

Substituting D, E in equation (62) finally gives

$$BF \simeq \left[\frac{n}{2\pi pq}\right]^{1/2} \exp\left(-\frac{(r-np)^2}{2npq}\right). \tag{64}$$

This is much easier to use in calculations than equation (61).

Evaluating equation (64) with the given values of n, r, p and q gives us $BF = 3.562 \times 10^{-36}$. With such a small value of BF, evidence for M_2 is overwhelming, i.e., that the dice are not "fair." Evidently in making dice, the 1 and the 6 faces are opposite each other, as are 2 and the 5 and the 3 and the 4 faces. In their manufacture, more mass is scooped out of the 6 and 5 faces causing the center of gravity to shift toward the one 1 or the 2 face, making the 5 or the 6 face more likely to settle updward.

The data for this problem was taken from Jaynes [43] who himself obtained the data from Czuber [21] who in turn got the results of Rudolf Wolf, a Swiss astronomer. In the 1860's Wolf rolled a die 20,000 times in such a manner so as not to systematically favor one face over the other. Frequency data for each face and an analysis using the Maximum Entropy Method is in [43]. Dice tossing experiments were also conducted by W. Weldon as reported in [69] and [30] and similar biases toward the 5 and the 6 face were noted in [48].

4 Markov Chain Monte Carlo

5 Bayes Nets

6 Definitions

The **pattern space** is that domain defined by the discretization of sensor data observing the real world. A vector $\mathbf{X}$ in R dimensional pattern space is represented by $[x_1\ x_2\ x_3\ \cdots\ x_r\ \cdots\ x_R]^T$.

Here, x_r is a scalar representing a particular value associated with the r^{th} dimension. We can say that $\mathbf{X}$ is comprised of the scalar values descriptive of a set of R measurements which the designer has determined will define pattern space.

For example, consider a camera that can record information on 4 channels - red, green, blue and infrared. The pattern space is 4 dimensional with each value x_r for $r = 1, 2, 3, 4$ corresponding to a channel value or reading.

have an explicit representation of $\mathbf{\Phi}$ at all, but only K. The restrictions on what functions can qualify as kernel functions is discussed in Burges [14].

What is gained is that we have moved the data into a larger space where the training data may be spread further apart and a larger margin may be found for the optimal hyperplane. In the cases where we can explicitly find $\mathbf{\Phi}$, then we can use the inverse of $\mathbf{\Phi}$ to construct the non-linear separator in the original space. Clearly there is a lot of freedom in choosing the Kernel function and recent work has gone into the study of this idea both for SVM and for other problems [79].

With respect to the curse of dimensionality, we never explicitly work in the higher dimensional space, so we are never confronted with computing the large number of vector components in that space.

For the results presented below, we have used the inhomogeneous polynomial kernel function

$$K(\mathbf{x}, \mathbf{y}) = (\mathbf{x} \cdot \mathbf{y} + 1)^d, \tag{143}$$

with $d = 7$, though we found little difference in our results for $d = 2, \ldots, 6$. The choice of the inhomogeneous polynomial kernel is based on other workers success using this Kernel function in solving the handwritten digit problem [10].

In fact there are principled ways to choose among kernel functions and to choose the parameters of the kernel function. Vapnik [92] has pioneered a body of results from probability theory that provide a principled way to approach these questions in the context of the Support Vector Machine.

12.7 Multi-Class Classifiers

Two simple ways to generalize a binary classifier to a classifier for K classes are:

1. Train K binary classifiers, each one using training data from one of the K classes and training data from all the remaining $K - 1$ classes. Apply all K classifier to each vector of the test data, and select the label of the classifier with the largest margin, the value of the argument of the sgn function in Eq. (141).
2. Train $\left(\begin{array}{c} K \\ 2 \end{array} \right) = K(K-1)/2$ binary classifiers on all pairs of training data. Apply all $K(K-1)/2$ binary classifiers to each vector of the test data and for each outcome give one vote to the winning class. Select the label of the class with the most votes. For a tie, apply a tie breaking strategy.

We chose the second approach, and though it requires building more classifiers, it keeps the size of the training data smaller and is faster for training.

A A Life of Thomas Bayes

Thomas Bayes was born in 1701 (or 1702, the exact year is unknown) to Joshua and Anne Bayes (née Carpenter). Even his place of birth is unknown but it is surmised [23] that it occurred in Hertfordshire. Bayes Senior (his date of birth cannot be ascertained either) was a Presbyterian minister, being ordained on 22 June, 1694. In some sources Joshua is described as being a Calvinist and in others and Arminian. He died on 24^{th} April, 1746.

While history records some details of Joshua's early education, one can only conjecture as to that of Thomas's. There is small evidence to suggest that Bayes may have been a student of John Ward in Tenter

Alley (Moorfields, London) and even of John Eames at Coward's academy [64]. It is however certain [14] that Bayes was in attendance at Edinburgh University starting 1720. While a student, Bayes at the very least studied Logic and Metaphysics under Colin Drummond and took courses in Divinity for which he delivered exegeses on the Gospel of Matthew [22]. Given our interest in his mathematical and statistical writings we would like to know whether Bayes studied mathematics while at the University. Unfortunately, we have no direct knowledge of this but a letter from John Ward dated 10 May, 1720 commends him on his course of studies stating in part [6]: "In occupying yourself simultaneously with both mathematics and logic you will more clearly and easily notice what and how much each of these excellent instruments contributes to the directing of thought and sensation." One may then surmise that Bayes did indeed study mathematics and may well have done so under James Gregory who also sponsored his entrance to the University Library as stated in *Leges Bibliothecae Universitatis Edinensis* [22]. He did not graduate from Edinburgh University, but this was not considered a bad thing at that time. The last record of his at the university dates from 1722.

After leaving Edinburgh University and returning to London, there is scant record of Bayes life. There are records of his church related activities dating from 1728 and 1732 but nothing else. In 1734 [15] he moved to Turnbridge Wells (thirty–five miles from London) where he became a minister at the Mount Sion chapel. During his London period and his turn to Turnbridge Wells, Bayes researched and published two works: *Divine Benevolence Or, An ATTEMPT to prove that the PRINCIPAL END Of the Divine PROVIDENCE* and *GOVERNMENT IS THE Happiness of his Creatures (1731)*, [16] and *AN INTRODUCTION TO THE Doctrine of Fluxions, And DEFENCE of the MATHEMATICIANS AGAINST THE OBJECTIONS of the Author of the ANALYST, so far as they are designed to affect their general Methods of Reasoning (1736).* While the first is of a religious nature, the second is a defense of the Newtonian methods of doing calculus against an attack by George Berkeley [17]. It is believed that Bayes was elected a Fellow of the Royal Society in 1742 on the strength of his second publication - though there may have been other unpublished communications [6] that could have played a role as well.

Thomas Bayes suddenly died on 7 April, 1761, though he had been in ill health for at least ten years prior to that. He is interred in the Bayes–Cotton family vault, Bunhill Fields, London, [18] shown in Figure 16. A much published portrait of Bayes [19] is in Figure 15.

If Bayes had only published the two aforementioned documents, he would have gone down as an extremely minor figure in the annals of mathematics. Indeed, the posthumous publication of his essay on the doctrine of chances [20] in 1763 [5] wasn't widely cited for many years after his death. His work on inverse probability was independently re–discovered and presented in far greater generality and clarity – along with the first modern statement of what we now know as Bayes theorem [21] – by Laplace about ten years later [53]. However, this way of doing probability died (or at least was suppressed) for many years while orthodox methods [22] held sway. It is only since 1950 - more than two hundred years after Bayes work - that Bayesian

[14] This is due to serendipitous scholarship by Andrew Dale as recounted in the preface to [23]. Dale was visiting the University of Chicago Library and noticed a catalogue of manuscripts from the Edinburgh University Library that contained admission information on Thomas Bayes [6].

[15] There is some uncertainty about this date as discussed in [6], page 15.

[16] The unusual capitalization is directly from the text while the italicization is mine.

[17] He is the "Author of the ANALYST" referred to in Bayes publication. His attack on Newtonian calculus was published under the jaw–breaking title of: *The Analyst; or, a Discourse Addressed to an Infidel Mathematician. Wherein it is examined Whether the Object, Principles, and Inferences of the Modern Analysis are more Distinctly Conceived, or more Evidently Deduced, than Religious Mysteries and Points of Faith by the Author of the Minute Philosopher (1734).*

[18] Interestingly enough, Bayes grave is within five minutes walking distance of the offices of the Royal Statistical Society.

[19] There is doubt that this is actually a portrait of Bayes. Bellhouse [6] discusses how he arrived at this conclusion.

[20] The paper was and was communicated to the Royal Society by his friend Richard Price who is also interred in the Bunhill–Fields burial grounds along with notables William Blake, John Bunyan and Daniel Defoe. Richard Price is mentioned in Bayes will in the following sentence [23]:"Also I give and bequeath to John Hoyle late preacher at Newington and now I suppose at Norwich and Richard Price now I suppose Preacher at Newington Green two hundred pounds equally between them or the whole to the Survivor of them." While the paper was published in 1763, there is no accurate date as to when it was first written.

[21] Bayes writing would be archaic to the modern reader and his mathematical approach is geometric, i.e., he calculates areas rather than use the simpler integral notation. Despite this, to the dedicated statistician, there is great value to reading his essay - or at the very least, modern interpretations of the essay [82], [38].

[22] Since the subject of this paper is Bayesian statistics, we will not discuss orthodox methods. For an extensive discussion on

REV. T. BAYES

Figure 15: Bayes or not Bayes?

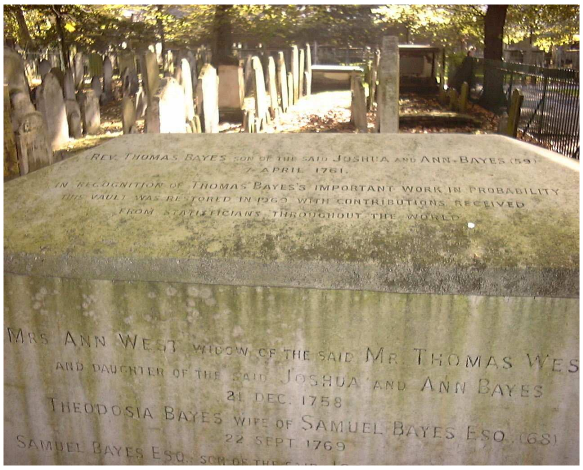

Figure 16: Bayes–Cotton grave Bunhill–Fields, London. The inscription reads in part: "In recognition of Thomas Bayes's important work in probability this vault was restored in 1969 with contributions received from statisticians throughout the world."

methods have been on the rise and that the name [82] "Bayes's Theorem is surely among the most frequently encountered eponyms in modern statistical literature."

B A brief analysis of Bayes Essay

> *Though subjected to Talmudic scrutiny in recent years, Bayes's paper had little impact at the time.*
>
> David Howie [41].

Bayes paper, "*An Essay towards solving a Problem in the Doctrine of Chances,"* was communicated to the Royal Society by his friend Richard Price. It is not clear how Bayes technical writings arrived in his possession, but it is evident that the introductory material by Bayes was either edited or completely rewritten by Price as stated in his opening letter.

The paper consists of the letter by Price, two sections, a scholium and an appendix. Every section needs to be read carefully because of the archaic writing style as well as the use of the Newtonian method of writing integrals as areas. Nevertheless, the problem that Bayes is trying to solve is succinctly stated:

> *Given* the number of times on which an unknown event has happende and failed: *Required* the chance that the probability of it happening in a single trial lies somewhere between any two degrees of probability that can be named.

The first section deals with Bayes' concept of probability of which the following is the most obscure:

> The *probability of any event* is the ratio between the value at which an expectation depending on the happening of the event ought to be computed, and the value of the thing expected upon it's happening.

Using Jeffreys [48] (page 32) notation, we have

$$P(A|I) = \frac{E(A,N|I)}{N}.$$

The probability is thus the ratio of how much you are willing to bet, $E(A,N|I)$, in order to win a prize of value N, where, A is the event in question [36]. Throughout his paper, Bayes does not give any indication of how $E(A,N|I)$, "ought to be computed." Also note that I have used the modern Bayesian habit of always conditioning on the background information I. Clearly, this definition permits both a frequency view of probability (as in games of chance) as well as the probability as degree–of–belief where frequencies might not be available or simply not apply.

There are a total of seven propositions in Section I of which the third and the fifth are critical. Bayes assumes two "subsequent events," say A and B, [23] and proves the following two statements:

$$P(B|A) = \tfrac{P(A\cap B)}{P(A)}, \quad 3^{rd} \text{ proposition,}$$

and

$$P(A|B) = \tfrac{P(A\cap B)}{P(B)}, \quad 5^{th} \text{ proposition.} \tag{144}$$

the pros and cons of each along with illustrative examples, see [46], [41].

[23] In Bayes third poposition it is not clear whether A and B are logically connected or temporally related, but in the fourth proposition, it is evident he is referring to a temporal connection. For the purposes of our discussion we assume that the third and the fifth proposition can be obtained by a simple exchange of A and B. For an extensive discussion of these two propositions, see [75].

The second section - the heart of the paper, contains two propositions (eight and nine) preceded by a thought experiment that Hacking [36], calls a “tandem set–up.” The two–stage experiment is ingenious since it posited a continuous parameter space - most previous probability models up to that point in time used discrete spaces.[24] Bayes considered a level square table $CDAB$ shown in Figure 17 and a ball W “be thrown on it.” He postulated that the ball can fall anywhere uniformly within the table and upon coming to rest a line os is drawn through the ball, parallel to CB.

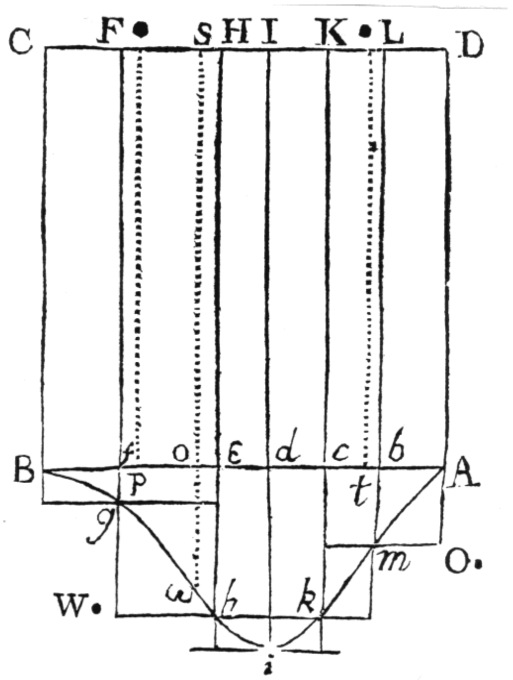

Figure 17: The square table CDAB used by Bayes in his thought experiment. The lower curve with base located on the line BA is the posterior density $f(\theta|X = p)$ which we discuss in the text.
. Note that we have used Bayes original figure which is freely available on the web.

In the second stage, the ball O is thrown n times and the number of times it comes to rest between os and AD is denoted by the random variable X. Bayes calls this event M and says that “the probability of the event M in a single trial is the ratio of Ao to AB. If we denote that the position of the ball W be denoted by θ, the problem then reduces to finding the position of θ accurately after the second–stage of the experiment. Using proposition three, and assuming event M occurs p times (and its complement occurs $q = n - p$ times), Bayes derives proposition 8:

$$P(b < \theta < f \cap X = p) = \int_b^f {}^nC_r\, \theta^p(1-\theta)^q d\theta.$$

The marginal distribution (integrating out the dependence on θ) is

$$P(X = p) = \int_0^1 {}^nC_r\, \theta^p(1-\theta)^q d\theta = \frac{1}{n+1}.$$

Finally, in proposition nine, Bayes uses proposition 5 (equation 144) to derive the “probability of its happening in a single trial lies somewhere between any two degrees of probability that can be named:”

$$P(b < \theta < f|X = p) = \frac{\int_b^f {}^nC_r\, \theta^p(1-\theta)^q d\theta}{\int_0^1 {}^nC_r\, \theta^p(1-\theta)^q d\theta}. \tag{145}$$

[24]Notable exceptions to this are discussed in [38].

Equation (145) is now seen to be a special case of equation (40) and what we derived by direct application of Bayes Theorem was derived by first principles by the Reverend.

Bayes extends his reasoning in his scholium to cases beyond his square table example. There is no need to elaborate upon this since we have directly applied this in our coin–tossing example in section 2.1. For further discussion on section two of Bayes paper and his scholium see [36], [82], [61], [62], [26].

Need to get Fishers book on statistics. Also include a criticism of Bayes paper. Get Shenyin's paper on Poincare where it is mentioned that Cournot was the first to use the term "Bayes Formula."

References

[1] J. Aczél. *A Short Course on Functional Equations*. D. Reidel,, Dordrecht., 1987.

[2] M.A. Aizerman, E.M. Braverman, and L. I. Rozoner. Theoretical foundations of the potential function method in pattern recognition. *Automation and Remote Control*, 25:821–837, 1964.

[3] Harry C. Andrews. *Introduction To Mathematical Techniques In Pattern Recognition*. Wiley-Interscience, New York, 1972.

[4] N. S. Bakhvalov. On approximate calculation of integrals. *Vestnik MGU, Ser. Mat. Mekh. Astron. Fiz. Khim.*,, 4:3–18, 1959. (In Russian).

[5] T. Bayes. An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society,*, 53:370–418, 1763.

[6] D. R. Bellhouse. The reverend thomas bayes frs: a biography to celebrate the tercentenary of his birth. *Statistical Science*, 2004. To be published.

[7] R. E. Bellman. *Dynamic Programming*. Princeton University Press, Princeton, N.J., 1957.

[8] J. W. Bishir and D. W. Drewes. *Mathematics in the Behavioral and Social Sciences*. Harcourt, Brace and World, Inc., New York, 1970.

[9] C. M. Bishop. *Neural Networks for Pattern Recognition*. Clarendon Press, Oxford, 1995.

[10] B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In *Fifth Annual Workshop on Computational Learning Theory*, pages 144–152. ACM, June 1992.

[11] F. Bowman. *Introduction to Bessel Functions*. Dover Publications, New York, 1958.

[12] S. Boyd and L. Vandenberghe. Convex optimization, 1997. Course notes from Stanford University for EE364, Introduction to Convex Optimization with Engineering Applications. Available at `http://www.stanford.edu/class/ee364/`.

[13] G. L. Bretthorst. *Bayesian Spectrum Analysis and Parameter Estimation*. Springer–Verlag, Berlin., 1988.

[14] C. J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, 1998. Also available at `http://svm.research.bell-labs.com/SVMdoc.html`.

[15] C. Cortes. *Prediction of Generalization Ability in Learning Machines*. PhD thesis, University of Rochester, 1995.

[16] C. Cortes and V. N. Vapnik. Support vector networks. *Machine Learning*, 20:1–25, 1995.

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

[18] G. Cowan. *Statistical Data Analysis.* Clarendon Press, Oxford, Oxford., 1998.

[19] R. T. Cox. Probability, frequency and reasonable expectation. *American Journal of Physics,*, 14:1–13, 1946.

[20] R. T. Cox. *The Algebra of Probable Inference.* John Hopkins Press, Baltimore, 1961.

[21] E. Czuber. *Wahrscheinlichkeitsrechnung.* Teubner, Berlin, 1908.

[22] A. I. Dale. *A History of Inverse Probability From Thomas Bayes to Karl Pearson.* Springer, New York., second edition, 1999.

[23] A. I. Dale. *Most Honourable Remembrance: The Life and Work of Thomas Bayes.* Springer, New York., 2003.

[24] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm (with discussion). *J. Roy. Statist. Soc. Series B*, 39:1–38, 1977.

[25] F. Y. Edgeworth. The philosophy of chance. *Mind*, 31:257–283, 1922.

[26] A. W. F. Edwards. *Likelihood.* Johns Hopkins University Press, Baltimore., second edition, 1992.

[27] S. Epp. *Discrete Mathematics with Applications.* Brooks/Cold Publishing Company, Pacific Grove., 1995.

[28] R. A. Fisher. *The Design of Experiments.* Oliver and Boyd, Edinburgh, 1995.

[29] R. Fletcher. *Practical methods of optimization (2nd edition).* J. Wiley, 1987.

[30] T. C. Fry. *Probability and its Engineering uses.* D. Van Nostrand Company, 1928.

[31] A. J. M. Garrett. Whence the laws of probability? In G. J. Erickson, C. R. Smith, and J. T. Rychert, editors, *Maximum Entropy and Bayesian Methods: Proceedings of the 17th International Workshop on Maximum Entropy and Bayesian Methods of Statistical Analysis,*. Kluwer Academic, 1998.

[32] A. Gelman and Nolan D. You can load a die, but you can't bias a coin. *The American Statistician,*, 56:308–311, 2002.

[33] R. L. Graham, D. E. Knuth, and O. Patashnik. *Concrete mathematics.* Addison–Wesley Publishing Company, Reading, Massachusetts., 1989.

[34] L. Greengard and J. Strain. The fast gauss transform. *SIAM J. Sci. Stat. Comput.*, 12(1):79–94, 1991.

[35] S. F. Gull. Bayesian inductive inference and maximum entropy. In G. J. Erickson and C. R. Smith, editors, *Maximum–Entropy and Bayesian Methods in Science and Engineering (Vol. 1)*, pages 53–74. Kluwer Academic, 1988.

[36] I. Hacking. *Logic of Statistical Inference.* Cambridge University Press, Cambridge, U.K., 1965.

[37] A. Hald. *A History of Probability And Statistics And Their Applications before 1750.* John Wiley and Sons Inc., New York, 1990.

[38] A. Hald. *A History of Mathematical Statistics from 1750 to 1930.* John Wiley and Sons Inc., New York, 1998.

[39] K. M. Hanson and D. R. Wolf. Estimators for the cauchy distribution. In G. R. Heidbreder, editor, *Maximum Entropy and Bayesian Methods*, pages 255–263. Kluwer Academic, 1996.

[40] R. V. Hogg and E. A. Tanis. *Probability and Statistical Inference.* Macmillan, New York., 1988.

[41] D. Howie. *Interpreting Probability.* Cambridge University Press, Cambridge., 2002.

[42] A. K. Jain and B. Chandraseharan. Dimensionality and sample size considerations in pattern recognition practice. In P. R. Krishnaiah and L. N. Kanal, editors, *Handbook of Statistics.* North–Holland, 1982.

[43] E. T. Jaynes. Concentration of distributions at entropy maxima. In R. D. Rosenkrantz, editor, *Papers on Probability, Statistics and Statistical Physics.* D. Reidel Publishing Co., 1983.

[44] E. T. Jaynes. Bayesian chirp and spectral analysis. In G. J. Erickson and C. R. Smith, editors, *Maximum Entropy and Bayesian Spectral Analysis and Estimation Problems.* D. Reidel, Dordrecht–Holland, 1987.

[45] E. T. Jaynes. How does the brain do plausible reasoning. In G. J. Erickson and C. R. Smith, editors, *Maximum Entropy and Bayesian Methods in Science and Engineering.* Kluwer Academic Publishers, 1988.

[46] E. T. Jaynes. *Probability Theory – The Logic of Science.* Cambridge, London, 2003.

[47] H. Jeffreys. Some tests of Significance, Treated by the Theory of Probability. *Proc. Cambridge Phil. Soc.*, 31:203–222, 1935.

[48] H. Jeffreys. *Theory of Probability.* Clarendon Press, Oxford, U.K., second edition, 1961.

[49] R. E. Kass and A. E. Raftery. Bayes Factors. *J. Am. Stat. Assoc.*, 90:773–793, 1995.

[50] T. Kohonen. *Self–organization and associative memory.* Springer–Verlag, Berlin, 1988.

[51] W. L. G Koontz and K. Fukunaga. Asymptotic analysis of a nonparametric estimate of a multivariate density function. *IEEE PAMI*, 21:967–974, 1972.

[52] K. Lange. *Numerical Analysis for Statisticians.* Springer, New York., 1999.

[53] P. S. Laplace. Mémoire sur la probabilité des causes par les évènements. *Mèm. Acad. R. Sci. Paris (Savants Étranger)*, 6:621–656, 1774. This was translated by Stigler in *Statistical Science*, Vol. 1, pp. 359–378, 1986.

[54] J. C. Lemm. *Bayesian Field Theory.* Johns Hopkins Press, Baltimore., 2003.

[55] T. Leonard and J. S. J Hsu. *Bayesian Methods: An Analysis for Statisticians and Interdisciplinary Researchers.* Cambridge University Press, Cambridge, Bonnie England., 1999.

[56] D. J. C. MacKay. *Information Theory, Inference and Learning Algorithms.* Cambridge, London, 2003.

[57] A. Maurer. *The Philosophy of William of Ockham in the Light of Its Principles.* PIMS, Toronto, 1999.

[58] C. Maxfield. *Bebop to the Boolean Boogie.* HighText Publications Inc., Solana Beach, California., 1995.

[59] G. J. McLachlan and T. Krishnan. *The EM Algorithm and Extensions.* John Wiley and Sons, Inc., New York., 1997.

[60] X. L. Meng and S. Pedlow. Em: A bibliographic review with missing articles. *Proc. Statist. Comput. Sect. Am. Statist. Ass.*, 39:24–27, 1992.

[61] E. C. Molina. Bayes' theorem: An expository presentation. *Ann. Math. Statistcs*, 2:23–37., 1931.

[62] F. H. Murray. Note on a scholium of bayes. *Am. Mathematical Society Bulletin*, 36:129–132., 1930.

[63] J . A. Nelder and R. Mead. Simplex method for function minimization. *The Computer Journal*, 7(4):308–313., 1965.

[64] M. E. Ogborn. *Equitable assurances. The story of life assurance in the experience of the Equitable Life Assurance Society.* George Allen and Unwin, London., 1962.

[65] A. Papoulis. *Probability, Random Variables and Stochastic Processes*. McGraw–Hill, Inc., New York., 1991.

[66] J. B. Paris. *The Uncertain Reasoner's Companion*. Cambridge University Press,, Cambridge., 1994.

[67] E. Parzen. On estimation of a probability density function and mode. *Annals of Mathematical Statistics*, 33:308–319, 1962.

[68] E. Pearson and E. Johnson. *Tables of the Incomplete Beta Function*. Cambridge University Press, Cambridge., 1969.

[69] K. Pearson. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Phil. Mag.*, 50(V):157–175., 1900.

[70] A. J. Pettofrezzo. *Matrices and Transformations*. Dover, New York., 1978.

[71] S. J. Press. *Subjective and Objective Bayesian Statistics: Principles, Models and Applications*. John Wiley and Sons, New Jersey, 2003.

[72] W. H. et al. Press. *Numerical Recipes in C*. Cambridge University Press, Cambridge., 2^{nd} edition, 1992.

[73] R. A. Redner and H. F. Walker. Mixture densities, maximum likelihood and the em algorthm. *SIAM Review,*, 26:195–239, 1984.

[74] D. W. Scott. *Multivariate Density Estimation: Theory, Practise and Visualization.* John Wiley and Sons, Inc., New York., 1992.

[75] G. Shafer. Bayes's two arguments for the rule of conditioning. *The Annals of Statistics,*, 10:1075–1089, 1982.

[76] B. W. Silverman. *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London, 1986.

[77] D. S. Sivia. *Data Analysis: A Bayesian Tutorial*. Clarendon Press, Oxford, England., 1996.

[78] C. R. Smith and G. Erickson. From rationality and consistency to bayesian probability. In J. Skilling, editor, *Maximum Entropy and Bayesian Methods*. Kluwer Academic Publishers, 1989.

[79] A. J. Smola, B. Schölkopf, and R. J. Müller. The connection between regularization operators and support vector kernels. *Neural Networks*, to appear, 1998.

[80] Donald Specht. Generation of polynomial discriminant functions for pattern recognition. *IEEE Transactions on Electronic Computers*, EC-16:309–319, 1967.

[81] Donald Specht. Probabilistic neural networks. *Neural Networks*, 3:109–118, 1990.

[82] S. Stigler. Thomas bayes's bayesian inference. *J. R. Statist. Soc. A*, 145:250–258, 1982.

[83] S. M. Stigler. *The History of Statistics: The Measurement of Uncertainty before 1900.* The Belknap Press of Harvard University Press, Cambridge, Massachusetts., 1986.

[84] S. M. Stigler. *Statistics on the table: The history of statistical concepts and methods*. Harvard University Press, Cambridge, Massachusetts., 1999.

[85] R. A. Tapia and J. R. Thompson. *Nonparametric Probability Density Estimation*. Johns Hopkins University Press, Baltimore, 1978.

[86] M. E. Tarter and R. A. Kronmal. An introduction to the implementation and theory of nonparametric density estimation. *The American Statistician*, 30:105–112, 1976.

[87] J. F. Traub and A. G. Werschulz. *Complexity and Information*. Cambridge University Press, Cambridge, U.K., 1998.

[88] M. Tribus. *Rational Descriptions, Decisions and Designs*. Pergammon, New York., 1969.

[89] M. Tribus. An appreciation of richard threlkeld cox. In R. L. Fry, editor, *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*. American Institute of Physics, 2002.

[90] K. S. Van Horn. Constructing a logic of plausible inference: A guide to cox's theorem. *International Journal of Approximate Reasoning,*, 34:3–24, 2003.

[91] V. N. Vapnik. *Estimation of dependencies based on empirical data*. Springer, 1982.

[92] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995.

[93] V. N. Vapnik. *Statistical Learning Theory*. John-Wiley and Sons, Inc., 1998.

[94] V. N. Vapnik and A. Ja. Chervonenkis. *Theory of pattern Recognition*. Nauka, 1974. In Rusian.

[95] Philip D. Wasserman. *Advanced methods in neural computing*. Van Nostrand Reinhold, New York, 1993.

[96] C. F. Jeff Wu. On the convergence properties of the em algorithm. *The Annals of Statistics*, 11:95–103, 1983.

[97] A. Zellner. *An Introduction to Bayesian Inference in Econometrics*. John Wiley, New York, 1971.