

# A Probability Primer: Uncertainties & Applications

Peter Grindrod CBE CMath

June 1, 2009

## 1 Uncertainty and Making Decisions

In many situations, whether in business, society, the environment, the clinic, or the laboratory, we are able to monitor a wider and wider variety of performance parameters and events. For effective knowledge distillation, interpretation, prediction and proactive intervention we need to develop some methods and algorithms. When we look at a complex or complicated system and make observations, what do we learn? What do the observations tell us about the object of our attention?

The pace of change of many technologies means that concepts and methods should ideally be adaptive new inputs, be self-learning or self-tuning; and hence may be applied even as both the underlying nature of the activity and the process of observation evolve. Hence analysts must deal with evolving uncertainties, and express their estimation of these, and incorporate reasoning to make their understanding clear: they need methods that are genuinely much smarter.

By “smarter” we do not mean an ability for making algorithms (or software) easy to set up or to integrate across data sources (that is called deployability), or the ability to provide and transmit content and support for such alerts. All this is just efficient generation and use of “intelligence”– used in the sense of “military intelligence” (the well known oxymoron!). What we mean by “smarter” is the sense that the algorithms can reason in a way that is consistent with the way that we ourselves think; that the algorithms are logical, and extend logic to common sense inferential rules; that the algorithms are used to consistently and continuously update the current view as to what, is or is not, certain.

Moreover a one time decision (to change, to invest, to treat, to investigate, or terminate) may need to be taken for management reasons at some particular moment in time. In this case best estimates of an evolving situation are required (what will happen under alternative decision scenarios? what is happening right now?). There is no time for the analysts to say “there isn’t enough data” or “we don’t know whether something aberrant can happen - no danger thresholds have yet been exceeded”. Instead decision making

requires the support of our best current recommendations - even if there is little data. Designing algorithms to give good decision support requires us to draw the best possible inferences we can based on everything we know to date (our prior knowledge), and our current incoming information (observations, data) about events.

What may be surprising is that, properly viewed, probability theory is really is just an extension of logic. It is not simply about making estimates of chance events, but it is far more powerful. Rather than asking “*we know how the system works and what its uncertainties are: what is the chance of it producing an output like X?*”; we want to ask “*having observed an output like X, what does that tell us further about the how system works, and what its uncertainties are?*”.

A moment’s thought and you will see that almost all plausible reasoning – from common sense, to deduction – require us to grapple with the second type of question.

In business or in science this is ever more true, as we are able capture or monitor more things. We observe events and results, and then we ask how does that alter our view about what is happening?

A good introduction to plausible reasoning is a good grounding for any decision making. Careful use of language, inferences, and knowledge of uncertainties flushes out what we do not know (where our uncertainties reside and derive from); and points us to the information that we need to make better justified decisions. The literature abounds with apparent paradoxes – these usually come from poor application of just a few simple principles (legal, medical, engineering, commercial...) – or from using the wrong conditional models to update our world view.

Thus an understanding of the material presented here represents a skill for life.

## 2 Sources of Uncertainty

Uncertainties arise in different ways.

There is the uncertainty over the truth of many propositions about the underlying *system* that we are addressing, and over the events that we are observing. The car is safe. The witness is lying. The economy will improve. The climate is warming. England will win the world cup. What do our observations tell us about the system, and the truth of these statements?

Observations of any related events may change the uncertainty we have in asserting these propositions. To see this we need to *model* the consequences of the proposition, and alternative propositions, and see whether the observations are more consistent with one hypothesis (that a particular proposition is true) or another.

Modellers, and especially mathematical modellers, are often uncertain as to which mod-

els they may apply to a given situation though. They must make some decisions, some assumptions and some omissions, based on their experience and sometimes on pragmatics. They must choose a class of model. How can modellers compare one model against another? This is called *conceptual model uncertainty*. Every assumption modellers make in setting out a model introduces uncertainty and models which share assumptions are not independent: if a common underlying assumption becomes invalid then so are the models.

When we observe events that can or cannot be represented well by the behaviour of solutions to a model we become more or less convinced of its usefulness. Our uncertainty in the conclusions or predictions from the outputs decreases or increases accordingly.

Given a particular model it may well have a parameter or two which we do not know, or that we are uncertain about (yet we have some experience or ideas about). This is called *parameter uncertainty*. We need to use evidence or observation to make better estimates for the possible whereabouts of parameter values.

Our models may well contain variable or stochastic, space or time dependent, parameters or processes. This is called *variability*. You could think of these quantities as sophisticated parameters if you wish. These things need to be modelled too. This in itself introduces submodels, with submodel conceptual and parameter uncertainties. Variability (input uncertainty, just like parameter uncertainty) may add to total uncertainty over outputs, even given highly certain inputs.

So even with our prior assumptions and beliefs, and our models, and our methods of making them operational (solution methods, numerical calculations), we may well want to actually find some things out. These may be forecasts, predictions, or may be decisions based on our estimation of uncertain model outputs. The more that we know, or observe, the more we hope to evolve our representation of the uncertainties concerning such outputs.

Sometimes we have other challenges. For example, in watching an animal, a person, or a machine, and trying to understand the nature of, and constraints upon, its ability to reason with uncertainty and respond to input (perceived observations).

We might be building or hypothesising systems that could reason with information or possible facts. These are models where we have to represent a process of reasoning, where the currency itself - the state variable - is uncertainty and is represented over "possibility space". To build such systems that reason in a common sense way, that reason consistently, that can be both very uncertain or very certain about different things, is a challenge.

It is curious how ill suited the normal methods of representing and transmitting information in digital systems seems to be in this task. We all possess a brain (1kg or so of wetware) which has evolved to achieve such goals, at low cost in real time, with flexibility and adaptability. How is such functional response to be understood, if not as an organ adapted to the task of plausible reasoning, successfully or sometimes otherwise,

under uncertainty? Computers and brains are functionally “dimensional opposites”. We can solve PDEs on computers: but we cannot solve them effortlessly within our brains. Without effort our brains can recognize and discriminate distinct sensory inputs, but we find it hard to make our computers do well at such tasks without them performing feats of memory and calculation that we know our brains simply cannot do.

To face any or all of these challenges we should at first be very clear about the mathematics of uncertainty and reason.

### 3 Probability as an extension of logic

What a strange thing probability is. This point was almost entirely missed at school, where we were all taught an experimental view, with definitions derived from expectations: the probability of picking a heart from a pack of cards – try it out lots of times! We were encouraged to imagine doing the experiment over and over again – then the probability was simply the number of successful events divided by the number of attempts. But this approach is only ever valid in the limit of an infinite number of experiments: we haven't the time nor the imagination in complex cases – and what about situations where we can't actually make or imagine repeated experiments because the question in mind is too one-off, and/or we have only a very small amount of evidence?

In such cases “probability” still means something to us though – it can obviously represent our uncertainty surrounding the truth of a given proposition. A proposition is just a statement about some event which may happen, or has happened, or a statement about certain circumstances being correct:

The card I select will be a heart”,  
”It is raining”, ” I am a liar”,  
”The next president of the US will be a woman”,  
”Ghosts exist”, or  
”Mr X is a murderer” .

Look at these examples: in general the more interesting, controversial or salacious, a proposition is, the harder it is to imagine any thought experiments that could determine an agreeable absolute probability, which will satisfy all onlookers. The probability that such a proposition is true is entirely subjective - I may know more about the subject or have more information than you - so our opinions, and our assessments of the residual uncertainties, are personal to us. Subjective probability is really the currency of the science of logical reasoning. Even though our estimates for the probabilities may differ between us in absolute values, whenever any new evidence arrives, both of our subjective probabilities should be updated in a consistent manner. That is, if we shared the same starting point – our prior estimates are the same – then the new information and evidence should alter our estimates in the same way, so that we arrive at the same “posterior” estimate of uncertainty after we have accounted for the new evidence. This consistent updating process is called Bayesian Updating, named after the Reverend Thomas Bayes.

## 4 A crash course in plausible reasoning

So let's start probability theory all over again. Let us leave aside what we have learned up to now. Here is a crash course in probability theory for plausible reasoning. This is all we ever need(!) and we will hold to these ten central points.

**Resources:** three excellent and invaluable books to dip into are [1, 2, 3].

1. **Plausibility.** For any statement or proposition we refer to the “plausibility” of the statement meaning our belief that the statement is true. Sometimes we will have alternative, mutually exclusive, statements/propositions, usually called “hypotheses”, where we are certain that one of them is true, and yet we are uncertain as to which one. They may have different levels of plausibility associated with each of them.
2. **A Conditional and Subjective Quality.** The plausibility of a proposition, meaning an estimation of the truth of statement, is always conditional – it is dependent on every other piece of information we have that we have taken in to account. Necessarily then, it is subjective too - since my plausibilities are conditional on my knowledge and yours are conditional on yours. In some fairly simple and therefore rather special circumstance we can agree what the plausibilities are if we can agree what other information (called assumptions) we will both consider.
3. **Conditional Notation.** In order to stress the conditional nature of plausibility we write  $A|X$  to represent the plausibility of the proposition that  $A$  is true given all background knowledge  $X$  (which stands for everything that is known when we determine the plausibility that  $A$  is true). Now suppose we gain some new evidence  $E$ , that was not in  $X$ . Then we will write  $A|E, X$  to denote the new plausibility that  $A$  is true given  $X$  **and** given  $E$ .
4. **A Numerical Scale for Plausibilities.** Next we wish to measure plausibility on a numerical scale: that is we want to represent the plausibility of the truth of any proposition by a real number on a scale running from zero for false or untrue, up to one for certain or true. Any numerical function measuring plausibility must satisfy some simple constraints on its manipulation and combination though. For example, if there is more than one order in which to assemble hybrid propositions (for example,  $A$  and  $B$ ), then the function and its manipulation must produce consistent results. Under such simple constraints it turns out that there is a unique plausibility function,  $P$ , mapping propositions onto the interval zero to one, such that

$$P(A|X) + P(\text{not}A|X) = 1, \tag{1}$$

for all propositions  $A$ ; and

$$P(A \text{ and } B|X) = P(A|B, X).P(B|X), \tag{2}$$

for all propositions  $A$  and  $B$ . We call this function  $P$  the **probability**.

5. Notice that we have a sum rule and a product rule. The first condition (1) allows us to sum up probabilities over mutually exclusive events, and if, for example, a set of  $m$  such events,  $E_i$  say, are all equally probable, then each must probability  $P(E_i|X) = 1/m$ . The second condition (2) says that we must use multiplication to work out probabilities for hybrid propositions (logical “and”s).
6. The bald fact is that these two rules can be derived from scratch (as solutions to certain functional equations, by Cox – see Jaynes’ excellent book <sup>1</sup>), as a consequence of the requirements of plausible reasoning, without any recourse to frequentist repetition of experiments. This is not splitting hairs. Now we can apply probability theory to situations which are not equivalent to any repeatable thought experiments – and probability (though necessarily subjective in this context) represents our (un)certainty in any and all propositions we dare to consider in self consistent way.

*Have confidence: if we stick to (1) and (2) and their direct descendants we will not go wrong!*

7. If  $A$  is independent of  $B$ , then knowledge about  $B$  has no impact on our knowledge about  $A$ . So  $P(A|B, X) = P(A|X)$  and then we see that we can simply multiply up probabilities:  $P(A \text{ and } B|X) = P(A|X).P(B|X)$ . This is a consequence of the more general product rule (2) — but in schools is often taught first!
8. Next we present **Bayes’ theorem** which is named after the Reverend Thomas Bayes, 1702-1761: his theory of probability was published posthumously in 1764.

New evidence (an event,  $E$ , is observed), “ $E$  is true”, updates our estimates of probabilities in a constant and repeatable way. Suppose we wish to consider the probability that  $A$  is true. Then with no further evidence we have  $P(A|X)$ . Knowing further evidence,  $E$  also, we can consider two ways to write  $P(A \text{ and } E|X)$  from (2):

$$P(A \text{ and } E|X) = P(A|E, X).P(E|X) = P(E|A, X).P(A|X).$$

Hence from this we get the **Posterior** (post  $E$ ) probability,  $P(A|E, X)$  updated from the **Prior** probability,  $P(A|X)$ :

$$P(A|E, X) = P(E|A, X).P(A|X)/P(E|X). \tag{3}$$

This last equation is known as **Bayes’ theorem**. It is often written in different ways. Suppose we have a set of two or more alternative, mutually exclusive and exhaustive, propositions,  $A_i$  say. Then for each  $A_i$   $P(A_i|X)$  is the prior probability. Once we know that “ $E$  is true”, we can use (3) to update these. Since all such equations contain the same denominator,  $P(E|X)$  we often just write

$$P(A_i|E, X) \propto P(E|A_i, X).P(A_i|X). \tag{4}$$

---

<sup>1</sup>Jaynes, E.T. and Bretthorst, G.L. (2003) *Probability Theory: The Logic of Science* Cambridge University Press, Cambridge.

Then if the  $A_i$ s are exhaustive as well as exclusive, we can always normalize the set right hand sides they must sum to unity) to obtain an expression that avoids having to deal with the term  $P(E|D)$  which we may not know explicitly (though we can know it now !):

$$P(A_i|E, X) = \frac{P(E|A_i, X).P(A_i|X)}{(\sum_j P(E|A_j, X).P(A_j|X))}. \quad (5)$$

The terms  $P(E|A_i, X)$  used in the updating (from prior to posterior s for the  $A_i$ ) can be thought of as **model** terms. Under each separate hypothesis, that  $X$  and  $A_j$  is true, we must have such a model available for calculating this probability for  $E$ .

9. More notation. Sometimes it is easier and mathematically convenient to talk about **odds** rather than probabilities. If  $p$  denotes the probability  $P(A|E, X)$  for some proposition,  $A$ , conditional on  $X$  and some event data,  $E$ , say. Then the odds on  $E$  are simply given by

$$O(A|E, \dots, X) \equiv \frac{P(A|E, \dots, X)}{P(not A|E, \dots, X)} = p/(1 - p).$$

If we know the odds,  $O(A|E, \dots, X)$  we can easily find  $P(A|E, \dots, X)$  and vice versa.

Just as we have prior and posterior probabilities we can have the corresponding prior and posterior odds.

Now consider (3) as it is written, and again replacing  $A$  with its complement *not*  $A$ . Then taking the ratio we have Bayes' theorem written for odds:

$$O(A|E, X) = \frac{P(E|A, X)}{P(E|not A, X)}.O(A|X). \quad (6)$$

This formulation is very useful - it avoids the term  $P(E|X)$  again since it is a ratio of two equations each with a division by that term. It says the Posterior odds are equal to the prior odds multiplied by a ratio of the model terms. If we have a model  $P(E|A, X)$  and a single model for  $P(E|not A, X)$ , then we can use these directly. If we have split *not*  $A$  up into a greater number of mutually exclusive alternatives then this last formula gets a little more tricky.(See multiple hypothesis testing later in section 6).

The model ratio term in (6) is often called a **Bayes Factor**.

Before we go any further let us imagine the kind of calculations we might make. If we observe a number of events which are all independent, then we may apply (6) successively. Each posterior, after each observation, becomes the prior as we update to account for the next observation. Since the events are independent we may simply multiply together all the Bayes factors so as to get the overall posterior odds. Numerically, if the  $P(E|A, X)$ 's are very small then the odds are also small and then this may result in underflows. So it is very good practice to take the logarithm of the odds. We will introduce Log Bayes Factors (LBFs) in section 13.

10. **Summaries.** So in any problem the priors belong to each of us: they are subjective. We may select them in a number of ways based on what we know. Sometimes we select them to make life easy for ourselves.

The model terms (and the Bayes factors) must be fit for our purpose. This is mathematical modelling: for each hypothesis we must derive or assert a suitable distribution for observable events (over the set of possible values, continuous or discrete) under the assumption that it is true.  $X$ , our prior knowledge and skill, is required here!

Finally we have a posterior. A distribution of probabilities over a number of competing hypotheses. It is our job to **summarize** this posterior. This can be done in a number of ways. Perhaps with one or two simple parameters: the modal value (corresponding to the peak - most likely hypothesis) or mean/expected value. Or perhaps we should specify a credible set of values/hypotheses. Or even summarize by presenting the whole distribution for inspection.

## 5 Immediate Applications

Now we are ready to go to work.

### 5.1 Example Application 1

Suppose a man is arrested in a New York neighbourhood suspected of committing a knife murder only one hour earlier. The NYPD believe he has a probability of  $p = P(\textit{guilty}|X)$  of being the murderer, where  $X$  represents everything that they know. And the NYPD know a a lot. When he is searched he is found to be carrying is carrying a knife. Does that make him more likely to be guilty?

At first site you may think it does. But we need a little more information to decide the question.

What is  $P(\textit{carrying a knife}|\textit{not guilty}, X)$ ? One in 10 men of his age carry a knife like his in the tough neighbourhood where the suspect lives and was picked up. That is our model in this case:  $P(\textit{carrying a knife}|\textit{not guilty}, X)=1/10$  .

What is  $P(\textit{carrying a knife}|\textit{guilty}, X)$ ? The policemen know that following a knife murder almost all murderers will discard the weapon as soon as they can. After one hour they estimate  $P(\textit{carrying a knife}|\textit{guilty}, X)=1/50$  (only 1 in 50 murderers would still have the knife). That is our model in this case. Now we apply (5). We have

$$P(\textit{guilty}|\textit{carrying a knife}, X) = \frac{(1/50)p}{(1/50)p + (1/10)(1 - p)} = \frac{p}{5 - 4p}$$

Hence for any prior  $p$ , the discovery of the knife makes the suspect's guilt less likely.

Of course we could get a different result if we change the "models" around a bit. But the lesson is clear. We can't jump to any conclusion until we estimate the conditional probabilities for the new evidence, under the alternative hypotheses (using some distinct "models" to do so).

Note if we use (6) then things are much simpler in terms of odds:

$$\begin{aligned} O(\textit{guilty}|\textit{carrying a knife}, X) &= \frac{P(\textit{carrying a knife}|\textit{guilty}, X)}{P(\textit{carrying a knife}|\textit{not guilty}, X)}.O(\textit{guilty}|X) \\ &= \frac{1}{5}.O(\textit{guilty}|X) = \frac{p}{5}. \end{aligned}$$

There is a vast literature on Bayesian probability in legal reasoning and other types of inference. The reason for it is precisely because the results can be surprising (apparent paradoxes); and there are lots of cases where inferences are drawn without all the full information (necessary to derive the alternative conditional probabilities for the new evidence) being considered (jumping to conclusions!).

## 5.2 Example Application 2

**The Monte Hall problem.** Here is a variant of a rather famous problem. You are on a game show. There are five doors. Behind one door is a prize car, behind the others there is nothing.

The game show host asks you to select two of the doors. You do so. Then the host (who knows where the car is hidden) walks up to the remaining three doors and opens up two of them revealing nothing behind them. He then invites you to stick with your original choice of two doors or to swap them both for the single remaining door. You will receive any prize behind the door or the pair of doors that you have after this. Should you swap?

The prior odds that the car is behind one of the two doors you first selected are 2/3 (the prior probability is 2/5 of course). We will use (6).

What is the probability that the host could open two of the remaining three doors revealing nothing, assuming you already have already got the car behind one of your doors? It is one. He can easily do it. What is the probability that the host could open two of the remaining three doors revealing nothing, assuming you already have NOT got the car behind one of your doors? It is also one. He can easily do it since he knows which one of the three doors is hiding the car so he can open the other two. Let  $A$  denote the proposition that the car is behind either one of the two doors you initially select. Let  $E$  denote the event whereby the game show host opens up the two empty doors.

Then (6) become

$$O(A|E, X) = \frac{P(E|A, X)}{P(E|not A, X)} \cdot O(A|X) = \frac{1}{1} \cdot \frac{2}{3}.$$

So the model ratio term in (6) is one: the posterior odds are the same as the prior odds. You still have a 2/5 chance of already having the car. But something has changed. The alternative (*not A*) has now been narrowed down to the car being behind a single door. The probability that the car is behind that door is 3/5 (the only remaining possibility if you haven't got the car). Hence you should swap - you should give up your two doors for the final door.

*Sloppy reasoning* would just say that originally all of the doors had one in five chance, so why give up two chances for just one? You should stick! Or that at the end we have three doors left, all are equally likely so two chances are better than one. Stock! But the single door that is offered in the swap has changed in status. Extra knowledge that of the host has been used to select that door by opening and revealing the other two doors to be empty. That door has been selected in a special way with extra insight and is not a randomly selected door from the five (as yours were). Its status has changed. It now has a 3/5 probability of hiding the car because the game show host knew where the car was from the start and he selected it to remain closed.

If the host DID NOT know where the car really was then this changes things completely.

In that case when he opens up a random pair of doors, selected from the three that you have not selected, he risked revealing the car and you would have lost immediately - game over! - but fortunately that didn't happen.

So the probability that the host could open two of the remaining three doors revealing nothing, assuming  $A$ , is one. He can easily do it. But the probability that the host could open two of the remaining three doors revealing nothing, assuming *not*  $A$ , is  $1/3$ . He had a two thirds chance of revealing the car and ending the game. So the model ratio term in (6) is 1 divided by  $1/3$ : equal to three. Thus the posterior odds are 2 (the prior odds being the same,  $2/3$ ):

$$O(A|E, X) = \frac{P(E|A, X)}{P(E|not A, X)} \cdot O(A|X) = \frac{1}{1/3} \cdot \frac{2}{3} = 2.$$

Hence in that case you should stick with the pair and not swap.

### 5.3 Example Application 3: Is my friend a cheat?

This example is given in D'Agostini's excellent book.<sup>2</sup>

You meet a new friend. You decide to go out for a drink. She suggests tossing a coin to see who buys each round of drinks. She tosses, you call, and you lose. You buy the drinks. Then she suggest you do it again for the next round. Again she tosses, you call, and you loose. You buy the drinks. This happens for each round of drinks you have. Is she cheating you somehow? When do you call a halt to this?

Suppose you are very trusting – you have never met a girl who would cheat you out of drinks. So your prior  $p = P(Cheat|X)$  is small, maybe  $1/100$  or  $1/1000$ . The prior odds on her being a cheat are  $O(Cheat|X) = P(Cheat|X)/P(Not Cheat|X) = p/(1 - p)$ .

If she is a cheat then  $P(You lose|Cheat, X) = 1$ . That is our model under the cheat hypothesis.

If she is not a cheat then  $P(You lose|Not Cheat, X) = 1/2$ . That is our model under the not-a-cheat hypothesis.

Thus, using (6), after the first loss the odds are

$$O(Cheat|\{loss\}, X) = 2 \cdot \frac{p}{(1 - p)}.$$

After  $m$  consecutive losses the odds are

$$O(Cheat|\{loss, loss, \dots, loss\}, X) = 2^m \cdot \frac{p}{(1 - p)}.$$

---

<sup>2</sup>G. D'Agostini, Bayesian reasoning in data analysis: A critical introduction, World Scientific Publishing, 2003

The posterior has the slightly more complicated form,

$$P(\text{Cheat}|\{\text{loss}, \text{loss}, \dots, \text{loss}\}, X) = \frac{2^m p}{(2^m p - p + 1)}.$$

Suppose  $p = 1/10$  how many tosses before you think it is more likely than not that she is a cheat? [4] How many if  $p = 1/100$ ? [7] Or if  $p = 1/1000$ ? [10]

### Exercise.

You friend is a pretty flagrant cheat. But what if she lets you win one in four tosses - by using her biased coin and calling heads herself each time? Suppose the sequence is  $S = \{\text{win}, \text{loss}, \text{loss}, \text{loss}, \text{win}, \text{loss}, \text{loss}, \text{loss}\}$ : how do you feel about her at the end of the evening after eight drinks? Even if  $p = 1/10$ , you still think  $P(\text{Cheat}|S, X) = 0.240\dots$  Does the sequence matter or just that you won twice in eight rounds?

Show (using (6)) that in this case after exactly  $m$  losses and  $n$  wins the odds are

$$O(\text{Cheat}|\{m \text{ losses}, n \text{ wins}\}, X) = \frac{(1/4)^n (3/4)^m}{(1/2)^{n+m}} \cdot \frac{p}{(1-p)}.$$

If cheats let you win sometimes it's much harder to accuse them!

## 5.4 Example Application 4: The Swine 'Flu Test

Suppose there is a new virus that is difficult to detect. But there is also test (blood test or something similar).

It is believed that 1 in 100 people have swine fever (SF).

The test is 98% accurate for those with SF:  $P(\text{Positive}|SF, X) = 0.98$  The test produces 2% false positives:  $P(\text{Positive}|notSF, X) = 0.01$ .

Before anybody takes a test their probability of having SF is therefore 0.01. Suppose he or she takes the test and gets a positive result: what is the probability that he or she has SF now?

The prior odds of having SF are 1/99. Applying (6) we have

$$O(A|E, X) = \frac{P(E|A, X)}{P(E|not A, X)} \cdot O(A|X) = \frac{.98}{.02} \cdot \frac{1}{99} \sim \frac{1}{2}.$$

So as a result of the prior population bias even getting a positive test means that the testee is still twice as likely NOT to have SF than to have it!

This example stresses the importance of NOT simply focusing on the model terms - which make the test look awesome. But to consider the bias on the conditions that

each hypothesis is likely to occur. Of course a positive test makes it more likely that the testee has SF: so it is useful as a screen. If we screened out all those members of a public sample who get positives then we would still need to do some more work before allocating expensive drugs or treatments.

## 5.5 Example Application 4: The transposed conditional

While writing these notes, today there was a good example of very sloppy plausible thinking in The Times (which had to be corrected by a letter from the President of the RSS, no less). Drinkers look away now please!

The Times report that a high number of patients in high dependency clinics (IHDC) with (near) Liver Failure (LF) were middle class drinkers (MCDs) who drink more than a few glasses of wine at home every evening. The article implied that MCDs had an increased probability of suffering from LF.

But we do not have enough information: we know  $P(MCD|IHDC, LF, X)$ . Suppose 9 out of 10 folk with LF in HDCs are MDCs. We have  $P(MCD|IHDC, LF, X) = .9$ .

The article sought to imply something about  $P(LF|MCD, X)$ : transposing the conditional and the consequence; and dropping another condition (IHDC)! It implied in particular that  $P(LF|MCD, X) > P(LF|X)$ , the probability of some random adult suffering with LF.

In fact this example is like that discussed in section 5.1, and it is still possible that  $P(LF|MCD, X)$  is less than  $P(LF|X)$ .

Let us make some further assumptions.

First we must deal with the IHDC proposition. In odds notation we already have

$$9 = O(MCD|IHDC, LF, X) = \frac{P(IHDC|MCD, LF, X)}{P(IHDC|notMCD, LF, X)}.O(MCD|LF, X).$$

The middle classes have *sharp elbows*, so our model for tendency is that middle class adults suffering LF are, say, 36 times more likely to be IHDC than somebody from the lower classes (notMDC) suffering from LF - the MCDs always visit their doctors when they are unwell. Perhaps the lower classes simply carry on drinking and die; or stay away from HDCs at any rate.

$$9 = 46.O(MCD|LF, X) : \text{ so } O(MCD|LF, X) = 1/4, P(MCD|LF, X) = 1/5.$$

Now directly from Bayes theorem again:

$$P(MCD|LF, X) = \frac{P(LF|MCD, X)}{P(LF|X)}.P(MCD|X).$$

Suppose that  $P(MCD|X) = 1/4$ , that is MCDs make up a quarter of all adults. Then we have

$$\frac{P(LF|MCD, X)}{P(LF|X)} = \frac{P(MCD|LF, X)}{P(MCD|X)} = \frac{1/5}{1/4} = \frac{4}{5}.$$

Well I just made up the extra facts that were missing in the article. But the point is clear. Even if  $P(LF|MCD, X)$  is large we must not confuse it with  $P(MCD|LF, X)$ . This error is so common it is called the “transposed conditional”.

## 5.6 A possible application: on-line casino monitoring

On-line casinos have a particularly exposure to various types of modes of player behaviour which may be fraudulent (deliberate losses to launder money or to draw cash from stolen card Ids, etc). These are a problem to the operators who must be able to unwind and refund such transactions as a condition of their merchant’s license (to use credit card payments etc) and their gaming licenses.

The possible application here is to employ a variant of multiply hypothesis testing in order to provide real time continuous monitoring (predicting a likely current mode of play for each player) and a periodic playing behaviour summaries.

The first problem would be to passively monitor selected virtual tables of players and detect the increased likelihood of fraud and various types of normal and aberrant behaviour, including naivety, idiocy, and so on. The input data is rolling, and monitors play on all live tables (or all live tables containing any new -less than n hours/hands - players), hand by hand; according to, for example, we might monitor age of table(hours); commitment (cumulative investment) of all players; nature and sign-in of players (landing, mode, demographics, region); no. of players on table; sequence/number of low probability losses; sequence/number of low probability wins; size of betting behaviours; volatility (deviations) of betting behaviour of all players; day part week part; and other

The objective would be to use these time dependent and static inputs, updated after every hand played, to indicate the probability that various types of fraudulent or cheating behaviour may be taking place. Hence to prioritise attention from available monitoring staff for possibly intervention etc. Depending upon the type of behaviour exhibited the play and players may be dealt with directly in a number of ways (from the unwinding of hands/transactions, to advice to attend learning courses, etc). All tables could be continuously ranked ordered according to the likelihoods than particular type of fraud is taking place.

Since the type of games played on-line will most likely evolve it is essential to have a monitoring system that can learn new games and new behaviour within games from supplied data, and that is not specific to the details of a particular game or particular aberration of playing behaviour (that is, be calibrated by new data sets). In practice there will be only a few staff tasked with on-line table monitoring and intervention, so

a system of prioritising tables (there may be thousands concurrently) is essential. The ability to demonstrate such an automated aberration and fraud detection process may in future become part of the requirements for grant of licenses.

The second problem would be to provide periodic (weekly perhaps) summaries, of the incidence of such activities, comparative key performance indicators, an indication of increases or decreases in losses, and the consequent need to increase surveillance or intervention resources. This would be a periodic application of the algorithm(s) applied retrospectively at the close of all tables giving advice as to the extent of normal and aberrant behavioral session or tables and indicating the extent to which such behaviour(s) is challenged, monitored or ignored by security, monitoring or interventionist staffing.

## 6 Bayesian Multiple Hypothesis Testing

In this section we consider a class of very general problems in which we must decide whether one of a number of distinct things is happening - possibly in real time as data come in.

Consider the problem of observing a sequence of events emanating from some unknown source. Suppose that the source may be in exactly one of  $N$  possible modes or, equivalently, may be of exactly one of  $N$  ( $\geq 2$ ) possible types.

We wish to use all of our knowledge about past (similar) events, about the behaviour of such sources in the alternative modes, and the newly observed events so as to make a decision as to which mode the source likely to be in.

This type of problem requires us to test simultaneously the  $N$  hypotheses,  $H_k$ , that the source is in mode  $k$  ( $k = 1, \dots, N$ ) against each other. After observing some events, we will have a current best (most likely) hypothesis and we may prefer one so much that we may take some decision consequent to it being correct. Let  $X$  denote our prior information of all events and all we know about the alternative modes. Let  $D$  denote the newly observed "event" data. When  $D$  changes our (posterior) estimates for the probability of each hypothesis  $P(H_k|D, X)$ , will be updated via Bayes' Theorem, depending upon both the priors and the set of  $N$  models for the "event" data,  $D$ , under the alternative hypotheses.

As before let  $O(H_k|X) = P(H_k|X)/(1 - P(H_k|X))$  and  $O(H_k|D, X) = P(H_k|D, X)/(1 - P(H_k|D, X))$  denote the prior and posterior odds that  $H_k$  is true.

Then using (3) we have

$$P(H_k|D, X) = \frac{P(D|H_k, X)P(H_k|X)}{P(D|X)} = \frac{P(D|H_k, X)(1 - P(H_k|X)) \cdot O(H_k|X)}{P(D|X)}$$

and

$$P(\text{not } H_k|D, X) = \sum_{j \neq k} P(H_j|D, X) = \sum_{j \neq k} \frac{P(D|H_j, X)P(H_j|X)}{P(D|X)}$$

and hence

$$O(H_k|D, X) = \frac{P(D|H_k, X)}{\sum_{j \neq k} P(D|H_j, X) \frac{P(H_j|X)}{(1 - P(H_k|X))}} \cdot O(H_k|X). \quad (7)$$

Compare this with (6). The terms  $P(H_j|X)$  (appearing within both the prior odds and within the denominator of the updating term) are the priors. If  $N = 2$  then notice that the updating term simplifies to a simple likelihood ratio:  $P(D|H_k, X)/P(D|H_j, X)$  (where  $j$  is not  $k$ ) as in (6): but this is not our general situation.

However this often causes confusion. If we have only two hypotheses ( $A$  and  $\text{not } A$ ), the odds are updated simply by multiplying by a Bayes factor which contains only model

terms (the ratio of the probability of the data under the alternative hypotheses). For  $N$  greater than 2, this simply does not happen, and to update the odds we must multiply by a factor that involves both the model terms and the priors. That is how it is!

### Exercise

What if the drinking friend, from the exercises in the last section, has three possible modes of operation – ( $H_1$ ) she is honest; ( $H_2$ ) she has a biased coin and wins three in four times; or ( $H_3$ ) she is somehow flagrantly cheating and you only win one in twenty times (when she make a silly mistake! Hic!). Then there are three alternative hypotheses to test. Suppose you think these have prior probabilities,  $P(H_k|X)$ , of .89, .1, and .01 respectively. First show that after you have  $m$  losses and  $n$  wins,

$$P(D|H_1, X) = (1/2)^{(n+m)}, \quad P(D|H_2, X) = (1/4)^n(3/4)^m,$$

$$P(D|H_3, X) = (1/20)^n(19/20)^m.$$

Use (7) to calculate updates for all three hypotheses after you loose 7 out of 10 rounds. In this exercise we have used a simply binomial model for each hypothesis, based on the facts of our drinking companies alternative strategies – so writing down th models was easy. And all three models and priors are need to update all three prior odds to the corresponding posterior odds.

Now consider the general set up: look at the updating terms in (7), that is, the ratios

$$\frac{P(D|H_k, X)}{P(D|not H_k, X)} = \frac{P(D|H_k, X)}{\sum_{j \neq k} P(D|H_j, X) \frac{P(H_j|X)}{(1-P(H_k|X))}}$$

which update the odds  $O(H_k|D, X)$  in (7) . In order to proceed, we simply need a set of models for the "event" data,  $D$ , under the alternative hypotheses,  $H_j$  for  $j = 1, \dots, n$ .

## 7 Discrete Distributions: Multinomial Models

A multinomial model is simply a way of describing a set of probabilities that some (random) variable is drawn from a discrete set of mutually exclusive and exhaustive alternatives.

Suppose  $\{B_j | j = 1, \dots, m\}$  denotes such a discrete set of alternatives possible classes or categories for an observable event, or quantity,  $b$ . A multinomial model for the random event  $b$  is a set of probabilities  $\{P_j | j = 1, \dots, m\}$ , such that

$$P(b \in B_j | X) = P_j, \quad \text{and} \quad \sum_{j=1}^m P_j = 1.$$

Suppose that we observe a number of such events, none of which effect any of the others (so the likelihood of each result is given by our model), we say they are independent. Suppose that for exactly  $n_j$  of these the result  $b$  is in  $B_j$ , then we have

$$P((n_1, n_2, \dots, n_m) | (P_1, P_2, \dots, P_m), X) = \prod_{j=1}^m P_j^{n_j}.$$

Hence we have a model for the likelihood of observed data, given the set of multinomial probabilities (that must sum to one).

If  $m = 2$  then we only talk about  $P_1 = p$ , say, since  $P_2 = (1 - p)$ , and we have a binomial distribution for the different outcomes (combinations of events) form a number  $N$  of experiments.

## 8 Continuous Distributions

Often we will wish to run a continuum of hypotheses against one another.

We will deal here with a distribution of probable values for some real parameter  $\lambda$ : the generalization to higher dimensions or other types of state space is obvious in almost all cases.

Suppose we have some real constant  $\lambda$  that is unknown. Let  $S$  be any subset of the real line and define the corresponding hypothesis,  $H_s$ , via

$$H_S = \text{“}\lambda \in S\text{”}$$

Then we introduce a **probability density function**, often called a **pdf**,  $f(\lambda|X)$ . This is a non negative real valued function,

$$P(H_S|X) = \int_S f(\lambda|X)d\lambda.$$

Each hypothesis is intimately linked to its set of hypothesised values,  $S$ . Clearly hypotheses are mutually exclusive if the intersection between their sets is empty (or of measure zero).

If  $S$  contains all possible values (the entire support of  $f$ ) then  $P(H_S|X) = 1$ . Hence  $f$  must be of unit mass as well as nonnegative.

Bayes theorem still applies – as it must. So for any new event or evidence,  $E$ , we can write

$$f(\lambda|E, X) = \frac{P(E|\lambda, X)f(\lambda|X)}{P(E|X)}.$$

This last follows by applying Bayes theory to  $P(H_S|X)$ , defined above, and its posterior counterpart,  $P(H_S|E, X)$ , and observing that  $S$  can be chosen arbitrarily.

Normalizing probability density functions is tedious.

We know that the total mass of any posterior pdf must be unity. So often we will prefer to deal with **non normalized density functions**, and write simply

$$f(\lambda|E, X) = P(E|\lambda, X)f(\lambda|X),$$

with the running proviso that we will own up and normalize such  $f$ 's whenever that is required.

Of course we also may stray into the area of  $f$ 's which are of infinite mass - not integrable - if we wish. These are called **improper density functions**. For example  $f(\lambda|X) = 1$  for all real  $\lambda$ , or  $f(\lambda|X) = 1/\lambda$  for all positive  $\lambda$ . This idea can indeed be very useful - since a posterior pdf may be integrable (thus proper) even when the prior is improper.

Suppose we have absolutely no prior information to constrain our thoughts about  $\lambda$ : how will we set our prior then? Improperly?

As successive events are observed we will evolve a posterior pdf. So a prior pdf should not rule any values out that we might have otherwise accepted later. Eventually, with *enough* observations, the prior becomes *less important*. Consult the references, especially [1], on this aspect.

Now consider an unknown real parameter  $\lambda$  which is used to *model* some observable  $z$ . We will have a model  $P(z|\lambda, X)$  which gives us

- a probability density function for  $z$ , given a value for  $\lambda$  – if  $z$  is continuous;
- a multinomial for  $z$ , given a value for  $\lambda$  – if  $z$  is categorical or discrete.

For example our model may assert that  $z$  is normally distributed about  $\lambda$  with unit variance, say. So given  $\lambda$  we can estimate a possible values for  $z$ .

Now suppose we observe an actual value  $z^*$ . What does this tell us about  $\lambda$ ? Well the event  $E$  is simply the observation itself: that  $z$  lies with any set  $S$  containing  $z^*$ .

So we have

$$f(\lambda|E, X) \propto P(E|\lambda, X).f(\lambda|X) = \int_S P(z|\lambda, X)dz .f(\lambda|X)$$

But  $S$  was arbitrary, about  $z^*$ : so we have

$$f(\lambda|E, X) \propto P(z^*|\lambda, X).f(\lambda|X).$$

Again we can normalize if we wish and write

$$f(\lambda|E, X) \propto \frac{P(z^*|\lambda, X).f(\lambda|X)}{\int P(z^*|\lambda, X).f(\lambda|X)d\lambda}.$$

Now let us start with some *pleasant* distribution for our prior ( $f(y|X)$ ), and some well chosen model  $P(z|\lambda, X)$  for our observable, and suppose that we observe a set of  $m$  independent measurements:  $E = \{z_1, \dots, z_m\}$ .

Then we will have the (improper) posterior

$$f(\lambda|E, X) = P(z_1|\lambda, X).P(z_2|\lambda, X)\dots P(z_m|\lambda, X).f(\lambda|X).$$

This will soon get rather hairy! When the models are algebraically complicated these become difficult to deal with. How will we summarize the posterior? We will probably struggle to find even its mean or mode.

There are two ways around this: a traditional approach as outlined in the next section – which uses some trickery to reduce the amount of algebra, by pragmatic choice of the prior; and the use of the computer to summarize and sample from the posterior, as written. This last possibility is only 30 or 40 years old and is the subject of much progress (see later sections).

## 9 Algebraic Convenience: Conjugate Priors

In the days before computers could be used so as to summarise and sample from posteriors, in order to avoid excessively complicated functions, a rather useful practice was developed: the use of **conjugate priors**.

The central and elegant idea is that if the model  $P(z|\lambda, X)$  is given, then rather than choose any prior,  $f(\lambda, X)$  for  $\lambda$ , if we had no better reason then we could make life very easy for ourselves by choosing  $f$  from a particular function family, such that the posterior would be from the same family. Such a family of functions is called a **conjugate prior** distribution for the chosen model distribution.

Let  $F(\lambda|\theta)$  be a family of pdfs (normalized or not) for  $\lambda$  parameterized by  $\theta$ . Then  $F$  is conjugate to the model,  $P(z|\lambda, X)$ , if the posterior is given by

$$F(\lambda|\hat{\theta}) = P(z|\lambda, X).F(\lambda|\theta),$$

where  $\hat{\theta} = \hat{\theta}(\theta, z)$ , is some well defined function. Notice that if the model is given then this last is a functional equation for  $F$  and  $\hat{\theta}$ .

For example, suppose  $z$  is a multinomial variable, with  $m$  categories/classes as in section 7. Then  $\lambda$  be the vector  $(P_1, P_2, \dots, P_m)$  of the unknown probabilities in our multinomial model.  $\lambda$  lives on the simplex:  $\lambda \geq 0$ , and  $\lambda.(1, 1, \dots, 1) = 1$ .

As we observe instances of the multinomial variable  $z$  we will change our opinion as to where the  $\lambda = (P_1, P_2, \dots, P_m)$  may lie.

Now or any reals  $\mathbf{s} = (s_1, s_2, \dots, s_m) \geq 0$  let

$$G(P_1, P_2, \dots, P_m, \mathbf{s}) = \prod_{i=1}^m P_i^{s_i},$$

be defined on the simplex  $\{P_i \geq 0, \sum P_i = 1\}$ . Strictly speaking we should have normalised  $G$  so that it integrates to one: but we can proceed with this improper form. Note that on this simplex  $G(P_1, P_2, \dots, P_m, \mathbf{s})$  has a maximum (modal value) at  $\mathbf{s}/\|\mathbf{s}\|$  (hint: use Lagrange multiplier to maximize  $G$  whilst constraining to the simplex). (If  $\mathbf{s} = 0$  then  $G$  is uniform.

Then suppose our prior “insight”,  $X$ , allows us to select some nonnegative real values for  $\mathbf{s}$  and to take the prior

$$f((P_1, P_2, \dots, P_m), X) = G(P_1, P_2, \dots, P_m, \mathbf{s}).$$

Now suppose that we observe some evidence,  $E$ , containing as set of independent instances of the categorical variable  $z$  with exactly  $n_i$  of them within  $C_i$ .

Let  $\mathbf{n} = (n_1, \dots, n_m)$ , then we have the posterior

$$f((P_1, P_2, \dots, P_m)|E, X) = \prod_{i=1}^m P_i^{n_i} . G(P_1, P_2, \dots, P_m, \mathbf{s}) = G(P_1, P_2, \dots, P_m, \mathbf{n} + \mathbf{s}),$$

which is of the same family as the prior. Hence  $G$  is the conjugate prior for the multinomial.

For many, many, observations the posterior this becomes peaked around its nodal value at  $\sim \mathbf{n}/\|\mathbf{n}\|$ .

Note that if  $m = 2$  and the multinomial is a binomial we usually write  $P_1 = p$  and  $P_2 = 1 - p$ ; and abuse the notation to write

$$G = G(p, \mathbf{s}) = p^{s_1}(1 - p)^{s_2}.$$

In this case  $G$  is called a beta distribution (when normalised) and will be discussed in more detail in the next section.

If we a priori think that a coin is likely to be a fair one we might select  $s_1 = s_2 = 10$ . But after observing  $Q$  consecutive heads ( $C_1$ ), and no tails, then we have the posterior  $G(p, (Q + 10, 10)) = p^{Q+10}(1 - p)^{10}$  which has a modal value at  $p = (Q + 10)/(Q + 20)$ .

In our next example of a conjugate prior suppose that we will observe/measure some real quantity  $z$ . As a model for  $z$  we will choose a Poisson distribution with intensity  $\lambda$ , some unknown parameter. We have the model distribution

$$P(z|\lambda, X) = \lambda e^{-\lambda z}.$$

Suppose next that we feel able chose the prior

$$f(\lambda|x) = \lambda^a e^{-\lambda b},$$

for some nonnegative constants  $a$  and  $b$ . Again we have not bothered to normalize this  $f$ . It has a modal peak at  $\lambda = a/b$  (Calculus!).

Then observing a single value of  $z$ , say at  $z^*$ , we have the posterior

$$f(\lambda|z^*, x) = P(z^*|\lambda, X).f(\lambda|X) = \lambda^{a+1} e^{-\lambda(b+z^*)}.$$

If we continue adding further independent observations, so that there are  $m$  in total, then the modal value approaches the inverse of the mean of the observed  $z$ -values: we will have  $\lambda_{mode} = (1 + a/m)/(\bar{z} + b/m)$ .

The two parameter family  $\lambda^a e^{-\lambda b}$  is the conjugate prior for the Poisson distribution.

## 10 Laplace and Laplace's law of succession

As we have observed earlier, many problems introduced within the subject of "probability" are of the following type: we know, or rather we are told, unambiguously about some experimental or chance situation and we asked to derive the probability of a certain

kind of result occurring. Almost all of the interesting and relevant problems that occur within business analytics are the exact inverse of this type of problem: we know what result, or observation, has occurred, and we need to make some statement about how this affects our knowledge about the experimental or chance situation which has yielded (which is, to us, more or less uncertain).

This “inverse” of the problem format is exactly what Bayes and Laplace had in mind: from the prior information and the new data what can we infer? It is a probability theory which can be used as an extended form of mathematical logic.

Suppose we have an urn containing some red and some white balls. All probability theorists just love urns containing balls – it is the fault of the Bernoulli’s (look them up!) – our prior information about them is that the balls are always well mixed in such urns and each has the same probability of being sampled though a single “draw”. But in this case we do not know how many balls of each type are in the urn – but we will though assume that both types of balls are contained. We draw a ball at random from the urn and examine its colour. We replace it and draw again. Suppose after  $N$  such draws, we have drawn a red ball exactly  $n$  times and a white ball exactly  $m = N - n$  times. What is the probability  $p$  that the next ball to be drawn (and any ball subsequent ball drawn independently) will be red?

We do not know  $p$  at this stage. But we can express our knowledge about the possible values that the parameter  $p$  might take by using a **probability density distribution**,  $f(p|x)$  say. As before this is a nonnegative function defined over all of the possible values of  $p$  (in this case  $(0,1)$ ) such that the probability that the true value of  $p$  lies within the interval between two constants  $a > 0$  and  $b > a$  (but less than one) is given by the integral of the density distribution over the interval:

$$P(a < p < b|X) = \int_a^b f(p|X)dp.$$

Note that we are using a probability distribution for the value of the parameter,  $p$ , which is itself a probability. Do not confuse these. The distribution represents our knowledge as to where  $p$  is likely to be found. But  $p$  itself is a parameter of the model situation, which, since it happens to describe some random chance happening, is itself a probability – but must take some exact – though to us unknown – parameter value. Semantics help here : the “chance”  $p$  is some parameter value which describes the model situation and hence how results are produced – if we only knew it!

When probability distributions for some parameter are integrated over all of its possible values we must obtain unity - since we are sure the parameter must take one such possible value. In our present case we have always

$$P(0 < p < 1|X) = \int_0^1 f(p|X)dp = 1.$$

If  $f$  has a maximum at some value  $p^*$ , then this represent the maximum likelihood value

or mode value for  $p$ . If the distribution is narrow and highly spiked then we must think that the true value for  $p$  lies very close to the mode value. Conversely if the distribution is rather flat with a large variance we must be unsure as to where the true value lies, and have no great preference for one estimate over another.

For any function of  $p$ , say  $G(p)$  which is given, the **expected value** for  $G$  is given by:

$$\langle G \rangle = \int_a^b G(p)f(p|X)dp.$$

In particular our expected value of  $p$  itself is just:

$$\langle p \rangle = \int_a^b pf(p|X)dp,$$

and our expected value of any power  $p^q$  is just:

$$\langle p^q \rangle = \int_a^b p^q f(p|X)dp.$$

Hence the variance of the distribution is

$$\sigma^2 = \langle p^2 \rangle - \langle p \rangle^2 = \int_a^b (p - \langle p \rangle)^2 f(p|X)dp.$$

Now let us return to thinking about our urn filled with some red and some white balls, where  $p$  is the probability that any ball to be drawn will be red. As our information about the possible true value of  $p$  changes, this will change the distribution, and hence our estimates. This information changes each time we draw a single ball.

At the start of this thought experiment we know only that  $p$  lies between zero and one. Let us assume a prior distribution,  $f_0(p|X) \equiv 1$ , that is uniform (and equal to one) for all values between zero and one. That is, before any balls have been drawn, we will assume that the probability that the model parameter  $p$  lies between  $a > 0$  and  $b > a$ , but less than one, is

$$P(a < p < b|X) = \int_a^b f_0(p|X)dp = (b - a).$$

Using Bayes formula after drawing  $n$  reds and  $m$  whites we have a posterior distribution

$$f(p|(n, m), X) \propto P((n, m)|p, X) \cdot f_0(p|X)$$

But  $f_0 \equiv 1$  so by integrating both sides of this equation we have

$$f(p|(n, m), X) = \frac{P((n, m)|p, X)}{\int_0^1 P((n, m)|p, X)dp}.$$

Now if  $p$  is considered known, then the probability of drawing each red independently is  $p$  and the probability of drawing each white independently is  $1 - p$ . Therefore we have the simple “model” for the event  $(n, m)$ :

$$P((n, m)|p, X) = p^n(1 - p)^m.$$

This is called the “binomial distribution” - the probability of drawing the various results  $(n, m)$ , given a value for  $p$ . Considered as a function of  $p$ , for  $(n, m)$  given, it is called the “beta distribution”. Above we see that our posterior distribution for  $p$  is just a normalised form of this function:

$$f(p|(n, m), X) = \frac{p^n(1 - p)^m}{\int_0^1 p^n(1 - p)^m dp}. \quad (8)$$

So what is our expected value  $\langle p \rangle$  for  $p$  - the expected probability of drawing a red ball in the next draw? We have the ratio of two integrals

$$\langle p \rangle = \frac{\int_0^1 p^{(n+1)}(1 - p)^m dp}{\int_0^1 p^n(1 - p)^m dp}. \quad (9)$$

Using integration by parts again and again we have:

$$\begin{aligned} \int_0^1 p^s(1 - p)^r dp &= \frac{r}{(s + 1)} \int_0^1 p^{(s+1)}(1 - p)^{(r-1)} dp \\ &= \frac{r(r - 1)}{(s + 1)(s + 2)} \int_0^1 p^{(s+2)}(1 - p)^{(r-2)} dp \\ &\vdots \\ &= \frac{r(r - 1) \dots 2 \cdot 1}{(s + 1)(s + 2) \dots (s + r - 1)(s + r)} \int_0^1 p^{(s+r)} dp \\ &= \frac{r!s!}{(s + r + 1)!}, \end{aligned} \quad (10)$$

where we use the usual factorial notation:  $s! = 1 \cdot 2 \cdot \dots \cdot (s - 1) \cdot s$ .

Using (10) twice in (9), we obtain:

$$\langle p \rangle = \frac{n + 1}{n + m + 2}. \quad (11)$$

This (and its generalizations) is called **Laplace’s Law of Succession**. Laplace first gave it in 1774 and it has played a major role in the story of Bayesian inference. Many books on probability theory try to ignore it: but it is remarkably useful for our purposes. It expresses an estimate for a probabilistic parameter based on observed success counts

(from observed results) AND upon other possible outcomes that we may or may not have observed yet. For large  $n + m$  it resembles a raw frequency fraction: yet for smaller sets it presents a very useful way to turn such counts into numerical values without excluding what we have not yet seen (a white ball in 1000 draws for example). It is based upon Bayes' theorem, and a principal of indifference in our selection of a prior (any value of  $p$  is equally acceptable — just tell me what it is!)

We can also see how good an estimate for the true value the expectation  $\langle p \rangle$  might be by considering the standard deviation - the square root of the variance - we have (again using (10)), and some rearrangement (exercise!):

$$\sigma^2 = \frac{\langle p \rangle (1 - \langle p \rangle)}{(n + m + 3)}.$$

As is often the case the standard deviation goes to zero like  $(n + m)^{-1/2}$  – increase the sample size by 100 to decrease the error by 10. But we are content with (11) even when  $n + m$  is small since it represents our expectation given our prior knowledge and the data we have; and we also now have an estimate,  $\sigma$ , as to how wide the resulting posterior distribution is.

Now let us consider some of the surprising things about Laplace's Law. First it does not depend on the number of balls in the urn – which we do not know. In fact the urn is just a mechanism for producing results – abstract it: we simply *generate* a red ball with some unknown probability  $p$ : what can we say about  $p$ ? The details of the shape of the urn or the number of balls in it do not matter. In our applications we will frequently need to estimate a parameter – the chance of a certain type result happening – based on a limited number of calibration results. Then we can use Laplace's Law.

Second it is never equal to zero or one. Even if we have never seen a white ball and have drawn  $N$  red balls in succession,  $\langle p \rangle = (N + 1)/(N + 2)$ . There is always a chance (based on our prior information that white balls can be in such urns) that the next draw will be white.

If we have two identical urns, but with distinct mixes of balls – and have drawn 100 red balls (no whites) from urn one and 1000 red balls (no whites) from urn two we can say that if we are to see a white ball appearing it is much more likely to come from urn one than urn two (because for the latter we have carried out more draws without a white appearing). This is consistent with our common sense.

**Exercise.**

Consider (8). Find the modal value for  $p$ . This is where  $p^n(1 - p)^m$  has a maximum. Consider how this differs from the expected value given by (11). What if we have made  $N$  draws and all of them are red balls?

**Example: Indicators of churn and direct marketing**

Here we consider how a customer's response to some direct marketing contact might be used as an indicator of her future intention.

Suppose we have two types of customers: customers that are about to defect, or leave our customer base, and customers that are not, and will remain loyal (at least for a while). Suppose we send customers emails regularly and seek take-ups to marketing offers.

Our test mail-outs shows that of 500 randomly selected customers e-mailed who remained loyal (for at least one more month), 300 took up the offer(s). Out of 100 randomly selected customers who subsequently defected (within a month), only 10 took up the offer. We want to know whether non take-up enhances the chance that a customer will defect (over the next month or so) – and by how much: should we follow up null-responders with a more directed contact?.

**Exercise**

First use Laplace's Law to show that we have the estimates

$$\begin{aligned}
 P(\text{take - up} | \text{will defect}, X) &= \frac{11}{102} \\
 P(\text{no take - up} | \text{will defect}, X) &= \frac{91}{102} \\
 P(\text{take - up} | \text{not defect}, X) &= \frac{301}{502} \\
 P(\text{no take - up} | \text{not defect}, X) &= \frac{201}{502}
 \end{aligned}$$

Suppose that 15% of all of our customers defect each month. Then for any customer chosen at random (with no further information available) we have the priors

$$P(\text{will defect}|X) = 0.15 \quad P(\text{not defect}|X) = 0.85$$

Show that if the customer is emailed and does not take up an offer, then

$$P(\text{will defect}||X) = \frac{\frac{91}{102} \cdot 0.15}{\frac{91}{102} \cdot 0.15 + \frac{201}{502} \cdot 0.85} = .28223 \dots$$

Hence a null response to the emailed offer increases the likelihood that the customer will defect from 15% to about 28%.

From a marketing point of view - to email everybody (100% of customers) and then follow up all null responders with direct contacts would require following up with a fraction

$$\frac{91}{102} \cdot 0.15 + \frac{201}{502} \cdot 0.85 = .474$$

of the entire customer base, 28.22% of whom are expected to defect.

**Exercise**

Generalize this last result to the case where we have a fraction  $\theta$  of customers who will defect each month. How does  $P(\text{will defect}||X)$  depend upon  $\theta$ ? What fraction of the customer base will be null responders?

The **general case of Laplace's Law of Succession** is given as follows. (Just think of urns containing balls of many,  $K$ , distinct colours.)

Consider an mechanism where there is a number  $K$  of possible mutually exclusive types of result,  $A_k$  say, for  $k = 1, \dots, K$ ; each generated by a corresponding causal process that remains constant. We suppose that each type of result,  $A_k$ , occurs with a probability  $p_k$  where  $\sum_{k=1}^K p_k = 1$ . Then suppose we have obtained  $N$  results, or have made  $N$  observations, and obtained  $A_k$  exactly  $n_k$  times (of course  $\sum_{k=1}^K n_k = N$ ). Then what estimates can we have for the  $p_k$ ?

This involves making some tricky integrals over sets in  $K$ -dimensional space. But the argument is analogous to that given above for the case  $K = 2$ . We obtain

$$\langle p_i \rangle = \frac{n_i + 1}{N + K}. \quad (12)$$

This generalizes (11) (simply put  $K = 2$ ,  $p_1 = p$ ,  $n_1 = n$ ,  $N = n + m$ ).

Equation (12) is very useful to us: whenever we want to express a probabilistic model (for use in Bayes' theorem) as a multinomial model, we need to estimate the probabilities based on some calibration data sets. These estimates avoid absolute zero (impossibility) and absolute unity (certainty), and allow for possibilities barely, or indeed never yet, observed in the data.

Now we come another very interesting feature. (12) depends upon  $K$  the number of results that are thought possible - even if some have never yet been observed. For example suppose we believe that our urn contains yellow, red and white balls; and we take ten draws obtaining four white and six red and zero yellow. Then we have

$$p_{red} = 7/13, \quad p_{white} = 5/13, \quad p_{yellow} = 1/13.$$

Next suppose that we are told that sometimes the urns may contain green balls also - as a result of a recent improvement in their manufacture and composition (we just have not seen any yet). If we believe this then we must recompute with  $K + 4$ . So

$$p_{red} = 7/14, \quad p_{white} = 5/14, \quad p_{yellow} = 1/14 \quad (\text{and } p_{green} = 1/14).$$

Hence  $p_{red}$  depends not just on ( $n_{red}$  and  $N$ ) but on the number of possibilities we are prepared to entertain.

Suppose next we are convinced that we can observe and discriminate between 1000 different colours and hues...

But perhaps this is a desirable feature since if we want to admit the possibility of rare types of result/event, for which there is little or no hard evidence, we have to allocate a small chance to their occurrence. Before any data is observed, if we have no prior bias then all such events are to be deemed equally likely. Hence Laplace moves us seamlessly from the prior state of knowledge (all equally possible), through the the small data situation, and on to the large data situation.

Unfortunately Laplace did not really help himself by using his law to calculate the probability that the sun will rise tomorrow given that it has done so for 5000 years!(without any prior knowledge of the workings of the solar system) The odds on tomorrow's sunrise are

$$p/1 - p = 1.83M.$$

Good. But any additional information will also alter our knowledge – and Laplace himself knew a great deal about the mechanics of the heavens – its hard for us to put that aside. Our knowledge of what may or may not cause the failure of the sun to rise means that this is not a good example: Laplace only meant this to represent a statement of knowledge given only the fact of  $n$  successes in  $N$  (independent) trials.

The debate that this formula stirred up has lead to 200 and more years of objections and woolly thinking. Nevertheless we stress that this formula is an extremely useful rule for us especially since it converts data into estimates for probabilistic model parameters that we can employ within models for random (biased) processes. In particular we can use it to calibrate multinomial models and, by extension, Markov models (that are a set state dependent multinomial models describing inter-state switching).

Pierre Simon Laplace (1749-1826) was a French mathematician born in Normandy. At the age of 19 he was keen to study science and so gave up the study of theology at the university in Caen and left for Paris where d'Alembert, impressed by Laplace, secured him a place at Ecole Militaire. By 1773 Laplace was admitted to the Academy of Sciences. His career then progressed both within mathematical circles and within public service. For example in 1799 Napoleon made Laplace the Minister of the Interior (but dismissed him six weeks later), and then raised him to the senate where he served as vice-president in 1803.

Later, recognizing the imminent demise of the republic, in 1814 Laplace voted for the overthrow of Napoleon and the restoration of the monarchy. Charles X made him a Marquis and he stayed in Paris until his death.

Celestial mechanics was his chief interest, working often in collaboration with Lagrange, but he found the time and inclination to write the *Theory Analytique de Probabilities* (1812) and *Essai philosophique sur les probabilities* (1814). The Laplace transform (a methods for calculating difficult integrals and thus solving certain differential equations) introduced within his work on probability has become a fundamental tool of applied analysis (therefore both loved and hated by students!).

## 11 Approximating a Continuous Density Function with a Multinomial

A particularly useful trick when there aren't a lot of observations present is to approximate a density function for a continuous variable with a multinomial distribution by partitioning up the support set.

Suppose that the observation of an event is represented by some vector,  $\mathbf{z}$ , of measurable quantities or classifications. We will call  $\mathbf{z}$  the features. And say that it lives in a feature space.

Suppose that there are  $J$  features,  $\mathbf{z}$  is of length  $J$ . Here we will assume that these measurements are real or at least integers. The extension to other types of variable (categorical for example) is obvious.

Now suppose that the feature space is partitioned into exactly  $m$  subsets, here called classes,  $C_1, \dots, C_m$ . In reality, any  $\mathbf{z}$  that is observed is drawn from an (unknown) density distribution  $f(\mathbf{z}|X)$  defined over the event space. We may use a multinomial model, induced by the partition  $\{C_r\}$ , to represent a simplified version of each of the  $f_k(\mathbf{x})$ . Set

$$P_r = P(\mathbf{z} \in C_r | X) = \int_{C_r} f(\mathbf{z}) d\mathbf{z},$$

to denote the probability that a random  $\mathbf{z}$  lies in  $C_r$  assuming all our prior experience  $X$ .

In practice we may not know  $f$  or the  $P_r$  very accurately. So we have to calibrate (estimate) the  $P_r$  based on some sets of observations. We may use Laplace's *law of succession* to do so based on a sample  $\tilde{\mathbf{z}}$ . This is discussed in section 10 on page 24. We have the estimate for  $P_r$  given by

$$\hat{P}_r = \frac{\text{number of } \tilde{\mathbf{z}} \in C_r + 1}{\text{number of } \tilde{\mathbf{x}} + m}.$$

This is a low resolution approximation to the full density function. Clearly its calibration is not sensitive to any errors in the  $\tilde{\mathbf{z}}$ 's that do not perturb them across the partition.

## 12 Example: Beatles or Stones?

In this section we want to present a monitoring application that will read the lyrics of a song constantly update the estimate as to whether the song was written by either the Beatles or the Rolling Stones. This problem is clearly analogous to a number of monitoring and decision problems in business, particular where we can monitor performance in real time. It has rather soft data - so we have to make some quantitative measurements under each hypothesis (that the lyrics were written by the Beatles or the Stones).

First we need some calibration data and we will deliberately keep things very simple. We selected five songs from each band (mid sixties and early seventies tracks only - vintage Stones!). Each song was chopped up into distinct 100-character length, “bite size”, pieces (including spaces and punctuation characters). These pieces will be called “bites”.

The lyrics obtained yielded 35 Beatles bites, 42 Stones bites, and 10 bites from another “mystery song” that will represent our input data, and where we want to attribute authorship.

Then for each “bite” we made counts of the number of vowels and the number of punctuation marks within them.

We normalised the counts so that all values were mapped onto the interval zero to one. So each bite became an  $\mathbf{z}$  on the unit square. For the vowel count we defined

$$z_{vowel} = \frac{\text{number of vowels}}{14} - 3,$$

For the punctuation mark count we defined

$$z_{punc} = \frac{\text{number of punctuations}}{7}.$$

Hence for each bite we set  $\mathbf{z} = (z_{vowel}, z_{punc})$ .

Next we generated five centroids as random vectors within the unit square. For both the calibration bites (under both hypotheses) and the bites from a mystery song, we mapped all  $\mathbf{z}$ s to the class corresponding to the centroid that was closest using the usual Euclidean (unweighted sum of squares) distance measure. In this way the  $\mathbf{z}$ 's were partitioned into five classes.

The use of centroids (and weights for the Euclidean metric) is a highly useful way of generating partitions.

The centroids used were

$$(0.542, 0.473)^T, (0.47, 0.524)^T, (0.521, 0.468)^T, (0.547, 0.485)^T, (0.467, 0.463)^T.$$

For the calibration sets of the Beatles and Stones data we used Laplace's Law of Succession to yield the following estimates of the multinomial probabilities of an  $\mathbf{z}$  being

within each cell conditional of each hypothesis (designating authorship):

$$(P_{Beatles,1}, \dots, P_{Beatles,5}) = (0.125, 0.075, 0.150, 0.025, 0.625)$$

$$(P_{Stones,1}, \dots, P_{Stones,5}) = (0.234, 0.128, 0.128, 0.0426, 0.468).$$

These multinomials are the two models:  $P(\mathbf{z}|H_{Beatles}, X)$  and  $P(\mathbf{z}|H_{Stones}, X)$ .

Next we selected our priors  $P(H_{Beatles}|X) = P(H_{Stones}|X) = 1/2$  to representing our starting point of indifference.

So the prior odds on the Beatles  $O(H_{Beatles}|X) = P(H_{Beatles}|X)/(1 - P(H_{Beatles}|X))$  is one.

Successively the ten bites from the mystery song were in classes 5, 3, 3, 5, 5, 5, 5, 5, 3, and 5. After each bite we recomputed the odds using (35) ( each time multiplying the previous odds by the updating term). We obtained

$$O(H_{Beatles}|Bites, X) = 1.34, 1.57, 1.84, 2.46, 3.28, 4.39, 5.86, 7.82, 9.19, 12.27$$

respectively.

After analysing the lyrics the method asserts that it is 12 times more likely to be a Beatles song than not.

The mystery song was *Strawberry Fields*.

This was a very simple example and had some of the bites from the mystery song appeared within classes 1 or 4 then the odds would have been reduced at the corresponding updating step. However it is important that we draw confidence from this example. It was simple and successful. And at any point we could have given our best estimate of attributed authorship – so this method can be used in real time monitoring situations.

In practice we should next make some attempt to optimize the model. The method of Log Bayes Factors would be a good one here (see later, section 13). Effectively in such a method we estimate the evidential value of an a bite appearing in each class. Then we take a population average of such values, so that we get a typical evidence value from a single bite. Clearly we want to choose a partition where this is large, and hence the model is likely to discriminate with as little data as possible. We defer this to the next section where we show how to consider alternative measures as discriminators. In this case they come from the imposition of a partition, but in more general situations they may be variables directly measured in experiments.

Note that this example was really easy to compute because we only had two hypotheses competing – recall that the updating formula (35) is particularly sweet in that case, since the updating multiplier to be applied to the odds is independent of the priors (the prior probabilities before each new bite is added).

Using the definitions supplied above for  $\mathbf{x}$ , based on 100 consecutive characters from the lyrics (including spaces and punctuation); the centroids; and the probabilities from the

calibration sets (again all given above), you can now run your own tests on your favourite Beatles or Stones songs. Using the lyrics to *Angie*, as a second “mystery” example, then successively the nine bites from this mystery song are in classes 1, 1, 4, 3, 1, 1, 1, 1, and 1. After each bite we obtained

$$O(H_{Beatles}|Bites, X) = 0.53, 0.28, 0.17, 0.20, 0.10, 0.06, 0.03, 0.02, 0.01$$

Notice the 4th bite is in class three, and is the only time the odds lean away from the Stones.

More generally give any measurable attribute (real, discrete, binary,..) observed under two (or possibly more) alternative hypotheses we would like to have a method measuring its power in discriminating, so that we could contrast different attributes regardless of their data-type. This is what we will do in the next section.

## 13 Supervised Discrimination Problems

Supposes we have two different types or sets of observations: say observations made (with certainty) under two alternative hypotheses. Suppose also that for each we measure a possibly large number of variables or attributes. Then we want to know which of these attributes is most useful in determining the set to which any new observation belongs. This is called supervised discrimination. Typically we have a calibration set for which we know the type to which each observation belongs, and we wish to use these to find out how we can discriminate the type of any future observation based solely on the bases of a few of the attributes.

If the observations are samples then often we will have relatively few of these (100s in medical tests). If the attributes are genomic, such as SNPs (typically trinomial variables: XY, XX, YY) representing the genotype of an individual at a single base on the genome), then we may have thousands or hundreds of thousands available. Which attributes are markers for the type (a functional type is called a phenotype in genomics)?

Here we will introduce a method that allows us to contrast different types of variable: binary flags, multinomials, or real numbers by taking an approach based on *evidence*.

This methods has been encoded into a software application “Judge” that is part of the Reading Analytics Workbench (RAW), and is available to scientists in many disciplines at University of Reading.

We use a Bayesian measure of the “evidence” that an attribute provides for distinguishing one set of classifications versus another.

To begin with we label one set observations as “cases”, and the other as “controls”, a nomenclature common in biological scenarios. The “evidence” in question is a data average of log likelihood ratios evaluated by Bayesian methods, each known as a log Bayes factor or “LBF”.

We calculate the LBF that each observation is a case as opposed to a control based on a particular measured attribute. There is a different LBF calculation for discrete (categorical- multinomial) and continuous (real) attributes.

We will average the LBF’s for each attribute over its available data to produce the available evidence that each attribute provides separately for the cases versus controls distinction. This evidence may be used to rank attributes as markers for cases versus controls, and a shuffle test allows the user to assess the significance of the evidential value of each attribute.

## 13.1 Bayes Factors

Bayes Factors arise when, given some new observation,  $E$ , we try to compare the probabilities of two mutually exclusive hypotheses,  $H_1$  and  $H_2$ , that “it is a case” and that “it is a control”, respectively.

As before follows we have

$$\frac{P(H_1|E, X)}{P(H_2|E, X)} = \frac{P(E|H_1, X)}{P(E|H_2, X)} \cdot \frac{P(H_1|X)}{P(H_2|X)}. \quad (13)$$

As usual we must also take account of any prior modelling information,  $X$ .

The first of the two right hand terms is a ratio of the observation likelihood given each hypothesis, and is known as a Bayes Factor (BF). The second ratio is the ratio of the priors of each hypothesis given what we “know” (or can assert) already (called the prior odds).

The value of equation (13) is that it provides us with a route to quantitative calculation. The conditional hypothesis probabilities are not directly calculable, but the conditional observation likelihoods and their ratio are. There is often too some idea of the prior odds; in some situations there is no a priori reason to favour one model over the other, and it may reasonably be taken as 1. In other situations the prior odds are very important, as with false positives in a medical test for a rare condition (see D’Agostini 2005 p.82).

## 13.2 Calibrating the models

Calculating a Bayes factor is straightforward for us now: we have dealt with different types of models already (section 7 and 8).

If we consider an attribute that is real,  $z$  say, then we will select a pdf to model  $P(z|H_i, X)$ . Given a calibration data set under each hypothesis we will need to use these to tie down any parameters that are “hidden” in this notation, but that are involved in specifying this pdf model. Suppose the model depends on some parameters,  $\theta$  say, that we have selected our model for  $z$  (using our  $X$ !). For example  $\theta$  might be the mean and variance of a normal distribution, or of log normal (if  $z$  must be positive); or an intensity for a Poisson process; and so on.

Let us write  $P(z|H_i, X) = P(z|H_i, X)[\theta]$  to make the  $\theta$  dependence explicit.

Then we can use Bayes to calibrate the model (as in section 8). Let  $f_0(\theta|X)$  be a prior pdf for  $\theta$ . Let  $D_i$  be a set of observations for  $z$  under  $H_i$ . Then we have the posterior for  $\theta$

$$f(\theta|D_i, X) = \prod_{z \in D_i} P(z|H_i, X)[\theta] \cdot f_0(\theta|X).$$

So we must summarise this posterior: we select a good single estimate for  $\theta$  to use going forwards. Let us choose the mode for simplicity.

Note that if  $f_0$  is equal to 1 everywhere (is improper); then selecting the mode is equivalent to “maximum likelihood” estimation. We simply find the mode – the single value for  $\theta$  that makes the observations most likely to have actually occurred.

Hence we may estimate a suitable  $\theta$  value to calibrate our model using both the case and control calibration data sets. Remember there is no need to normalise the pdf for  $z$  ( $P(z|H_i, X)$ ): we will take a ratio of pdfs anyway in the BF.

Alternatively we might consider an attribute that is categorical,  $z$  say, with  $m$  alternatives, then we will use a multinomial model and employ Laplace’s law of succession so as to estimate the multinomial probabilities. Again we simply do this for each hypothesis: using case and control calibration data sets respectively.

Of course we can also make hybrid models which are a combination of more than one measured attribute... two tri-nomial SNPs become a nine-nomial; or a categorical and a real become a location model. Be creative!

Now what we wish to do next is calculate a measure of the power of any observable  $z$  (and thus pair of models - one for cases and one for controls) to discriminate successfully between the two hypotheses, given some new data (the observable,  $z$ ) without any assignment (to cases or controls). We want to find the “best ” marker/attribute.

### 13.3 Log Bayes Factors

In a situation with many observations, if they are assumed independent then an overall Bayes factor may be formed by multiplying all the individual ones together. This is equivalent to successive applications of (13).

Taking the logarithm of (13) is highly convenient at this point since the log of the product of separate BF’s is simply the sum of the separate logarithms of each Bayes factor (called Log Bayes Factors or LBF’s). The sum of the LBF’s provides a measure of the total evidence that the observations offer in distinguishing one model from another.

In practice each observation of a case or control may consist of one or more individual measured attributes ( $Y_A, Y_B, \dots$ ), as for example when each attribute is a single SNP (in genetics), or a single measured quantity (in metabolomics). To manage this we consider each attribute in isolation from the others and evaluate an LBF for each single attribute separately. For an attribute  $Y$  let  $y$  be a measured value observed from either a case or a control.

Then we define

$$LBF(y) = \log \left( \frac{P(y|H_1, X)}{P(y|H_2, X)} \right).$$

$LBF(y)$  tells us the change in the log odds that a single observation is a case as a result of knowing the attribute  $Y$  has value  $y$ . On the other hand  $-LBF(y)$  tells us the change

in the log odds that the observation is a control as a result of knowing the attribute  $Y$  has value  $y$ . The change in sign for the LBF is a consequence of the BF for a control hypothesis being the inverse of the BF for a case hypothesis (swap  $H_1$  with  $H_2$  in the BF ratio).

Let us consider a data set,  $E$ , containing a total of  $N$  cases and controls, each with their corresponding measured  $y$ -values, say  $y_i$  for  $i = 1, \dots, N$ . Let  $r_i=0$  for the cases and  $r_i=1$  for the controls. Suppose there are  $N_{ca}$  cases and  $N_{co}$  controls.

We will consider the odds on a hybrid hypothesis

$$H = \bigcap_{i=1}^N H_i,$$

where each individual hypothesis refers to an individual observation:

$$H_i = \{“y_i \text{ is a case” if } r_i = 0\}, H_i = \{“y_i \text{ is a control” if } r_i = 1\}.$$

In fact we know that  $H$  is true for this data set. But suppose we did NOT know this. Let us see how knowledge of the  $Y$ -attribute, all of the  $y$ -values, might improve the odds.

There must be some prior odds for  $H$ . This uses no knowledge of the attribute  $Y$ . And it plays no role in the following.

Then we can calculate the posterior for  $H$  (assuming independence): we simply sum the LBFs (under each separate hypothesis for each corresponding observation). Summing the LBF is equivalent to the product of the corresponding BFs. We have

$$O(H|E, X) = \sum_{i, \text{ Cases}} LBF(y_i) - \sum_{i, \text{ Controls}} LBF(y_i) + O(H|X).$$

Then consider the sum of “signed LBFs” in this last formula:

$$\sum_{i=1}^N (-1)^{r_i} LBF(y_i) = \sum_{i, \text{ Cases}} LBF(y_i) - \sum_{i, \text{ Controls}} LBF(y_i).$$

This is the change in the log odds for the whole set of the correct hypotheses,  $O(H|\dots, X)$  as a result of observing the set of  $y_i$ 's.

Since for this data set we actually know that  $H$  is true we wish to find an attribute for which this sum is as large as possible. Knowledge of such an attribute provides the best evidence that increases our certainty in the hybrid hypotheses  $H$  (which we know to be true in this case).

But since we wish to compare this term across different attributes (comparing apples with oranges; reals with categoricals, etc), and we must take care to make sure the evidence is fair to all: often for practical reasons some measurements are not available in all observations. To allow fair comparison between attributes we average each “signed LBF” contribution over its available observations/data.

This gives us a data normalized value as a final evidential value,  $ev$ :

$$ev(Y) = \frac{1}{N} \sum_{i=1}^N (-1)^{r_i} LBF(y_i). \quad (14)$$

Remark: Conceptually one can see that the observed arithmetic mean of the signed LBFs is equivalent to the log of the geometric mean of the corresponding BFs. It is thus the increase in the log odds,  $O(H|X)$  expected from a single measurement of the chosen attribute.

Hence the evidence value allows us to compare the expected power of alternative attributes  $Y$  to distinguish cases from controls.

### 13.4 Example

For example suppose we have 100 cases ( $H_1$ ) and 80 controls ( $H_2$ ). And an attribute called “sex” with values 0 and 1. Suppose 60 of the cases are 0’s and 40 are 1’s, whilst 30 of the controls are 0’s and 50 are 1’s.

Then we may use the models:

$$P(0|H_1, X) = 6/10, \quad P(1|H_1, X) = 4/10, \quad P(0|H_2, X) = 3/8, \quad P(1|H_2, X) = 5/8.$$

So that

$$LBF(0) = \log(8/5), \quad LBF(1) = \log(16/25)$$

.

Then

$$\begin{aligned} ev(sex) &= \frac{60 \log(8/5) + 40 \log(16/25) - 30 \log(8/5) - 50 \log(16/25)}{180} \\ &= \frac{30 \log(8/5) - 10 \log(16/25)}{180} = 0.103. \end{aligned}$$

Now suppose we have another attribute called “County” with values Oxon and Berks and Bucks. Suppose 50 of the cases are Oxon, 30 are Berks and 20 are Bucks; whilst 30 of the controls are Oxon, 30 are Berks and 20 are Bucks.

Then we may use the models:

$$P(Oxon|H_1, X) = 5/10, \quad P(Berks|H_1, X) = 3/10, \quad P(Bucks|H_1, X) = 2/10,$$

$$P(Oxon|H_2, X) = 3/8, \quad P(Berks|H_1, X) = 3/8, \quad P(Bucks|H_2, X) = 2/8,$$

so that

$$LBF(Oxon) = \log(4/3), \quad LBF(Berks) = \log(4/5), \quad LBF(Bucks) = \log(4/5).$$

Then

$$\begin{aligned} ev(County) &= \frac{50 \log(4/3) + 30 \log(4/5) + 20 \log(4/5) - 30 \log(4/3) - 30 \log(4/5) - 20 \log(4/5)}{180} \\ &= \frac{20 \log(4/3)}{180} = 0.032. \end{aligned}$$

Hence the County attribute is helpful but NOT nearly as useful as the Sex attribute in discriminating between cases and controls: on average it will not move the log odds as far in the correct direction.

### Significance

The data normalized evidence gives a fair way to compare the power of different attributes, but does not indicate the significance of each on its own. To assess this we may evaluate the evidence from the data after "maliciously" shuffling the measured values of an attribute amongst its available case plus control data. After each shuffle we get evidence that is only coincidental. We can then assess the significance of the original evidence by how it compares with the highest coincidental evidence from number of shuffles. For example if the genuine evidence is 2.5, and the best coincidental evidence from 20 shuffles is less (e.g. 1.2), it is clear that the genuine evidence is significant with approximately 95% confidence.

### Calculation of LBF's

This depends entirely on the models chosen for the attribute:  $P(y|H_1, X)$  and  $P(y|H_2, X)$ . But, once these are specified and calibrated from each set of cases and controls respectively, then everything proceeds as above. Just as in our example we use the set of cases and controls to determine any hidden parameters in these models: and then use them again within each component hypothesis within  $H$ .

For example reals might be represented/modelled by a standard distribution (Gaussian, Poisson,...). The models will contain parameters which themselves need to be determined using the observation data set given under each hypothesis. Then the LBF for each possible value follows.

Bowman and Delrieu (2005) give a good general explanation of LBF's and their usage for categorical and continuous attributes, and should be used as a reference for this underlying method when presenting results in publications.

## 14 Dynamical systems: unfolding a bifurcation

Here we follow the application presented in [5].

### 14.1 The Pitchfork

We consider an observable dynamical system, for which a mathematical model is available that has already yielded to a bifurcation analysis. We shall assume that (locally) the equilibrium solutions have been identified and are represented by a single well defined, real coordinate,  $x$ , which has been found to satisfy the pitchfork bifurcation equation:

$$x^3 - \lambda x = 0. \tag{15}$$

Here  $\lambda$  is a real system parameter, possibly controllable, though known theoretically.  $x$  will have been defined so that it measures the deviation away from some parameterised family of “trivial” equilibrium solutions. The parameter  $\lambda$  will have been shifted so that as  $\lambda$  moves through 0 there is a change in the linear stability of the trivial solution, due to a simple real eigenvalue moving through the origin: and hence the bifurcation occurs at the origin.

Typically the pitchfork bifurcation arises as follows: pretty much regardless of the type of dynamical system at hand. One considers an asymptotic expansion of non-trivial solutions near to the bifurcation point, and  $\lambda = 0$ , usually expanded in powers of a smaller parameter, which turns out to be  $\lambda^{1/2}$ . This is dominated by the first order term, the solution to which has a free amplitude coordinate,  $x$ , parameterising the centre manifold on which the loss of stability occurs. The second order term may next be resolved in terms of the first term. Then the third order term satisfies an equation which only has solutions provided that the inhomogeneous part (depending on both  $x$  and  $\lambda$ ) satisfies a solvability (orthogonality) condition (by applying the Fredholm alternative, for example). This last term yields the pitchfork bifurcation condition, above. This is a standard type of analysis for systems of ordinary and partial differential equations (see Chow and Hale, 1982, Golubitsky and Schaeffer, 1985 and Henry, 1981).

We will also assume that both observed values for  $x$  and  $\lambda$  are derivable (measurable) within our observable system. Typically one may only observe values along a stable branch of solutions for a dynamical system: so the data points near to (15) may not be available along all branches.

The symmetry of the pitchfork bifurcation equation is inherited from that assumed of the original system (the underlying modelling assumptions). External forcing terms (external fields, non uniformity, or inhomogeneities) assumed negligible within the model might erode this in practice. Also, the scaling of the bifurcation solutions (i.e. of  $x$ ) and the system parameters at the bifurcation point (fixing  $\lambda = 0$  there) will depend on the assumed values for the full system’s parameters, which may themselves vary or more likely be imprecisely known for any real system.

Hence, in practice, for any real observable system at hand, the behaviour will be altered by the addition of small extra terms. We consider the following unfolded version of the pitchfork:

$$x^3 - (\lambda + \varepsilon_2)x(1 + \varepsilon_1)^2 + \varepsilon_3 + \varepsilon_4x^2 = 0. \quad (16)$$

Here,  $\varepsilon_1$  has the effect of slightly re-scaling the amplitude of non-trivial solution branches, and  $\varepsilon_2$  has the effect of shifting the bifurcation point from zero to  $\lambda = -\varepsilon_2$ . In the preceding bifurcation analysis, these terms may be removed (scaled and shifted) and do not alter the unfolded bifurcation structure: but in contrasting experimental data with theory they are relevant tuning parameters to be determined, so we include them here. The parameters  $\varepsilon_3$  and  $\varepsilon_4$  are two terms that break the symmetry of the pitchfork in distinct ways:  $\varepsilon_3$ , removes the actual bifurcation point itself – the bifurcation curve splits into two components and becomes disconnected;  $\varepsilon_4$ , introduces a bias so that non-trivial solutions exist and coalesce (annihilating each other) away from the trivial one.

## 14.2 Estimating an unfolding

Suppose that we are to observe measured experimental data for  $x$  and  $\lambda$  from the real system, and must estimate actual values for the unfolding parameters  $(\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4)$ , as more and more measured data is obtained.

We shall consider a probability density function for any real vector of parameters, say  $\theta$ , about which there is an uncertainty. As usual we let  $X$  denote all of our prior knowledge about the situation(s) in which  $\theta$ , about which there is an uncertainty, is involved ( $X$  includes our theoretical bifurcation analysis and generally has subjective elements). Then our prior knowledge as to the whereabouts of  $\theta$ , about which there is an uncertainty, is represented simply by the prior conditional density function, denoted by  $P(\theta|X)$ . When new data, say  $D$ , is available, we update this to produce the non-normalised posterior density function:

$$P(\theta|D, X) = P(D|\theta, X).P(\theta|X). \quad (17)$$

As usual the middle term,  $P(D|\theta, X)$ , is our model, telling how likely the data is to be observed given any specific values for the parameters,  $\theta$ . We stress here that we can use this to modify our expectations about  $\theta$ , even when we have little or no hard data. We will be reasoning consistently and do not have to wait until *enough* data is available to begin estimating  $\theta$ : rather this is an ongoing task.

It is usual to summarize the behaviour of the posterior, in some fashion, in order to estimate our current uncertainty about  $\theta$ . For example, we might give the best present estimate for  $\theta$  as the modal estimate, (where the posterior is maximized, requiring some calculus!); or the expected value for the posterior (this would require some integration!). We may also use the posterior to indicate credible regions or bounds.

For our present purposes we will be content simply to examine the mode. Using these estimates for the unfolding parameters (and a further parameter describing the *noise* in the data) we can use (18) to give the corresponding current estimate for the bifurcation

structure. In fact, in our case below, the MLE estimate for  $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4)^T$  also turns out to be the expected value as we will choose a symmetric prior which is the conjugate to the model distribution.

First we introduce a model,  $P(D|\boldsymbol{\theta}, X)$ . Initially we assume that all of the  $\varepsilon$ 's are small, so as to have a model which is linear in  $\boldsymbol{\varepsilon}$ ; and we must also have to introduce a further parameter,  $\sigma$ , in order to model the error distribution implicit within the model.

Linearising the left hand side of (16) with respect to  $\boldsymbol{\varepsilon}$ , we obtain:

$$(x^3 - \lambda x) - (2\lambda x, x, -1, -x^2)^T \cdot \boldsymbol{\varepsilon} \approx 0. \quad (18)$$

Now suppose that we observe many pairs of values for  $x$  and  $\lambda$ , say  $(x_i, \lambda_i)$ . From (18) we assume that they satisfy

$$(x_i^3 - \lambda_i x_i) - (2\lambda_i x_i, x_i, -1, -x_i^2)^T \cdot \boldsymbol{\varepsilon} = e_i, \quad (19)$$

where the errors,  $e_i$ , are assumed to be independently normally distributed about zero with variance given by  $\sigma^2$ .

For a set of  $N$  such independent observations, we let  $\mathbf{y} = (y_1, y_2, \dots, y_N)^T$ , where  $y_i = (x_i^3 - \lambda_i x_i)$ , and let  $\mathbf{A}$  be the  $(N \times 4)$  design matrix with the  $i$ th row given by the row vector  $(2\lambda_i x_i, x_i, -1, -x_i^2)$ . Also set  $\mathbf{e} = (e_1, e_2, \dots, e_N)^T$ . Then, in vector notation, (19) becomes

$$\mathbf{y} - \mathbf{A} \cdot \boldsymbol{\varepsilon} = \mathbf{e}. \quad (20)$$

Hence the observed data  $(\mathbf{A}, \mathbf{y})$ , denoted by  $D$  earlier, gives rise to the  $e_i$ , depending on the value of  $\boldsymbol{\varepsilon}$ . Our model is completed by making a hypothesis that the  $e_i$  are normally distributed about zero with variance given by a further parameter  $\sigma^2$ . We therefore have the conditional probability, for the observed data,  $\mathbf{D} = (\mathbf{A}, \mathbf{y})$ , given the parameters,  $\boldsymbol{\theta} = (\boldsymbol{\varepsilon}, \sigma)$ :

$$P((\mathbf{A}, \mathbf{y}) | (\boldsymbol{\varepsilon}, \sigma), X) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left(\frac{-(\mathbf{y} - \mathbf{A} \cdot \boldsymbol{\varepsilon})^T \cdot (\mathbf{y} - \mathbf{A} \cdot \boldsymbol{\varepsilon})}{2\sigma^2}\right) \quad (21)$$

Next we consider our prior knowledge, denoted by  $X$ , gained from the mathematical analysis of our model, and all our previous experience in such matters. We expect to see a pitchfork bifurcation occur where both  $x$  and  $\lambda$  vanish. However, we suspect that one to four of our unfolding (and tuning) parameters may not be zero. Hence let us take our prior knowledge,  $X$ , to be that the full unfolded pitchfork will most likely be relevant, and our uncertainty as to the unfolding parameters may be represented by a prior distribution over  $(\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4)$  space.

The expected values for these parameters given by this distribution should perhaps be  $(0, 0, 0, 0)$  (unless we have any further prior experience, or knowledge, such as a feeling about the likely effects of terms such as gravity which the modeller has ignored). Perhaps the simplest, pragmatic, choice for our prior is a multivariate Gaussian distribution centred on the origin, with a covariance matrix,  $\mathbf{C}$ . Many other priors could be proposed

since this represents our uncertainty as to the exact nature of the modelling assumptions and their relevance to the actual experiment at hand: the less confidence we have, the wider the variances and the fatter the tails will be in the prior distribution. We could also incorporate constraints into the prior - co-dependences of the tuning parameters on more fundamental constraints, or even hard constraints on any of the values ( $\varepsilon_1 > -1$ , for example). However, since we wish to employ this formulation in order to use whatever small or large amount of experimental data we obtain in order to estimate possible values for the unfolding parameters, we shall keep the prior to be as simple as possible.

Therefore, before any data has been observed, and on the basis of our modelling, we expect that the parameters  $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4)^T$  will have values given by  $\boldsymbol{\varepsilon}_0 = (0, 0, 0, 0)^T$ ; and we will choose to employ the covariance matrix  $\mathbf{C}$  to represent the acceptability of all other various possible values. So we have our prior information for the possible values of  $\boldsymbol{\varepsilon}$ , given by:

$$P_{\boldsymbol{\varepsilon} \text{ prior}}(\boldsymbol{\varepsilon}|X) = \frac{1}{(2\pi)^2 |\mathbf{C}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}_0)^T \mathbf{C}^{-1}(\boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}_0)\right). \quad (22)$$

In O'Hagan and Forster (2004), the use of this distribution is suggested, on the grounds that it is the conjugate of that in (21) with respect to  $\boldsymbol{\varepsilon}$  (see earlier section).

This simply means that when (21) and (22) are multiplied together, the result is a distribution (the factor in the posterior governed by  $\boldsymbol{\varepsilon}$ ) of the same general form as (22). Specifically we obtain:

$$P((\mathbf{A}, \mathbf{y})|(\boldsymbol{\varepsilon}, \sigma), X) \cdot P_{\boldsymbol{\varepsilon} \text{ prior}}(\boldsymbol{\varepsilon}|X) \propto \exp\left(-\frac{1}{2}(\boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}^*)^T \mathbf{V}(\boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}^*)\right)$$

where  $\mathbf{V} = \left(\frac{\mathbf{A}^T \cdot \mathbf{A}}{\sigma^2} + \mathbf{C}^{-1}\right)$  and  $\boldsymbol{\varepsilon}^* = \mathbf{V}^{-1} \cdot \left(\frac{\mathbf{A}^T \cdot \mathbf{y}}{\sigma^2} + \mathbf{C}^{-1} \cdot \boldsymbol{\varepsilon}_0\right)$ . (23)

For a given  $\sigma$ , the mode and the expected value for  $\boldsymbol{\varepsilon}$ , are thus coincident at  $\boldsymbol{\varepsilon}^*$ .

Finally, having introduced  $\sigma$ , we must assert some acceptable prior distribution for its possible values. The variance  $\sigma^2$  cannot realistically be as large as the variance in the (to be) observed values in  $\mathbf{y}$ : so, for example, we might assert that  $\sigma$  is distributed by a log normal distribution, with  $\ln \sigma$  having an a priori expected value,  $\mu_\sigma$  (say 1/3 of a typical (or expected) value for  $\ln(\text{var}(\mathbf{y}))$ ): set  $\mu_\sigma = \frac{\ln(\text{var}(\mathbf{y}))}{3}$ ). By asserting a standard deviation,  $\eta$ , for the normal distribution of  $\ln \sigma$  equal to, say, 3, we may effectively allow for this assumption about  $\mu_\sigma$  to be in error by an order of magnitude and more either way. Hence we might choose:

$$P_{\sigma \text{ prior}}(\sigma|\mathbf{X}) = \frac{1}{\sigma} \frac{1}{\sqrt{2\pi}\eta^2} \exp\left(-\frac{(\ln \sigma - \mu_\sigma)^2}{2\eta^2}\right), \quad (24)$$

where the constants  $\mu_\sigma$  and  $\eta$  are known.

Now we may apply Bayes' rule with  $\boldsymbol{\theta} = (\boldsymbol{\varepsilon}, \sigma)$ . The joint posterior distribution for  $\boldsymbol{\varepsilon}$  and  $\sigma$ , after the experimental data observations are considered, is thus given by the multiple of the three distributions in (21), (22) and (24).

Rather than deal with this full distribution directly, for now we shall derive the conditions which determine the posterior modal values for  $\boldsymbol{\varepsilon}$  and  $\sigma$ , from this distribution.

Forming the product taking logarithms, and partially differentiating, firstly with respect to  $\boldsymbol{\varepsilon}$  and then with respect to  $\sigma$ , we obtain directly:

$$\boldsymbol{\varepsilon} = (\mathbf{A}^T \mathbf{A} + \sigma^2 \mathbf{C}^{-1})^{-1} (\mathbf{A}^T \mathbf{y} + \sigma^2 \mathbf{C}^{-1} \boldsymbol{\varepsilon}_0) \quad (25)$$

$$0 = \mathbf{e}^T \mathbf{e} - \sigma^2 N + \sigma^3 \frac{d(\ln(P_{\sigma \text{ prior}}(\sigma | \mathbf{X})))}{d\sigma} \quad (26)$$

The first of these, (25) (exactly as in (23)), is a kind of generalization of the usual normal equations obtained in a simple MLE regression analysis: as the data becomes more prevalent, the first terms in each of the brackets will grow, whilst the terms in  $\sigma^2$ , will hopefully converge. However (25) remains valid for little or no data; regardless of whether  $\mathbf{A}^T \mathbf{A}$  is invertible, or usefully pseudo invertible. Note that, in our case,  $\boldsymbol{\varepsilon}_0$  is zero: but we have kept it alive here for completeness. The error term  $\mathbf{e}$  in (26) depends directly on  $\boldsymbol{\varepsilon}$ : via  $\mathbf{y} - \mathbf{A} \boldsymbol{\varepsilon} = \mathbf{e}$ . So by substituting for  $\boldsymbol{\varepsilon}$  from (25) into (26) we obtain a single nonlinear equation for  $\sigma$ , which we may solve easily, by bisection for example.

If we adopt (24) as the prior for  $\sigma$ , then (26) becomes

$$0 = \mathbf{e}^T \mathbf{e} - \sigma^2 N - \sigma^2 \left( 1 + \frac{(\ln \sigma - \mu_\sigma)}{\eta^2} \right) \quad (27)$$

Notice (see O'Hagan and Forster, 2004) also from (26), that if our prior estimate for the modal value for  $\sigma$  happens to be correct, or if the prior for  $\sigma$  is a constant (the use of an improper prior reflecting our indifference), then the last term in (26) vanishes and the modal  $\sigma^2$  is simply the variance of the observed errors.

From (23) the modal value for  $\boldsymbol{\varepsilon}$  is also the expected value, and we can see that the covariance matrix in the distribution for  $\boldsymbol{\varepsilon}$  is simply the inverse of  $\mathbf{V}$ , so further information about the likely values for  $\boldsymbol{\varepsilon}$  is at hand, if we wish.

### 14.3 An Example

First we show (in Figure 1) the pitchfork bifurcation plotted as a curve,  $\lambda$  versus  $x$ , and a chosen unfolding with  $\boldsymbol{\varepsilon} = (0.02, 0.01, 0.02, 0.01)^T$ . Next we show (in Figure 2) some slightly noisy data: 20 points generated on the left hand branch. We shall consider them added into the analysis one at a time, in the order of  $\lambda$  increasing.

Finally, in Table 1, we show values obtained for  $\boldsymbol{\varepsilon}$ , computed as each successive point is added. Here set the prior assuming that each  $\varepsilon_i$  was assumed independent with a standard deviation of 0.01; hence

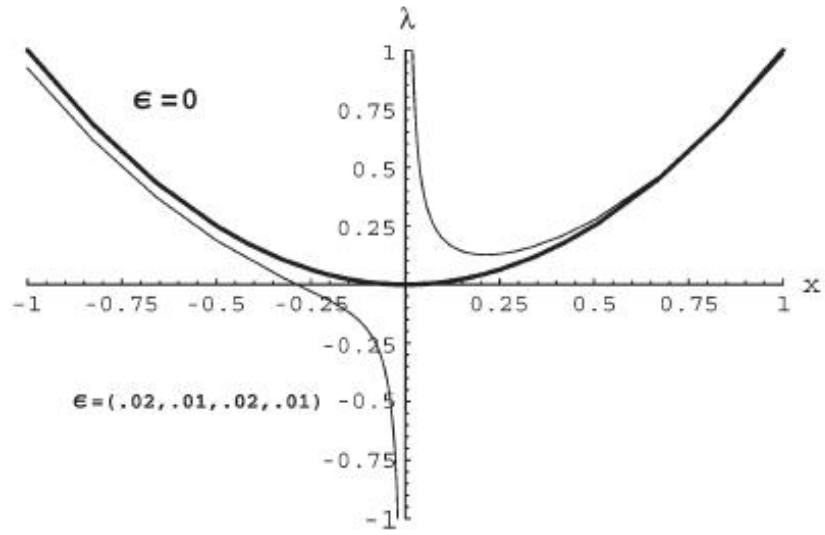


Figure 1: Pitchfork bifurcation with a chosen unfolding

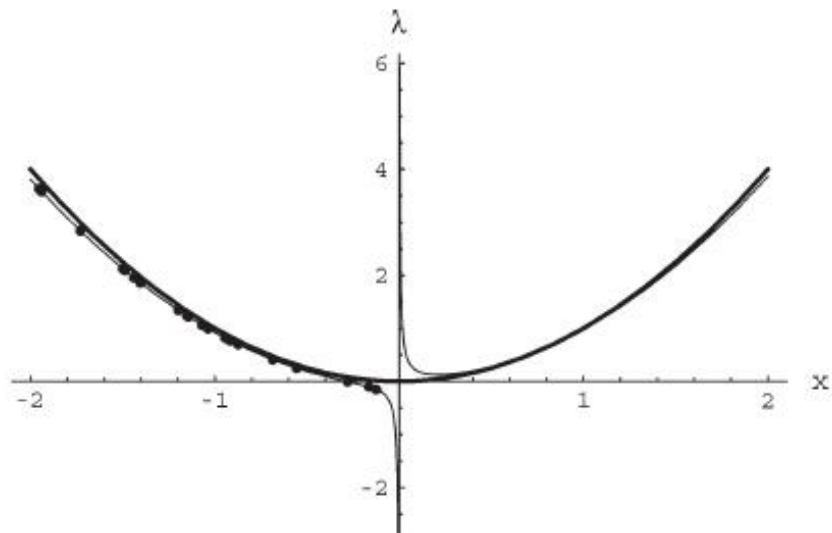


Figure 2: Data generated from our chosen unfolding

$$\mathbf{C} = \begin{pmatrix} 0.0001 & 0 & 0 & 0 \\ 0 & 0.0001 & 0 & 0 \\ 0 & 0 & 0.0001 & 0 \\ 0 & 0 & 0 & 0.0001 \end{pmatrix}.$$

No. of data points	$\varepsilon_1$	$\varepsilon_2$	$\varepsilon_3$	$\varepsilon_4$
0	0.00000000	0.00000000	0.00000000	0.00000000
1	0.00420671	0.00057908	0.00029639	0.00113141
2	0.00702109	0.00097372	0.00050022	0.00189543
3	0.00820411	0.00120166	0.00063669	0.00227392
4	0.00869562	0.00134239	0.00073621	0.00246870
5	0.00913940	0.00147218	0.00082878	0.00264663
6	0.00949148	0.00158540	0.00091265	0.00279545
7	0.00978683	0.00168729	0.00099023	0.00292531
8	0.00991306	0.00175518	0.00105134	0.00299702
9	0.01001460	0.00181608	0.00110836	0.00305855
10	0.01011108	0.00187510	0.00116403	0.00311771
11	0.01017867	0.00192511	0.00121449	0.00316418
12	0.02010993	0.01150625	0.01924807	0.00902476
13	0.02005592	0.01070382	0.01966117	0.00952820
14	0.02003611	0.01043498	0.01979279	0.00970497
15	0.02003158	0.01038282	0.01981576	0.00974244
16	0.02000503	0.01006421	0.01995975	0.00996635
17	0.01999562	0.00996018	0.02000424	0.01004281
18	0.01999371	0.00994185	0.02001129	0.01005741
19	0.01999509	0.00995362	0.02000720	0.01004739
20	0.01999576	0.00995899	0.02000545	0.01004263

Table 1: Estimated parameter values after each generated data point is added

## 14.4 An example using experimental data

This data [8] comes from an experiment in electroconvection in nematic liquid crystals. An electric field is applied to a liquid crystal with certain characteristics and when the “E-field” reaches some threshold value, convection is observed in the liquid crystal. Here  $\lambda$  measures the “E-field” applied and  $x$  measures the convection.

Firstly, Figure 3 shows the data together with the pitchfork bifurcation. Figure 4 shows

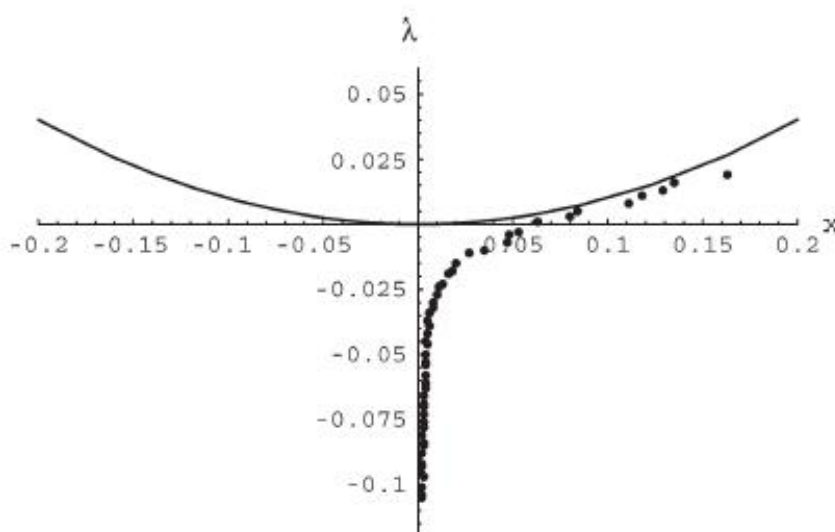


Figure 3: Real data with the pitchfork bifurcation

a fitted curve for  $\lambda$  which uses our estimates for the unfolding parameters, together with the data, including the unobserved branches of the bifurcation curve. Finally, Table 2 shows values for  $\varepsilon$ , computed as each point is added.

No. of data points	$\varepsilon_1$	$\varepsilon_2$	$\varepsilon_3$	$\varepsilon_4$
0	0.00000000	0.00000000	0.00000000	0.00000000
1	0.00000744	0.00019575	-0.00120095	-0.00003191
2	0.00205781	0.03046538	0.00398840	-0.00916645
3	-0.00582571	0.02310871	0.00279121	-0.00961714
4	-0.00486679	0.01750771	0.00200519	-0.01204971
5	-0.00590635	0.01204169	0.00123650	-0.01390697
6	-0.00113494	0.00592339	0.00059779	-0.02349996
7	0.00052503	0.00259609	0.00026456	-0.02952479
8	0.00365527	-0.00016404	0.00003573	-0.03633499
9	0.00677698	-0.00403441	-0.00024675	-0.04811532
10	0.00862809	-0.00662147	-0.00040567	-0.05775556

Continued on next page

Table 2: Estimated parameter values after each real data point is added

11	0.00900316	-0.00888101	-0.00055389	-0.06583826
12	0.00965076	-0.01077582	-0.00064572	-0.07447880
13	0.00775784	-0.00975918	-0.00058994	-0.07062793
14	0.00579412	-0.00852231	-0.00052795	-0.06551469
15	0.00461114	-0.00810602	-0.00050471	-0.06407483
16	0.00332980	-0.00758848	-0.00047809	-0.06209204
17	0.00212582	-0.00706456	-0.00045271	-0.05996663
18	0.00070715	-0.00639117	-0.00042225	-0.05706126
19	-0.00043744	-0.00595138	-0.00040192	-0.05524444
20	-0.00173114	-0.00539867	-0.00037810	-0.05281331
21	-0.00273399	-0.00509440	-0.00036412	-0.05159838
22	-0.00413104	-0.00456589	-0.00034243	-0.04923438
23	-0.00556162	-0.00405888	-0.00032204	-0.04695840
24	-0.00644189	-0.00388280	-0.00031374	-0.04633487
25	-0.00742585	-0.00364450	-0.00030345	-0.04538138
26	-0.00859457	-0.00333574	-0.00029096	-0.04406422
27	-0.00937350	-0.00324625	-0.00028628	-0.04384809
28	-0.01027271	-0.00308020	-0.00027907	-0.04323157
29	-0.01107052	-0.00297126	-0.00027396	-0.04289654
30	-0.01183603	-0.00288905	-0.00026985	-0.04269506
31	-0.01254349	-0.00285261	-0.00026742	-0.04272372
32	-0.01323921	-0.00284550	-0.00026607	-0.04290509
33	-0.01394194	-0.00285538	-0.00026534	-0.04317743
34	-0.01475018	-0.00275876	-0.00026104	-0.04288393
35	-0.01549564	-0.00269048	-0.00025774	-0.04273035
36	-0.01622168	-0.00263593	-0.00025493	-0.04264636
37	-0.01691371	-0.00260240	-0.00025286	-0.04266880
38	-0.01758978	-0.00258686	-0.00025142	-0.04278521
39	-0.01826474	-0.00258254	-0.00025037	-0.04296212
40	-0.01930833	-0.00246554	-0.00024563	-0.04258470
41	-0.02004304	-0.00249257	-0.00024560	-0.04294690
42	-0.02079197	-0.00252241	-0.00024566	-0.04332736
43	-0.02169203	-0.00244322	-0.00024224	-0.04312652
44	-0.02250869	-0.00238151	-0.00023941	-0.04300414
45	-0.02330274	-0.00232804	-0.00023687	-0.04292250
46	-0.02405121	-0.00228645	-0.00023473	-0.04289656
47	-0.02522426	-0.00236159	-0.00023606	-0.04362577
48	-0.02593988	-0.00233433	-0.00023440	-0.04367202
49	-0.02663339	-0.00231604	-0.00023304	-0.04376334
50	-0.02732124	-0.00230180	-0.00023182	-0.04387570

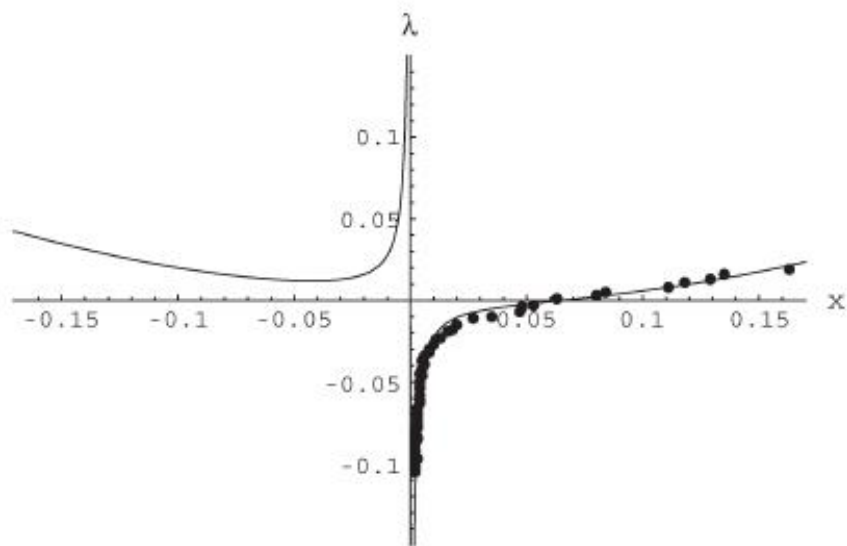


Figure 4: Real data with a fit using our estimates of the parameter values

## 14.5 Discussion

A key step in our solution methodology was the linearization of the unfolding that produces a simplified set of equations (as in (18)). This naturally limits the unfolding parameters to some neighbourhood of the origin, regardless of the values observed for  $x_i$  and  $\lambda_i$ . Suppose we do not wish to do this and instead that we define

$$g_i(\boldsymbol{\varepsilon}) = (\lambda_i + \varepsilon_2)x_i(1 + \varepsilon_1)^2 - \lambda_i x_i - \varepsilon_3 - \varepsilon_4 x_i^2 \quad (28)$$

and let

$$\mathbf{g}(\boldsymbol{\varepsilon}) = (g_1(\boldsymbol{\varepsilon}), \dots, g_N(\boldsymbol{\varepsilon}))^T. \quad (29)$$

We can then replace (20) by

$$\mathbf{y} - \mathbf{g}(\boldsymbol{\varepsilon}) = \mathbf{e} \quad (30)$$

and equations (25) and (26) become

$$\sigma^2 \mathbf{C}^{-1}(\boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}_0) = (\nabla_{\boldsymbol{\varepsilon}} \mathbf{g}(\boldsymbol{\varepsilon}))^T \cdot (\mathbf{y} - \mathbf{g}(\boldsymbol{\varepsilon})) \quad (31)$$

$$0 = (\mathbf{y} - \mathbf{g}(\boldsymbol{\varepsilon}))^T \cdot (\mathbf{y} - \mathbf{g}(\boldsymbol{\varepsilon})) - \sigma^2 N + \sigma^3 \frac{d(\ln(P_{\sigma \text{ prior}}(\sigma | \mathbf{X})))}{d\sigma}. \quad (32)$$

Again, if we choose (24) as the prior for  $\sigma$ , then (32) becomes

$$0 = (\mathbf{y} - \mathbf{g}(\boldsymbol{\varepsilon}))^T \cdot (\mathbf{y} - \mathbf{g}(\boldsymbol{\varepsilon})) - \sigma^2 N - \sigma^2 \left(1 + \frac{(\ln \sigma - \mu_{\sigma})}{\eta^2}\right). \quad (33)$$

Now we are faced with solving two non-linear equations, (31) and (33). We cannot easily find  $\sigma$  as before, by bisection for example, and we do not have an explicit expression for  $\boldsymbol{\varepsilon}$  in terms of  $\sigma$ . Solving these equations numerically proves to be difficult and time consuming in practice.

A different way of numerically producing estimates for the unfolding parameters, while avoiding linearization, is to find an estimate for the mean of the joint posterior distribution for  $\boldsymbol{\varepsilon}$  and  $\sigma$  (given by the multiple of (21), (22) and (24), with  $\mathbf{A} \cdot \boldsymbol{\varepsilon}$  replaced by  $\mathbf{g}(\boldsymbol{\varepsilon})$  in (21)). By using a standard MCMC method (see Metropolis *et al*, 1953) one may produce a sample population directly from the joint posterior distribution.

## 14.6 A further example

Twenty noisy data points were generated from a chosen unfolding of the pitchfork bifurcation with  $\boldsymbol{\varepsilon} = \tilde{\boldsymbol{\varepsilon}} = (0.01, 0.01, 0.01, 0.01)^T$ . This data is shown, with the chosen unfolding, in Figure 5. Using all twenty data points, our MCMC method gives  $\boldsymbol{\varepsilon}^* = (0.020297, 0.001013, 0.001642, 0.010814)^T$  as the estimate for  $\boldsymbol{\varepsilon}$ . Figure 6 shows the data together with the pitchfork using our estimated parameters.

The error terms are given by  $\mathbf{e} = \mathbf{y} - \mathbf{g}(\boldsymbol{\varepsilon})$  for a given  $\boldsymbol{\varepsilon}$ . For  $\boldsymbol{\varepsilon} = \tilde{\boldsymbol{\varepsilon}}$ ,  $\mathbf{e}^T \cdot \mathbf{e} = 1.104$ , whereas for  $\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}^*$ ,  $\mathbf{e}^T \cdot \mathbf{e} = 0.875$ , indicating that the unfolding parameters estimated by the MCMC method are actually a better fit for the noisy data.

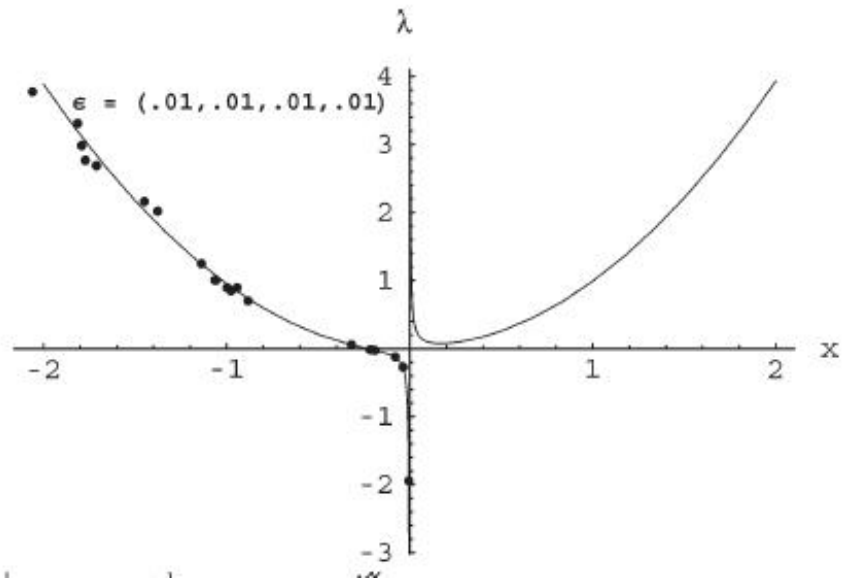


Figure 5: Twenty noisy data points with chosen unfolding

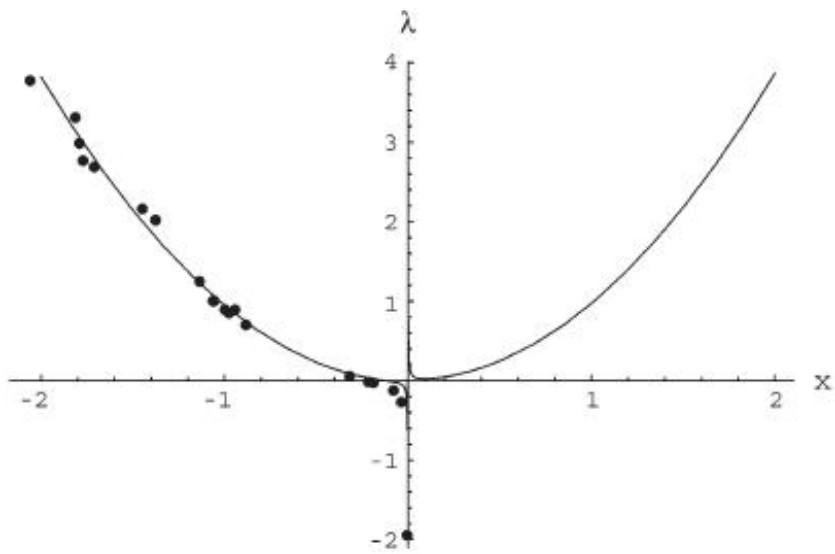


Figure 6: Data with unfolding estimated by MCMC

## 15 The Bayesian brain and Lindenmayer grammars

Problems of perception, learning and discrimination (and consequent decisions, or actions) based on sequentially received sensory inputs fit naturally into an empirical Bayes framework going back to the ideas of Locke. An approach whereby distinct internal “models” of distinct, specific *causes*, or *sources*, are adapted so as to match sensory input was suggested by Mumford [11] and is the basis for feedback or hierarchical schemes of neuronal processing. Recently Friston and Stephan [12] argue that such conditioning of a model to the input data can be thought of as a minimisation of free energy: consequently such evolving models are constantly tuning-in so as to minimise *surprise* (which is, colloquially speaking, the error between similar future sensory inputs and the model’s current expectations of them). There is a clear evolutionary advantage if brains can anticipate more, and be less surprised by their environment. These papers achieve more though, since they suggest how and where such models can evolve within the brain, and they point to hierarchical mechanisms which can lead to such conditioning in the light of fresh sensory input.

Here we take these ideas a step further so as to discuss some simple discrimination problems, rather than the conditioning of a single model in isolation.

In general we wish to consider the situation where sensory input is received from just one of a number of alternative sources, and to investigate whether the empirical Bayes machinery may be implemented, at low computational and storage cost, so as to infer which of the possible source-models is most likely to have generated that sensory input. We wish to consider how multiple models can compete to be inferred as the most likely source of the received input.

Specifically we shall apply these ideas to a problem of discrimination between artificial grammars. These are simple models of structured languages. The discrimination problem is thus “which of the languages that I know, or maybe one I don’t know, am I currently hearing?”. The structure of the input data sequences is both subtle and complex, and we have a wide class of such artificial grammars generated by suitable class of Lindenmeyer grammars.

We will assume that a Bayesian brain has calibrated separate models for separate grammars under circumstances where the source (the particular grammar) in each case was unambiguous. Then we wish to consider whether a Bayesian multiple hypothesis testing approach is fit for the purpose of discriminating between these grammars (and indeed a further “none of the above”, unknown, grammar) and attributing some new input to one, most likely, grammar. At the same time we wish to consider whether the characteristics (the models’ details) can be held at a relatively low resolution, so as to reduce both the computation/cognitive load and the memory requirements, and yet remain effective even in the presence of some errors within the input signals.

## 15.1 Lindenmayer grammars

Consider the following grammars (L systems) acting on  $\{0, 1\}$ , each producing a sequence of strings, starting out from 0:

$$0 \rightarrow w_0 \quad 1 \rightarrow w_1$$

where  $w_0$  is a given string of length  $a$ , containing exactly  $b$  ones and  $(a - b)$  zeros; and  $w_1$  is a string of length  $p$ , containing exactly  $q$  ones and  $(p - q)$  zeros. For the Fibonacci grammar we have  $w_0 = 1$  and  $w_1 = 01$ , so  $(a, b, p, q) = (1, 1, 2, 1)$ ; and re-application of these rules generates the successive strings 0, 1, 01, 101, 01101, 10101101, 0110110101101, and so on.

Let  $\rho_n$  denote the density of ones in the  $n$ th string. Then  $\rho_0=0$ , and

$$\rho_{n+1} = g(\rho_n) \equiv \frac{\rho_n(q - b) + b}{\rho_n(p - a) + a}.$$

In all cases this iteration has a unique and stable fixed point inside  $[0,1]$ . This is given by

$$\rho_\infty = \frac{q - a - b + \sqrt{4b(p - a) + (-a - b + q)^2}}{2(p - a)}$$

So  $\rho_\infty$  is dependent only on the pair  $p - a$  and  $q - a - b$ .

Hence the cases  $(a, b, p, q) = (1, 1, 2, 1), (2, 1, 3, 2), (3, 1, 4, 3)$  are all grammars that produce successive strings with identical asymptotic densities; though at any  $n$ th iteration the density of short substrings may be highly variable and only approximate to  $\rho_\infty$ .

Consider how a receiving party, called a *receiver*, might be able to distinguish between grammars, one from another, on the basis of substring samples; or be able to recognise that any such given sample is likely or unlikely to have been generated by a particular grammar.

The optimal approach would require the solution of the “**inverse problem**”: given a sampled continuous substring, taken from an unknown part of some string, then determine the smallest pair (in total length,  $a + p$ , say)  $(w_0, w_1)$  that could have generated that sample. Or characterise all possible pairs that could have generated that sample.

Any approach to this problem seemingly must involve a tactical search over  $(w_0, w_1)$ -space. We will assume that this task would take an amount of organisation and resources (addressable memory, and/or computational effort/time) as to render it impractical. Such inverse problems are open.

If the receiver makes any errors in collecting samples of (sub)strings then the situation is worse, since some of the more obvious features of the strings deterministically generated by the rules may become violated.

Instead we shall focus on a Bayesian approach both to leaning the idiosyncratic features of grammars, and to resolving recognition and discrimination problems. This form of

reasoning, and its implementation below, is entirely consistent with the recent ideas of Friston et al. Following that approach we assume that any “sensed”, sampled, substring is represented by a set of easily defined and extractable “performance metrics”, or features. Then we represent grammars via suitable distributions over **feature space**. This in turn allows the receiver to have an evolving representation of any such grammars, based on subjective experience to date. Any sampled substrings from a grammar cause an update from the prior to the posterior distributions.

## 15.2 Features

We have seen that a *receiver* monitoring (sampled) substrings, on the basis of density alone (distribution frequency for the underlying alphabet) cannot distinguish between grammars. Though density is an obvious and useful metric for shorter samples, especially at relatively earlier iterates of the corresponding L-system.

Distinct grammars may also possess rather specific autocorrelation structures. Sampling a continuous substring from within a string generated by each grammar we may estimate the lag-correlation (autocorrelation) structure. The Fibonacci grammar shows natural peaks in the autocorrelation whenever the lags are equal to the integers from the standard Fibonacci sequence (which correspond to the lengths of the previous strings, and so therefore there are substrings that must start similarly).

We consider three examples, for each of which  $\rho_\infty = (\sqrt{5} - 1)/2$ :

Example 1, the Fibonacci grammar,  $w_0 = 1$ ,  $w_1 = 01$ ,  $(a, b, p, q) = (1, 1, 2, 1)$ ;

Example 2,  $w_0 = 10$ ,  $w_1 = 011$ ,  $(a, b, p, q) = (2, 1, 3, 2)$ ;

Example 3,  $w_0 = 010$ ,  $w_1 = 1011$ ,  $(a, b, p, q) = (3, 1, 4, 3)$ .

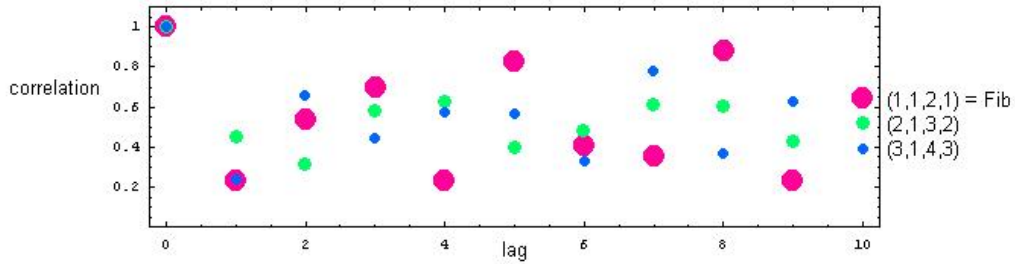


Figure 1. Autocorrelation of string samples from three example grammars, each with density  $(\sqrt{5} - 1)/2$ , and  $(a, b, p, q)$  as given.

It is possible that a simple *receiver* might characterise sampled input sequences by the occurrence or non-occurrence of certain substrings. For example, within any substring generated by the Fibonacci grammar, there can never be two consecutive zeros or three consecutive ones.

The simplest way to usefully adopt such an approach is to restrict attention to consecutive pairs of symbols. Let the transition matrix containing the transition probabilities for successive symbols, which can be estimated from any training, or sampled, sub-string, be as follows.

$$A = \begin{pmatrix} P(\text{next} = 0 | \text{current} = 0) & P(\text{next} = 1 | \text{current} = 0) \\ P(\text{next} = 0 | \text{current} = 1) & P(\text{next} = 1 | \text{current} = 1) \end{pmatrix}$$

Then for the Example 1 (the Fibonacci grammar with  $w_0 = 1$ ,  $w_1 = 01$ ), Example 2 (with  $w_0 = 10$ ,  $w_1 = 011$ ), and Example 3 (with  $w_0 = 010$ ,  $w_1 = 1011$ ) we estimate

$$A = \begin{pmatrix} 0 & 1 \\ .625 & .375 \end{pmatrix}, \begin{pmatrix} 0.275\dots & .724\dots \\ .446\dots & .553\dots \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ .618\dots & .381\dots \end{pmatrix}$$

respectively. Notice that on this basis Example 3 is extremely similar to the Fibonacci grammar. The rows of  $A$  always sum to unity, so  $A_{1,2}$  and  $A_{2,2}$  could be useful independent features to estimate from given substrings.

These examples taken together form up useful test cases, since if any receiver cannot distinguish between all three, it is likely to be focusing on the density distribution of symbols alone. If any receiver can distinguish Example 2 but not between Examples 1 and 3, it is likely that successive pairs (and particularly the lack of “00”s) within the samples is important.

### 15.3 Grammars represented by distributions over feature space

Let us assume that on receipt of any input string, the receiver extracts a number of features such as those above and collects them in a vector  $\mathbf{y}$  that belongs to some feature space,  $Y$ . For the moment we will be nonspecific: the definition of  $\mathbf{y} \in Y$  might rely

on features such as density, short lag autocorrelation structure, transition probabilities, and so on. The components of  $\mathbf{y}$  may be real, integer, or categorical and the "feature space",  $Y$ , is the set of all possible values of  $\mathbf{y}$ .

Suppose a receiver observes a set of messages (substrings), represented by data  $D_{cal,k} = \{\mathbf{y}_i \in Y\}_{i=1}^Q$ , generated by a certain grammar, called  $G_k$  (the index  $k = 1, \dots, n$  will be useful when we deal with more than one possible grammar simultaneously  $n \geq 2$ ). Let  $H_k$  denote the corresponding hypothesis that these  $\mathbf{y}_i$ 's "have been generated by grammar  $G_k$ ".

As usual we let  $X$  denote our prior information (if any) about the nature substrings and grammars we encounter. We will use  $D_{cal}$  as calibration data with which to model for observations, as suitable a low resolution distribution over  $Y$ , under each  $H_k$ .

Now suppose that the  $Y$  has been partitioned into exactly  $m$  subsets, here called classes,  $C_1, \dots, C_m$ . In theory at least, any  $\mathbf{y} \in Y$  received under  $H_k$  is actually drawn from a density distribution, say  $f_k(\mathbf{y})$ , defined over  $Y$ . Applying the ideas from section 11, we may employ this partition to represent each  $f_k$  as a multinomial.

Set

$$P_{k,r} = P(\mathbf{y} \in C_r | H_k, X) = \int_{C_r} f_k(\mathbf{y}) d\mathbf{y},$$

to denote the actual probability that a random sample  $\mathbf{y}$  lies in  $C_r$ , assuming  $H_k$  (and  $X$ ).

Suppose exactly  $Q_r$  out of the  $Q$  elements of  $D_{cal,k}$  are within class  $C_r$ . Then, using Laplace's Law of Succession, we have the estimates

$$\hat{P}_{k,r} = \frac{Q_r + 1}{Q + m}.$$

for the  $P_{k,r}$ . Note for each  $k$  these sum to unity over the classes,  $r$ ; and that  $P_{k,r} > 0$  even if there are no calibration observations are in  $C_r$ . Similarly  $P_{k,r} < 1$  even if all observations are in  $C_r$ .

Hence given a set of calibration data sets,  $D_{cal,k}$  for  $k = 1, \dots, n$ , we have derived a corresponding set of multinomial models.

Now if  $D = \{\mathbf{y}_i\}_{i=1}^q$  is some new set of features extracted from received substrings having exactly  $q_r$  of the  $q$   $\mathbf{y}_i$ 's within each  $C_r$ , then, assuming independence, we have the estimate

$$P(D | H_k, X) = \prod_{r=1}^m \hat{P}_{k,r}^{q_r}. \quad (34)$$

## 15.4 Bayesian multiple hypothesis testing

Now we are ready to apply the theory from section 6 directly.

Recall that have

$$O(H_k|D, X) = \frac{P(D|H_k, X)}{\sum_{j \neq k} P(D|H_j, X) \frac{P(H_j|X)}{(1-P(H_k|X))}} \cdot O(H_k|X). \quad (35)$$

The terms  $O(H_j|X)$  are the priors, and are subjective (how prevalent is each grammar in such situations?).

If a single substring,  $D = \{\mathbf{y}\}$ , is observed such that, say,  $\mathbf{y} \in C_r$ , then we may use (35) to update the odds  $O(H_k|D, X)$  (for each  $k$ ):

$$O(H_k|D, X) = \frac{\hat{P}_{k,r}}{\sum_{j \neq k} \hat{P}_{j,r} \frac{P(H_j|X)}{(1-P(H_k|X))}} \cdot O(H_k|X), \text{ where } \mathbf{x} \in C_r. \quad (36)$$

Hence are grammars are now each represented by multinomial models over a partitioned feature space, and our evolving information is represented by the odds  $O(H_k|D, X)$  on each hypothesis  $H_k$  (as  $D$  evolves).

## 15.5 Example

Here we consider a recognition/discrimination task using the three example grammars introduced above, corresponding to hypotheses  $H_1$ ,  $H_2$ , and  $H_3$ , respectively, together with a back stop,  $H_4$ , corresponding to “none of these three grammars”.

Sampled substrings were all 10 symbols in length. For each we defined a feature space corresponding to the metrics  $\mathbf{y} = (\rho, A_{1,2}, A_{2,2})$ , as defined above. Hence we chose  $Y$  to be the unit cube. We selected an entirely random partition of  $Y$  into  $m = 20$  classes. For the three given example grammars we calibrated their multinomials using 100 independently sampled substrings in each case, applying Laplace’s law of succession.  $H_4$  corresponds to an empty calibration set (and hence a uniform prior multinomial). The resulting multinomials for  $H_1$  to  $H_4$  are shown in Figure 2. Only a few of the classes are observed to be populated by any of the grammars.



Figure 2. Multinomials, calibrated under H1 to H4.

Using these we introduced a new source, providing successive substrings, again each of 10 symbols, one at a time. The resultant evolution of the log-odds on each hypothesis is shown in Figure 3. In fact the new source was indeed generating strings from Example 3, under  $H_3$ . Note that the evolution is not entirely monotonic: some of the substrings suggest  $H_1$  may be true.

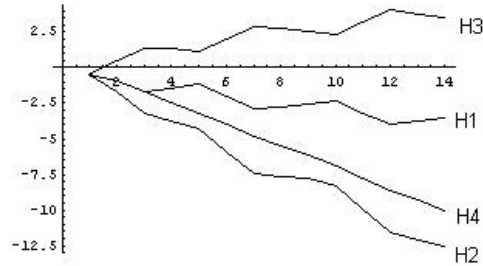


Figure 3.  $\text{Log}_{10}O(H_k|D, X)$  as successive new substrings are received

Next we show the results of four experiments, feeding in new substrings under each separate hypothesis. In the case of  $H_4$  we used a new grammar with  $w_0 = 011$  and  $w_1 = 1011$  to generate the substring samples.

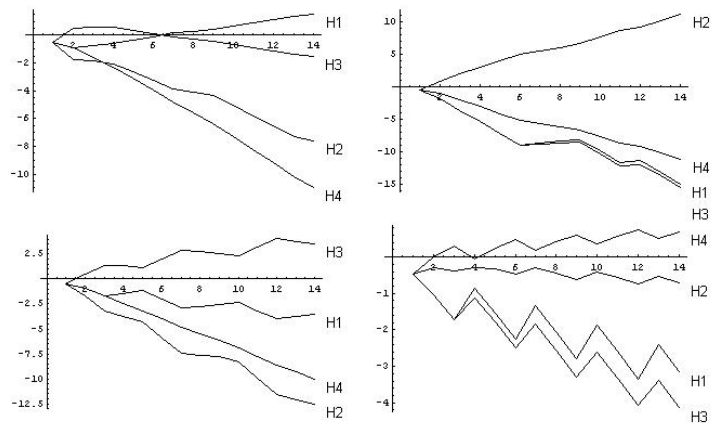


Figure 4.  $\text{Log}_{10}O(H_k|D, X)$ : four examples inputs,  $D$ , under each  $H_j$ .

Note how the receiver “changes its mind” in the first example after five newly sampled substrings (recall  $G_1$  and  $G_3$  are very similar).

## 15.6 Robustness to errors

Since this approach is probabilistic it can be deployed straightforwardly when errors are present. We repeated the experiments, using the four hypotheses calibrated in the last section, with the modification that errors were introduced into the received, unattributed,

samples (generated under  $H_1$ ,  $H_2$  or  $H_3$ ). At a 5% level (with one in twenty symbols altered at random) the most likely hypothesis, after the receipt of 10 substring samples, is incorrect on 6% of all occasions. When the error rate is increased to 10%, then after 10 samples the most likely hypothesis is incorrect on 40% of all occasions.

## Acknowledgements

I would like to acknowledge the huge amount of help that my former colleagues at Numbercraft Ltd and Quintessa Ltd gave me in my work over many years in diverse areas of quantitative assessment inference, and modelling under very different types of uncertainty.

I would like to thank Clive Bowman for his patient advice and many conversations with me about discrimination, Bayes factors, and much, much more.

I would like to thank Doug Saddy for introducing me to L-systems and their usage in generating artificial grammars.

I would like to thank Sam Clarke for his work with me on the unfolding bifurcation explained here.

I would like to acknowledge Selin Hekimoglu and Tom Mullin for supplying the data used in Example 14.4.

Finally I would like to thank my colleagues and friends at the University of Reading for encouraging me to burst into print with this short course for our graduate students.

## References

- [1] Jaynes, E.T. and Bretthorst, G.L. *Probability Theory: The Logic of Science*. Cambridge University Press, Cambridge, 2003
- [2] G. D'Agostini, Bayesian reasoning in data analysis: A critical introduction, World Scientific Publishing, 2003.
- [3] O'Hagan, A. and Forster, J. (004 *Kendall's Advanced Theory Of Statistics. Volume 2B: Bayesian Inference*.
- [4] O. Delrieu O and C.E. Bowman C, Visualizing gene determinants of disease in drug discovery, *Pharmacogenomics* 2006 Apr;7(3):311-29.
- [5] S Clarke, P Grindrod, A Bayesian estimation of unfolded pitchfork bifurcation structure based upon experimental data, *IMA Journal of Applied Mathematics*, 72, 395-404, 2007.
- [6] Chow, S.N. and Hale, J.K. (1982) *Methods Of Bifurcation Theory*. Springer-Verlag, New York.
- [7] Golubitsky, M. and Schaeffer, D.G. (1985) *Singularities and Groups in Bifurcation Theory*. Vol. 1. Springer-Verlag, New York.

- [8] Hekimoglu, S. (2003) Manchester Centre for Non-linear Dynamics, Department of Physics and Astronomy, University of Manchester, Oxford Road, Manchester, M13 9PL, UK. Personal Correspondance.
- [9] Henry, D. (1981) *Geometrical Theory of Semilinear Parabolic Equations*, Lecture Notes in Math. 840. Springer-Verlag, New York.
- [10] Metropolis, N. *et al.* (1953) *Equations of State Calculations by Fast Computing Machines*, Journal of Chemical Physics, 21:1087-1091.
- [11] D. Mumford. On the computational architecture of the neocortex. II The role of cortico-cortical loops. *Biological; Cybernetics*, 66: 241-251, 1992.
- [12] K.J. Friston and K.E. Stephan, Free energy and the brain, *Synthese*, 159: 417-456, 2007.