CURVE-FITTING FOR BAYESIANS?

GORDON BELOT

Abstract. Bayesians often assume, suppose, or conjecture that for any reasonable explication of the notion of simplicity a prior can be designed that will enforce a preference for hypotheses simpler in just that sense. But it is shown here that there are simplicity-driven approaches to curve-fitting problems that cannot be captured within the orthodox Bayesian framework.

1. Introduction

Many philosophers (and others) take a form of no-frills subjective Bayesianism to provide an analysis of rationality: rationality consists in starting life with a probability measure (a *prior*) giving one's credences at birth and in updating this measure by conditionalization to give one's credences at other times, given the evidence that one then possesses.[1] And that is all: on this approach, the relevant probability measure encodes everything there is to say about an agent's credal state—in particular, the relative plausibility for the agent of any two propositions depends only on the probabilities assigned to those propositions by the relevant measure. For convenience, I will call this approach *orthodox Bayesianism.*[2]

This is a very permissive view. Rational agents facing a shared body of evidence may disagree wildly in their doxastic attitudes so long as they began life with suitably differing priors. This liberality is often thought to be one of the strengths of the orthodox view (and its near relatives). A major goal of the Bayesian school is to show that

Forthcoming in BJPS Revised version of July 2016

The originally posted version of this paper featured an appendix concerning the Bayesian automatic Ockham's razor. The argument of the appendix involved a fallacy that was both embarrassing and pointless. The material on the automatic razor is excised here. A corrected version will form a free-standing paper.

[1]Here and throughout, probability measures are countably additive.

[2]The label is tendentious: arguably, the truly orthodox allow merely finitely additive priors (see fn. 15 below). Note that in the hands of most statisticians—and of some philosophers—Bayesian methods are pursued in a pragmatic spirit as tools of statistical inference rather than being taken to provide an account of rationality.

(at their best) scientists behave like Bayesian agents.[3] And here the variety of rational responses to shared evidence is crucial—for example, it underwrites a straightforward way to uphold the rationality of each of several parties to a scientific dispute.[4]

My goal here is to show that there is a sense in which the orthodox view is not as liberal as it is generally taken to be. It is often assumed, supposed, or conjectured that for any reasonable explication of the notion of simplicity a prior can be designed that will enforce a preference for hypotheses simpler in just that sense.[5] In Section 3 it is shown that this is false—there are simplicity-driven approaches to curve-fitting problems that cannot be captured within the orthodox Bayesian framework. In Section 4 I give some grounds for thinking that the most widely-discussed generalizations of the orthodox framework are of no help in getting around this problem—at least if they aim to replace the orthodox approach as an all-purpose account of rationality. The damage is assessed in Section 5. Section 2 lays some preliminary groundwork.

2. A Curve-Fitting Problem

Here is a highly idealized picture of one aspect of the scientific method. One begins with a set of hypotheses, $\mathcal{H}$, concerning the nature of some system. As one gathers data concerning this system, some hypotheses in $\mathcal{H}$ are ruled out by the data. At any stage of inquiry, however, a large number of hypotheses remain in the running. If pressed to select the most plausible one, a scientist will rely on background knowledge, judgements of prior probability, theoretical virtues, favourite statistical tests, and so on.

Elementary discussions of the scientific method often focus on a special case of this general picture: curve-fitting. A scientist is interested in the dependence of physical quantity Y on physical quantity X. Let us call the function F that encodes this dependence the *mystery function.* Data comes in the form of ordered pairs (x, y) consisting of a value x of X and the corresponding value $y = F(x)$ of Y (in the noise-free case) or a value y expected to be close to $F(x)$ (in the case of noisy data). After each data point is revealed, the scientist is expected to make a conjecture: to choose the function in $\mathcal{H}$ that is the most plausible candidate to be the mystery function, given the data seen.

[3]For examples, discussion, and references, see, e.g., (Bovens and Hartmann [2003]; Earman [1992]; Horwich [1982]; and Howson and Urbach [2006]).

[4]See, e.g., (Franklin [1990], §6.1; Howson and Urbach [2006], §8.f; or Salmon [1990]).

[5]For conjectures along these lines, see (Putnam [1979, p. 302; and Howson [2000], p. 206). For some relevant positive results, see (Juhl [1993], [1996]).

Our focus here will be on a special case of curve-fitting. We will assume that the data shown to the curve-fitter are noise-free—so that the curve-fitter should at each stage conjecture a function whose graph passes through the available data points.[6] X and Y will both range over the rational numbers.[7] There are, then, just a countable number of data points of the form $(x, F(x))$. Each will be shown to the curve-fitter at some stage of inquiry. The space of hypotheses $\mathcal{H}$ will be the space of continuous rational-valued functions over the rational numbers.[8]

There are lots of schemes for approaching problems of this sort. Here is a representative one, personified.

> Polly: At each stage of inquiry, choose as your conjecture the lowest-degree polynomial whose graph passes through the data points seen so far (i.e., the unique linear function if there is one; if not, the unique quadratic function if there is one; if not, the unique cubic function if there is one; etc).

This admittedly simple-minded method will be the initial focus of our attention.

3. No Bayesian Polly

Polly's approach to the curve-fitting problem can be thought of as being driven by simplicity considerations: she considers polynomial functions simpler than other functions, considers linear functions simpler than quadratic functions, quadratic functions simpler than cubic functions, and so on.[9] And at each stage of inquiry she conjectures that the mystery function is the simplest hypothesis consistent with the data.

Can there be a Bayesian Polly? That is, is there some probability measure on our $\mathcal{H}$ that, when conditionalized on any finite data set for

[6]This assumption will be relaxed below in Section 5 and in the Appendix.

[7]The use of rational rather than real numbers here allows us to restrict attention to countable sets at certain crucial points—see fnn. 10 and 27 below. The consequences of lifting this assumption will be discussed in Section 5 below.

[8]That is: in order for a function $f : \mathbb{Q} \to \mathbb{Q}$ to be in $\mathcal{H}$, it must be the case that for every rational x and for every $\varepsilon > 0$ there is a $\delta > 0$ such that for every rational x' such that $|x - x'| < \delta$, $|f(x) - f(x')| < \varepsilon$. In fact, just about any reasonable space of rational-valued functions on the rational numbers would do here.

[9]It takes two parameters to specify a linear function of the form $f(x) = a_1x + a_0$ ($a_1 \neq 0$), three parameters to specify a quadratic function of the form $f(x) = a_2x^2 + a_1x + a_0$ ($a_2 \neq 0$), and so on. The intuition that lower-degree polynomials are simpler than high-degree polynomials is widely shared: see, e.g., (Lewis [1994], p. 479; Hempel [1966], §4.4; and Poincaré [1952], p. 50).

our curve-fitting problem always assigns maximal probability to the function that Polly puts forward as her conjecture when shown the same data?[10]

No. Consider the function $Q(x) = x^2$. If shown the right data, Polly is willing to conjecture this function. So the prior of a Bayesian Polly would have to assign Q some positive probability δ. Now suppose that the first two data points that Polly is shown lie on the parabola $y = x^2$. Of course, if shown that data Polly will not advance Q as her conjecture—she will rather put forward the linear function $L(x)$ whose graph goes through those two data points. So a Bayesian Polly's prior would have to assign L a probability no less than δ. But there are infinitely many pairs of points lying on the parabola $y = x^2$ that could be the first two data points that Polly and her would-be Bayesian analogue see, each such pair lying on a distinct line—so a Bayesian Polly's prior would have to assign probabilities no less than δ to each of infinitely many linear functions in $\mathcal{H}$. But that is impossible—there is only a single unit or probability to be split up among all of the polynomial functions.

Notice that the problem at hand is not a picky one or one limited to small data sets.[11] What we have seen is that if a prior assigns positive probability to a quadratic polynomial, then amongst the infinitely many linear polynomials whose graphs intersect that of the given quadratic in two points, only finitely many can be assigned as high prior probability as is the given quadratic. So if a prior counts a given quadratic as a live hypothesis, then it in fact prefers that quadratic to the vast majority of its linear competitors. And, by parallel reasoning, if a prior considers a given cubic as a live hypothesis, then it in fact prefers that cubic to the vast majority of its linear and quadratic rivals, and so

[10]Since $\mathcal{H}$ is uncountable, one expects that a typical prior will assign probability zero to typical hypotheses when conditionalized on typical data sets—so it would require some fancy footwork to make interesting sense of the notion of the hypothesis assigned maximal posterior probability by a given prior relative to a given data set. But in the present context, we can assume that the prior of a Bayesian Polly would assign probability one to the set of polynomial hypotheses. So we can restrict attention to those priors that distribute their unit of probability over $\mathcal{H}$ by assigning positive probabilities to countably many polynomial hypotheses. In this context, it makes perfect sense to ask which polynomial function is assigned highest probability.

[11]Indeed, if our space of hypotheses were the integer-valued functions on the integers, we could show the following: let B be a Bayesian agent whose prior assigns non-zero probability to each polynomial; then for typical functions in the space of hypotheses, there is an order in which the data points could be revealed that would lead to B and Polly disagreeing infinitely often in their conjectures. For the relevant notion of typicality and the technique of proof, see, e.g., (Belot [2013]).

on. So any prior that assigns positive probability to each polynomial prefers each polynomial of degree two or more to the vast majority of its lower-degree rivals—and so fails radically to simulate Polly for data sets of arbitrary size.[12]

4. Prospects for a Generalized Bayesian Polly

Orthodox Bayesianism aims to provide a comprehensive account of rationality: agents are rational if and only if they have credal states representable by probability measures that they update by conditionalization.

The argument above shows that on this account polynomial curve-fitting is irrational. It is natural to ask whether any of the standard generalizations of the orthodox approach provide accounts of rationality compatible with this method of curve-fitting. So far as I can see, the answer is: No. At any rate, each of the five most widely-discussed generalizations either have the feature that they judge Polly to be irrational or they are implausible as accounts of rationality.[13]

4.1. **Imprecise Credences.** One way to generalize the orthodox picture is to use sets of probability measures to represent rational credal states.[14] In effect, under this approach agents can be thought of as being guided by committees of ordinary Bayesian agents. But this is of no obvious help here, since it tends to make it harder rather than easier to simulate Polly's curve-fitting behaviour—a committee of Bayesian agents who disagree with one another about which hypothesis is rendered most plausible by a given body of data will at least sometimes find themselves unable to agree on a conjecture (or will resort to randomization in order to choose conjectures)—which will make it impossible for them to always simulate Polly's behaviour.

4.2. **Merely Finitely Additive Probability Measures.** Let A_1, A_2, ... be a countable family of mutually exclusive propositions. The

[12]And, of course, any prior that fails to assign each polynomial positive probability *also* fails to simulate Polly for data sets of arbitrary size—namely, those consisting of data points lying on the graph of a polynomial that it has ruled out *a priori.*

[13]Readers who disagree with my judgements about plausibility here can read this paper as providing an argument in favour of one or more of the approaches discussed below that are capable of simulating Polly's behaviour. Lexicographic probabilities are not discussed here because in light of the results of (Halpern [2010]) it seems likely (but not, perhaps, certain) that they are subject to the same problems as infinitesimal-valued probability measures.

[14]For discussion and references, see (Joyce [2010]).

probability measures employed on the orthodox Bayesian approach satisfy

> COUNTABLE ADDITIVITY: the probability of a disjunction of some or all of the A_k is equal to the sum of the probabilities of those A_k.

Finitely additive probability measures generalize probability measures by requiring additivity to hold only for finite sets of propositions. It is often maintained that the most defensible version of the Bayesian account of rationality countenances rational agents with credal states representable by merely finitely additive probability measures.[15] What is the advantage of such generalized measures? They allow one to assign non-zero weight to a countable set, even while assigning zero weight to each member of the set (probability measures do the same thing for uncountable sets).

But this extra flexibility is of no help in constructing a generalized Bayesian Polly. Suppose, for example, that a Bayesian Polly were to assign a given linear hypothesis probability zero. Then that agent would have to assign probability zero to each quadratic function whose graph shared two points with that of the given linear function (after all, these two points could be the first two data points seen—and if our Bayesian Polly is to simulate Polly, she can't consider a quadratic function more plausible than a linear function after seeing two data points). And similarly, the agent would have to assign prior probability zero to any cubic function whose graph intersects that of the given linear function in three points. And so on. There is nothing special about our example of a linear polynomial: if you want to simulate Polly and if you assign probability zero to a given polynomial, you must also assign probability zero to many higher-order polynomials whose graphs intersect that of the given polynomial. But now the problem is that if you see data consistent with any of these polynomial hypotheses to which you assign prior probability zero, this data will also be consistent with many higher-order polynomials to which you likewise assign prior probability zero. There would be then be no Bayesian grounds for singling out the lowest-degree polynomial consistent with the data as most plausible given the data seen—it is simply one of many hypotheses assigned probability zero.

[15]See, e.g., (Arntzenius *et al.* [2004]; de Finetti [1972], Chapter 5; Howson and Urbach [2006], §2.d; Kadane *et al.* [1986]; or Savage [1962], §3.4). For a dissenting voice, see (Skyrms [1983], §3.4). For the bearing of the assumption of countable additivity on the problem of skepticism, see (Juhl and Kelly [1994]; Kelly [1996], Chapter 13).

4.3. **Hierarchical Bayesianism.** In hierarchical Bayesian approaches, the space of hypotheses is taken have additional structure beyond that required in the orthodox approach.[16] There are many things that can be done with this additional structure.[17] In our simple setting, a natural way to attempt to construct a hierarchical Bayesian Polly would be to consider an agent whose prior is concentrated on the polynomials with rational coefficients, and who takes these to fall into a hierarchy of families in the obvious way (the linear polynomials, the quadratic polynomials, the cubic polynomials, and so on), and who updates by conditionalization. This agent keeps track of the prior and posterior probability of each family as well of as each individual polynomial. So far so good—but if her algorithm for curve-fitting has her just conjecturing whichever individual polynomial has the highest posterior probability conditional on the data seen, then she will immediately be subject to the objection of Section 3 above. However, a hierarchical Bayesian agent has other options.[18] In our setting, the natural method of curve-fitting for a hierarchical Bayesian agent would be to determine, after each data point is revealed, which family has highest posterior probability, and then to put forward as her conjecture whichever member of that family has highest posterior probability within that family.[19]

But no procedure of this kind can simulate Polly. Consider again by way of illustration $Q(x) = x^2$. Any aspiring hierarchical Bayesian Polly will have to assign Q positive prior probability. Let L be a linear polynomial whose graph intersects that of Q in two points. If those two points are the first data points seen, then Polly will advance L as her conjecture. If our aspiring hierarchical Bayesian Polly is able to follow suit, it can only be because after seeing the first two data points she considers the family of linear polynomials to be more plausible than the family of quadratic polynomials. But since the first two data points have ruled out all linear polynomials other than L, it must then be the case that our agent assigns higher prior probability to L than to Q. But of the infinitely many linear polynomials whose graphs intersect that of Q in two points, only finitely many can be assigned higher prior probability than Q—so in the vast majority of cases, when the first

[16]The following discussion benefitted immeasurably from the remarks of an anonymous referee.

[17]See, e.g., (Gelman *et al.* [2013], Chapter 5; and Henderson *et al.* [2010]).

[18]See (Henderson *et al.* [2010], §2) on the role and treatment of higher-order hypotheses in the hierarchical approach.

[19]A procedure of this kind is discussed (but not endorsed) by (Henderson *et al.* [2010], p. 183).

two data points seen lie on Q, our hierarchical Bayesian will end up advancing Q (or some other higher-order polynomial) rather than L as her conjecture. And likewise, *mutatis mutandis,* for other higher-order polynomials.

4.4. **Primitive conditional probabilities.** At the heart of Bayesianism is the idea that for every prior P and every possible body of evidence E, should agents with prior P come to possess evidence E they ought to be in a credal state represented by $P(\cdot|E)$. What are our agents to do if $P(E) = 0$? The standard definition of conditional probability gives no answer.[20] But there are many cases in which it seems like agents ought to have well-defined credences after conditionalizing on a proposition of probability zero.[21]

This motivates the suggestion that we should be able to help ourselves to certain facts about conditional probabilities in addition to those underwritten by the standard definition.[22] Essentially, in order to represent our credal state prior to seeing any evidence, we need not only an official prior probability measure P, but also, for each B such that $P(B) = 0$, a further measure P_B that is in effect the prior we will switch to in the unexpected event that we should learn B.[23]

This gambit makes it possible to construct generalized Bayesian agents that simulate Polly. Consider a prior that assigns positive weight to each linear polynomial in such a way that the proposition that the mystery function is linear is given probability one; backed up by a measure that assigns positive weight to each quadratic polynomial in such a way that the proposition that the mystery function is quadratic is given probability one, to be used in case the data show that the mystery function is not linear; backed up by a measure that assigns positive weight to each cubic polynomial in such a way that the proposition that the mystery function is cubic is given probability one, to be used in case the data show that the mystery function is not quadratic; and so on. An agent with this system of priors will output the same conjecture as Polly for any multi-point data set. But note that such agents are absolutely mad—willing, before having seen any

[20]Conditional probabilities are normally defined via: $P(A|B) := P(A\&B)/P(B)$, when $P(B) > 0$.

[21]For instance, if A is the proposition that a certain coin toss came up heads while E is the proposition that a spinner came to rest in a certain position, it seems pretty clear that reasonable agents will take $P(A|E) = P(A)$, even though their prior probability for E will be zero.

[22]For discussion and references, see (Hájek [2011]).

[23]See (Halpern [2010], §§2 f.) for a framework in which this can be made precise.

data, to bet their lives against any stake whatsoever on the proposition that the mystery function is linear. So to the extent that the point behind the challenge to simulate Polly was to challenge Bayesians to show that they could accommodate agents with reasonable attitudes towards curve-fitting, this is a rather hollow victory.[24]

4.5. **Infinitesimal-valued probability measures.** There is a big, awkward difference between the way that countable and uncountable sets are treated in the standard framework.[25] There is a sense in which it is possible to spread probability evenly over an uncountable set like the unit interval—but each point must be assigned zero probability, and some sets of points cannot be assigned a probability at all. On the other hand, although it is possible to assign non-zero probability to each subset of a countable set, it is impossible to do so in a way that is spread evenly over the members of that set. The impetus for considering infinitesimal-valued probability measures—gizmos just like probability measures except that they take their values in extensions of the ordinary real number system that include infinitesimal numbers—comes from a desire to overcome this awkwardness: whether one is considering a spinner that might come to rest at any point on a circle or a lottery in which any natural number might be drawn, one can assign non-zero probability to each subset of the space of hypotheses under consideration, in a way that assigns the same infinitesimal probability to each individual hypothesis.

Of course, one can also define infinitesimal-valued measures that assign different weights to different hypotheses. Thus, in our case one might assign the set of polynomial functions probability one; assign each linear function the same infinitesimal probability ε; assign each quadratic function probability $\varepsilon/2$; assign each cubic function probability $\varepsilon/3$; and so on.[26] Such a prior perfectly simulates Polly's curve-fitting behaviour.

[24]The problem at hand is not special to this particular way of simulating Polly. The whole point of introducing primitive conditional probabilities is to allow us to construct agents that know how to respond should they see a data set to which they assigned zero prior probability—but before seeing any data, such agents are of course willing to bet their lives against seeing such a data set.

[25]Concerning each of the points that arise in this paragraph, see (Skyrms [1983]) for discussion and references.

[26]This can be done, e.g., by using an arbitrary enumeration to identify the rational polynomials with the natural numbers, then adapting the machinery of (Benci *et al.* [2013], §5.2).

However, it is far from clear that priors taking infinitesimal values play any role in representing rational agents. Indeed, there are situations in which any agent who assigns infinitesimal probability to each member of a countable set of alternatives is obliged to behave in a deeply irrational manner. Example: such an agent can start out believing that there is a one in six chance that the fair die just rolled came up six and then learn which ticket won a certain lottery—and end up assigning probability within an infinitesimal of one to the proposition that the die came up six, no matter which ticket won, although the lottery is held no matter how the die comes up (see Pruss [2012], pp. 82 f.).

5. How Damaging?

The argument of Section 3 shows that the conjecture made by Putnam and others is false—there *are* notions of simplicity that cannot be captured in the orthodox Bayesian framework. The discussion of Section 4 provides grounds for thinking that none of the most widely-discussed generalizations of the orthodox approach constitutes a viable response.

Fine. But is there an objection here to the orthodox Bayesian approach and its near relatives? This is a delicate question. As noted above, many Bayesians aim to show that (at their best) scientists behave like Bayesian agents. So there is an objection here if Polly's approach to her curve-fitting problem is judged to be interestingly similar to things that happen in scientific practice. There are of course ways of eschewing this judgement. I canvass the three most salient—and hope to leave the reader convinced that there is indeed something for Bayesians to worry about here.

5.1. **Who cares about Polly?** Polly is not a great candidate to be a rational agent—her response to some data sets will strike anyone as bizarre (e.g., no matter what data she sees, she will never conjecture that the true function is $f(x) = |x|$). But Polly's basic strategy is typical of a wide variety of techniques used in scientific contexts. And it is the basic strategy rather than her particular approach that causes the trouble here.

Note that for any method of addressing our curve-fitting problem, the set of functions that method is willing to conjecture is always countable.[27]

[27]A method of curve-fitting can be thought of as a function from the space of possible finite data sets to $\mathcal{H}$ that maps each data set to a function consistent with

So if a method responds to every two-point data set by conjecturing that the mystery function is linear, then essentially the same argument given above in Section 3 shows that that method cannot be simulated by a Bayesian agent. And standard curve-fitting procedures do have this feature.

Or, again, if a method proceeds as Polly's does, by segmenting the family of conjecture-worthy functions into a hierarchy of (nonempty) subfamilies (at least one of which is infinite), then always selecting as its conjecture the unique lowest-ranking function consistent with the data, then that method too will fall under the argument of Section 3. And standard curve-fitting techniques do have this structure.

5.2. **Who cares about *this* curve-fitting problem?** The focus on functions that take rational numbers as arguments and values is unusual (to say the least). But it allows the basic problem to be sharply isolated: a Bayesian Polly would have to consider all polynomial functions as live candidates, while considering every linear function to be more probable than every quadratic function—so standard curve-fitting techniques are inconsistent with Bayesianism when the magnitudes of interest take their values in the rational numbers.

It is natural to worry that the same sort of problem persists in a subtler form if one works with real-valued rather than rational-valued quantities. But even if that should turn out not to be the case, I think that a problem would remain here for Bayesians. Many scientists claim to be agnostic about the fine structure of space, time, and physical magnitudes—e.g., about the question whether they are more faithfully modelled by the rational numbers or by the real numbers. According to Bayesians, such scientists are making a mistake in combining this agnosticism with commitment to standard curve-fitting techniques. For according to the Bayesian analysis of rationality, commitment to standard curve-fitting techniques for ordinary empirical problems rationally forbids certain beliefs about the fine structure of physical magnitudes (such as that they have the structure of the rational numbers). That is, to put it mildly, a surprising consequence of a theory of rationality—and, to my mind, an unwelcome one.

5.3. **Who cares about curve-fitting?** Curve-fitting problems of the kind considered here are highly stylized models of scientific inquiry. But

---

it. Since the set of finite data sets is countable in our setting, the image of the map encoding a given curve-fitting scheme must also be countable.

they reflect fairly accurately the workings of some parts of science—especially those concerned with discovering the structure of individual systems rather than the discovery of laws of nature.[28]

However, one element of the curve-fitting problem discussed above is highly suspicious—the assumption that data are noise-free. One might hope that if this assumption were dropped, the problem would go away. That is not the case. Let us allow some sort of noise in our data—if we attempt to sample the value that the mystery function F takes at x, we often observe a value y that differs somewhat from $F(x)$. Let us model this as follows: there is a probability measure σ defined on the rational numbers such that $\sigma(0) > \sigma(x)$ for $x \neq 0$; when one samples the value of F at x, the probability of getting outcome y is $\sigma(y - F(x))$. This is a natural way to generalize the usual Gaussian distribution of measurement errors to the present context.

There are lots of techniques for polynomial curve-fitting in this sort of setting. The hallmarks of such techniques are that they always conjecture polynomial functions and that in looking for the curve that best fits a given data set, they pit against each other a desire to fit the data accurately (to posit a conjecture that makes the data highly likely) and a desire to posit polynomials of lower degree. Neither consideration is absolute. When shown two data points, such methods will posit the linear function whose graph passes through them. Shown more data, they will stick for a time with linear hypotheses, even though that means imperfect fit with the data. But if shown appropriate data, they will eventually put forward conjectures of arbitrarily high-degree—e.g., for any polynomial, if shown enough data points lying on the graph of that polynomial, they will eventually conjecture it, no matter how high its degree.

It follows from all of this that if a Bayesian agent is to simulate a technique of polynomial curve-fitting for noisy data, then it must assign non-zero prior probability to each polynomial. So let us proceed as usual. Consider the quadratic $Q(x) = x^2$. Suppose that this is assigned prior probability $\delta > 0$. Consider two points that lie on the graph of Q and suppose that these are the first two data points shown to our curve-fitters. Suppose that our Bayesian agent considers some linear polynomial L to be more plausible than Q after seeing this evidence E. That is, suppose that $P(L \mid E) > P(Q \mid E)$. Next, note that

$$\frac{P(L \mid E)}{P(Q \mid E)} = \frac{P(L)}{P(Q)} \cdot \frac{P(E \mid L)}{P(E \mid Q)}.$$

[28]See, e.g., (Parker [1994]).

Now, no hypothesis can make this evidence more likely than Q does (since the data points lie on Q). So the second quotient on the right hand side is no greater than one. So our supposition that $P(L \mid E) > P(Q \mid E)$ implies that $P(L) > P(Q)$. But of course, there can only be finitely many L with this feature—so of the infinitely many linear hypotheses that Polly prefers to Q when she sees the right sort of data, our agent prefers Q to all but finitely many when shown this sort of data. And, of course, this argument can be generalized to apply to higher-order polynomials and larger data sets. So the impossibility of a Bayesian Polly is not an artifact of the assumption of noise-free data.

The bottom line: there is a challenge here for Bayesians who want to uphold the rationality of scientific practice.

Acknowledgements

Early versions of this paper were presented in Ann Arbor, Dubrovnik, Fort Wayne, New Brunswick NJ, and Rochester—thanks to all those present. For helpful comments and discussion, thanks to Frank Arntzenius, Jim Joyce, Laura Ruetsche, and several helpful anonymous referees.

References

Arntzenius, F., Elga, A., and Hawthorne, J. [2004]: 'Bayesianism, Infinite Decisions, and Binding', *Mind,* **113**, pp. 251–83.

Belot, G. [2013]: 'Bayesian Orgulity', *Philosophy of Science,* **80**, pp. 483–503.

Benci, V., Horsten, L., and Wenmackers, S. [2013]: 'Non-Archimedean Probability', *Milan Journal of Mathematics,* **81**, pp. 121–51.

Bovens, L. and Hartmann, S. [2003]: *Bayesian Epistemology,* Oxford: Oxford University Press.

de Finetti, B. [1972]: *Probability, Induction, and Statistics: The Art of Guessing,* New York: Wiley.

Earman, J. [1992]: *Bayes or Bust? A Critical Examination of Bayesian Confirmation Theory,* Cambridge MA: MIT Press.

Franklin, A. [1990]: *Experiment: Right or Wrong?,* Cambridge: Cambridge University Press.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., and Rubin, D. [2013]: *Bayesian Data Analysis* (third edition), Boca Raton: CRC Press.

Hájek, A. [2011]: 'Conditional Probability', in P. Bandyopadhyay and M. Forster (*eds*), *Philosophy of Statistics,* Amsterdam: Elsevier, pp. 99–136.

Halpern, J. [2010]: 'Lexicographic Probability, Conditional Probability, and Nonstandard Probability', *Games and Economic Behavior* **68**, pp. 155–79.

Hempel, C.G. [1966]: *Philosophy of Natural Science,* Upper Saddle River: Prentice–Hall.

Henderson, L., Goodman, N., Tenenbaum, J., and Woodward, J. [2010]: 'The Structure and Dynamics of Scientific Theories: A Hierarchical Bayesian Perspective', *Philosophy of Science,* **77**, pp. 172–200.

Horwich, P. [1982]: *Probability and Evidence,* Cambridge: Cambridge University Press.

Howson, C. [2000]: *Hume's Problem: Induction and the Justification of Belief,* Oxford: Oxford University Press.

Howson, C. and Urbach, P. [2006]: *Scientific Reasoning: The Bayesian Approach* (third edition), Chicago: Open Court.

Joyce, J. [2010]: 'A Defense of Imprecise Credences in Inference and Decision Making', *Philosophical Perspectives,* **24**, pp. 281–323.

Juhl, C. [1993]: 'Bayesianism and Reliable Scientific Inquiry', *Philosophy of Science,* **60**, pp. 302–19.

Juhl, C. [1996]: 'Objectively Reliable Subjective Probabilities,' *Synthese,* **109**, pp. 293-309.

Juhl, C. and Kelly, K. [1994]: 'Realism, Convergence, and Additivity', in D. Hull, M. Forbes, and R. Burian (*eds*), *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association* Volume One, Chicago: University of Chicago Press, pp. 181–89.

Kadane, J., Schervish, M. and Seidenfeld, T. [1986]: 'Statistical Implications of Finitely Additive Probability', in P. Goel and A. Zellner (*eds*), *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti,* Amsterdam: North–Holland, pp. 59–76.

Kelly, K. [1996]: *The Logic of Reliable Inquiry,* Oxford: Oxford University Press.

Lewis, D. [1994]: 'Humean Supervenience Debugged', *Mind,* **103**, pp. 473–90.

Parker, R. [1994]: *Geophysical Inverse Theory,* Princeton: Princeton University Press.

Poincaré, H. [1952]: *Science and Hypothesis,* Mineola: Dover.

Putnam, H. [1979]: 'Probability and Confirmation', in H. Putnam, *Mathematics, Matter and Method* (second edition), Cambridge: Cambridge University Press, pp. 293-304.

Pruss, A. [2012]: 'Infinite Lotteries, Perfectly Thin Darts, and Infinitesimals', *Thought,* **1**, pp. 81–9.

Savage, L. [1962]: *Foundations of Statistics,* London: Methuen.

Salmon, W. [1990]: 'The Appraisal of Theories: Kuhn Meets Bayes', in A. Fine, M. Forbes, and L. Wessels (*eds*), *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association.* Volume Two, Chicago: University of Chicago Press, pp. 325–32.

Skyrms, B. [1983]: 'Zeno's Paradox of Measure,' in R. Cohen and L. Laudan (*eds*), *Physics, Philosophy, and Psychoanalysis: Essays in Honor of Adolf Grünbaum,* Dordrecht: Reidel, pp. 233–54.

Department of Philosophy, University of Michigan, belot@umich.edu