AN AUTOMATIC OCKHAM'S RAZOR FOR BAYESIANS?

GORDON BELOT

Abstract. It is sometimes claimed that the Bayesian framework automatically implements Ockham's razor—that conditionalizing on data consistent with both a simple theory and a complex theory more or less inevitably favours the simpler theory. It is shown here that the automatic razor doesn't in fact cut it for certain mundane curve-fitting problems.

1. Introduction

It is sometimes alleged that, across an array of interesting cases, the Bayesian framework automatically implements Ockham's razor: conditionalizing on data accounted for equally well by both a simple theory and a complex theory more or less inevitably favours the simpler theory.[1]

Roughly speaking, the idea is as follows. Suppose that we are able to account for the data seen so far using members of a smaller family of hypotheses (with fewer adjustable parameters) as well as members of a larger family of hypotheses (with more adjustable parameters). Within the smaller family we expect that the live hypotheses are fairly similar to one another compared to how similar the live hypotheses are to one another in the larger family—that is just an expected byproduct of the difference in the number of adjustable parameters. But this is to say that the smaller family in effect makes sharper predictions about what future data will look like than does the larger family. So if new data bear out the predictions of both families, the posterior probability of the smaller family should be boosted more dramatically than the posterior probability of the larger family. If the two families started out with even roughly equal prior probability, the smaller family will soon pull ahead—and stay there so long as it is capable of accounting for the data decently well. Something along these lines is indeed true in certain special cases—such as when each family of hypotheses is finite, or when the smaller family contains only a single hypothesis.[2]

To make the point vivid, consider the case of curve-fitting. Suppose that we are shown three data points that happen to be collinear. The idea is that this sort of data set ought to favour the theory that the true curve is linear at the expense of the theory that the true

Forthcoming in *Erkenntnis*.

[1]See, e.g., Rosenkrantz (1983, p. 82), Jefferys and Berger (1992), McKay (2003, ch. 28), White (2005), and Henderson *et al.* (2010, §4). It will be assumed throughout that Bayesian priors are probability measures—and in particular that they are both normalized and countably additive.

[2]For these cases, see, e.g., Henderson *et al.* (2010, §4) and Kelly and Glymour (2004, §4.4). For claims that the automatic razor should function beyond these special cases, see, e.g., Rosenkrantz (1983, p. 82) and White (2005, §3).

curve is, say, a cubic.[3] For consider the situation between the revelation of the second and third data points. The theory that the true curve is linear is essentially betting its life on the third data point being more or less collinear with the first two, while the theory that the true curve is a cubic is at best agnostic on this question. When the third data point is revealed to be in truth collinear with the first two, Bayesian conditionalization rewards the boldness of the linear theory by boosting its posterior probability at the expense of theories, like the cubic theory, that did not stick their necks out.

The aim of the present note is to show that this plausible-sounding line of reasoning is mistaken. Although the automatic razor functions well when everything in sight is finite, it is easy to construct a curve-fitting problem in which the range from which possible data points are sampled is infinite and in which conditionalization does *not* exhibit a systematic tendency to favour smaller families of hypotheses over larger ones. In particular, for problems of this kind, there is a sense in which typical data sets consisting of three collinear points confirm the theory that the true curve is a cubic at the expense of the theory that it is linear.

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 B on physical quantity A. Let us call the function F that encodes this dependence the *mystery function.* Data come in the form of ordered pairs (x, y) consisting of a value x of A and the corresponding value y of B expected to be close to $F(x)$. After each data point is revealed, the scientist is required 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.

We will specialize here to the case in which x and y range over the rational numbers and the space of hypotheses $\mathcal{H}$ under consideration is the space of polynomial functions in x with rational coefficients.[4] Note that we do not restrict x, y, or the coefficients of polynomials to bounded intervals of the rationals. We will assume that there is some fixed probability measure σ defined on the rational numbers that determines the data seen as follows: if the mystery function is F and the value of F is sampled at x, then the probability of seeing (x, y) is $\sigma(F(x) - y)$. We will make only one assumption about the form of σ: it takes its maximum value $\bar{\sigma}$ at zero (so although it may not be likely that one

[3]Here and throughout, these theories are to be understood as being incompatible: a polynomial of degree k is required to have a non-zero coefficient for x^k.

[4]The basic phenomenon that drives the argument of §3 below arises whether we work with real or rational variables and polynomial coefficients: whatever probability a prior assigns to the linear polynomials, it assigns almost all of this probability to some bounded subset of the space of linear polynomials, and hence all but rules out linear polynomials of relatively large slope or with relatively large intercepts (thanks to an anonymous referee for this way of putting the point). The restriction to rational variables and coefficients allows the consequences of this phenomenon to be brought out in an especially stark fashion.

will see the true value $F(x)$ when sampling at x, it is more likely that one will see this value than that one will see any other given value).

We will consider a Bayesian agent who has a prior probability distribution, Pr, defined over the space of hypotheses $\mathcal{H}$—for any hypothesis h in $\mathcal{H}$, $Pr(h)$ measures our agent's credence, prior to seeing any evidence, that the mystery function is h. For convenience, we will count a prior as admissible only if it assigns positive weight to each hypothesis in $\mathcal{H}$. We work in a context (such as gravitational wave astronomy) in which Nature chooses the order in which the values of x are sampled (and we will assume that no value is ever sampled twice). Our agent has no opinion at all about the order in which the values of x are liable to be sampled—but also does not think that the order in which they are sampled provides any relevant evidence about the identity of the mystery function.

In this setting, the following provides the natural model of our agent's response to evidence. Suppose that the first n values of x sampled are given by $\Delta = \{x_1, x_2, \ldots, x_n\}$. Then a possible data set will have form $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$. We will say that such a D is a data set *based on* Δ. If our agent knows that Δ gives the first values of x to be sampled, then her credences will be encoded in a probability measure Pr_Δ that assigns probabilities to pairs of the form (h, D) where h is a hypothesis in $\mathcal{H}$ and D is a data set based on Δ. $Pr_\Delta(h, D)$ for $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ is calculated in the obvious way:

$$Pr_\Delta(h, D) := Pr(h) \cdot \sigma(h(x_1) - y_1) \cdot \ldots \cdot \sigma(h(x_n) - y_n)$$

(recall that $\sigma(h(x_k) - y_k)$ is the probability of getting value y_k when sampling at x_k if the true value at x_k is $h(x_k)$).

With the joint probability distribution $P_\Delta(h, D)$ in hand, we can go on to define various marginal and conditional probabilities such as $Pr_\Delta(h)$, $Pr_\Delta(D)$, $Pr_\Delta(h|D)$, and $Pr_\Delta(D|h)$ in the usual way. For any Δ and any h in $\mathcal{H}$, $Pr_\Delta(h) = Pr(h)$—so our agent does indeed consider the order in which values of x are sampled to provide no relevant evidence concerning the identity of the mystery function. Informally, we can think of $Pr_\Delta(\cdot)$ as $Pr(\cdot|\Delta)$, but this is merely a heuristic (since Pr does not assign probabilities to propositions like Δ, the conditional probability $Pr(\cdot|\Delta)$ is, strictly speaking, undefined).

3. The Razor Malfunctions

If one wants to understand the extent to which something like the envisioned automatic Bayesian razor really works, it is natural to ask whether conditionalization generically favours the simpler theory over more complex alternatives, for data sets that are accommodated equally well by both.[5]

Let us consider a concrete special case. We use $\mathcal{H}_1$ to denote the set of linear polynomials of the form $\ell(x) = a_1 x + a_0$ ($a_1 \neq 0$) and $\mathcal{H}_3$ to denote the set of of cubic polynomials of the form $c(x) = a_3 x^3 + a_2 x^2 + a_1 x + a_0$ ($a_3 \neq 0$). Consider any Δ consisting of three values of x and any D based on Δ consisting of three collinear data points. We have

$$\frac{Pr_\Delta(\mathcal{H}_3|D)}{Pr_\Delta(\mathcal{H}_1|D)} = \frac{Pr_\Delta(\mathcal{H}_3)}{Pr_\Delta(\mathcal{H}_1)} \cdot \frac{Pr_\Delta(D|\mathcal{H}_3)}{Pr_\Delta(D|\mathcal{H}_1)}.$$

[5]For a related point made in a somewhat different context, see Seidenfeld (1979, pp. 414 f.).

In order for the automatic razor to do its job, the second quotient on the right hand side must be less than one. In that case, Pr_Δ views D as confirming the theory that the mystery function is linear relative to the theory that it is a cubic. Ideally, one would like to show that, for suitably plausible priors, every three-point collinear data set D favoured $\mathcal{H}_1$ over $\mathcal{H}_3$, in the sense that $Pr_\Delta(D|\mathcal{H}_3) < Pr_\Delta(D|\mathcal{H}_1)$—so that the only way that Pr_Δ could assign higher posterior probability to $\mathcal{H}_3$ than to $\mathcal{H}_1$ is if Pr (and hence also Pr_Δ) assigned higher prior probability to $\mathcal{H}_3$ than to $\mathcal{H}_1$.[6] More realistically, one might hope that all but finitely many of the countably infinitely many possible D under consideration had this feature.

We will show, however, that for *any* admissible prior Pr, there are infinitely many data sets D consisting of three collinear points such that $Pr_\Delta(\mathcal{H}_3|D) > Pr_\Delta(\mathcal{H}_1|D)$ (for the Δ on which D is based).

> Claim: Let Pr be an admissible prior and let c_0 be a cubic polynomial. Then there is an $r > 0$ (depending only on Pr and c_0) such that if Δ is a set of three values of x at least one of which has absolute value greater than r, and D is any data set based on Δ consisting of three collinear points lying on c_0, then $Pr_\Delta(\mathcal{H}_3|D) > Pr_\Delta(\mathcal{H}_1|D)$.

In short: we claim that for any prior Pr and any cubic c_0, there is a sense in which typical data sets consisting of three collinear points lying on c_0 render $\mathcal{H}_3$ more probable than $\mathcal{H}_1$ by Pr's lights. For if the x-axis carries its usual metric structure, then no matter how large r is, the interval $J := [-r, r]$ is finite in extent while its complement is infinite in extent—so only very special data sets result from sampling only within J.[7]

The Claim above is easily established. Let Pr, c_0, Δ, and D be as in the Claim. As emphasized above, our agent considers the values at which x is sampled to be irrelevant—so $Pr_\Delta(c_0) = Pr(c_0)$ and $Pr_\Delta(\mathcal{H}_1) = Pr(\mathcal{H}_1)$. Further, since D consists of three collinear points lying on c_0, $Pr_\Delta(D|c_0)$ is just $\bar{\sigma}^3$ (where $\bar{\sigma}$ is the probability of finding the true value of the mystery function when sampling at any value of x). So if we define

$$\varepsilon := \frac{Pr_\Delta(c_0) \cdot Pr_\Delta(D|c_0)}{Pr_\Delta(\mathcal{H}_1)},$$

then ε depends on Pr and c_0 but not on a D or Δ. We then have:

$$\begin{aligned} \frac{\varepsilon}{Pr_\Delta(D|\mathcal{H}_1)} &= \frac{Pr_\Delta(c_0)}{Pr_\Delta(\mathcal{H}_1)} \cdot \frac{Pr_\Delta(D|c_0)}{Pr_\Delta(D|\mathcal{H}_1)} \\ &= \frac{Pr_\Delta(c_0|D)}{Pr_\Delta(\mathcal{H}_1|D)} \\ &\leq \frac{Pr_\Delta(\mathcal{H}_3|D)}{Pr_\Delta(\mathcal{H}_1|D)}. \end{aligned}$$

So in order to show that $Pr_\Delta(\mathcal{H}_1|D) < Pr_\Delta(\mathcal{H}_3|D)$ it suffices to show that $Pr_\Delta(D|\mathcal{H}_1) < \varepsilon$.

[6] For a claim that something along these lines does in fact hold in contexts like ours, see Rosenkrantz (1983, p. 82).

[7] Bayesians might rather rely on a Pr-relative notion of typicality of data sets at this point. But such a notion is not easy to come by in our context, since Pr doesn't assign probabilities to the proposition that the first data points are given by D or that the first values of x sampled are given by Δ.

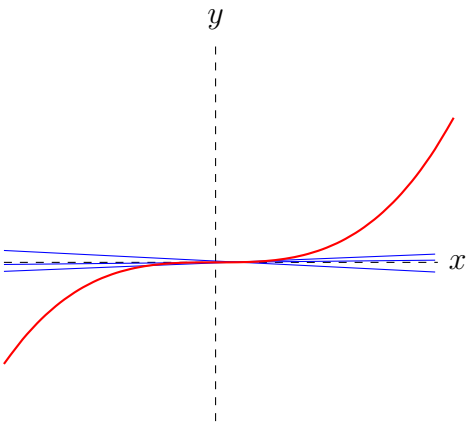

FIGURE 1. Up to a choice of units for the axes, every graph of the cool kids and a cubic looks like this: the y intercept of each curve is indistinguishable from zero; and far from the origin, the cubic soars far above/below the cool kids.

That is not difficult (given a suitable assumption about Δ). Here is the idea. We break the linear polynomials making up $\mathcal{H}_1$ into two groups: a finite set (the *cool kids*) of linear polynomials that collectively eat up almost all of $Pr_\Delta(\mathcal{H}_1)$, and the remaining infinite set of linear polynomials (the *uncool kids*). Since there are only finitely many of them, if we go out far enough towards $\pm\infty$ along the x-axis the graph of c_0 will be far above or below the graphs of all of the cool kids (see Figure 1). So if our data sets involve sampling at sufficiently large values of x, the chance of getting any data points that lie on c_0 if the data points are being generated by one of the cool kids is as small as we like. And of course the remaining uncool kids are collectively so unlikely that the chance of getting data points lying near one of them is also ignorably small. So $Pr_\Delta(D|\mathcal{H}_1) < \varepsilon$ as desired.

Here are the details. Enumerate the linear polynomials in decreasing order of probability conditional on $\mathcal{H}_1$: $\ell_1, \ell_2, \ldots$. Choose N large enough so that $\sum_{i=1}^{N} Pr_\Delta(\ell_i|\mathcal{H}_1) > 1 - \frac{1}{2}\varepsilon$. As a consequence we have:

$$\begin{aligned} \sum_{i=N+1}^{\infty} Pr_\Delta(D|\ell_i) Pr_\Delta(\ell_i|\mathcal{H}_1) &\leq \sum_{i=N+1}^{\infty} Pr_\Delta(\ell_i|\mathcal{H}_1) \\ &< \frac{\varepsilon}{2} \end{aligned}$$

(in the first line we use the fact that each $Pr_\Delta(D|\ell_i) \leq 1$; in the second, our choice of N above).

Next, notice that because $c_0(x) \to \pm\infty$ as x^3 while the $\ell_i(x) \to \pm\infty$ as x, the graph of c_0 is arbitrarily far above or below the graphs of each of $\ell_1, \ldots, \ell_N$ for sufficiently large values of x. So there is an r such that if $|x| > r$, then if the true value of mystery function at x is given by $\ell_i(x)$ $(i = 1, 2, \ldots, N)$, then the probability of getting a point on c_0 if sampling at x is less than $\frac{1}{2N}\varepsilon$ (σ is a probability measure on the rationals, so $\sigma(y) \to 0$

as $y \to \pm\infty$). So if at least one of the data points in D satisfies $|x| > r$, then we have:

$$\begin{aligned}\sum_{i=1}^{N} Pr_\Delta(D|\ell_i) Pr_\Delta(\ell_i|\mathcal{H}_1) &\leq \sum_{i=1}^{N} Pr_\Delta(D|\ell_i) \\ &< \sum_{i=1}^{N} \frac{\varepsilon}{2N} \\ &= \frac{\varepsilon}{2}\end{aligned}$$

(in the first line, we use the fact that each $Pr_\Delta(\ell_i|\mathcal{H}_1) \leq 1$; in the second, our choice of r above).

So if at least one of the data points in our set D of three collinear data points on c_0 satisfies $|x| > r$, then $Pr_\Delta(D|\mathcal{H}_1) < \varepsilon$, as desired:

$$\begin{aligned}Pr_\Delta(D|\mathcal{H}_1) &= \sum_{i=1}^{\infty} Pr_\Delta(D|\ell_i, \mathcal{H}_1) Pr_\Delta(\ell_i|\mathcal{H}_1) \\ &= \sum_{i=1}^{\infty} Pr_\Delta(D|\ell_i) Pr_\Delta(\ell_i|\mathcal{H}_1) \\ &= \sum_{i=1}^{N} Pr_\Delta(D|\ell_i) Pr_\Delta(\ell_i|\mathcal{H}_1) + \sum_{i=N+1}^{\infty} Pr_\Delta(D|\ell_i) Pr_\Delta(\ell_i|\mathcal{H}_1) \\ &< \frac{\varepsilon}{2} + \frac{\varepsilon}{2}\end{aligned}$$

(in the first line we use the law of total probability, in the second the fact that for each i, $Pr_\Delta(D|\ell_i, \mathcal{H}_1) = Pr_\Delta(D|\ell_i)$, the third line is book-keeping, the fourth follows from observations made above).

It will be clear from the method of proof that the assumptions that the data points are precisely collinear and that they lie precisely on the graph of c_0 could have been relaxed—and similarly that instead of cubics and linear polynomials, we could have used mth-order polynomials and kth-order polynomials for any $m > k$.

What, then, was wrong with the intuitive argument for the automatic Bayesian razor? The problem is that while it is true that before seeing the third data point, the theory that the mystery function is linear is betting its life on the third point being at least roughly collinear with the first two, it is also betting its life on a stronger proposition—that the three data points will at least roughly lie on the graph of one of the handful of linear functions that eat up almost all of the available probability. And losing a single wager in which you have staked your life can spell trouble.

Acknowledgements

This material was presented at the Ninth Workshop in Decisions, Games, and Logic at the University of Michigan and at the Workshop on Probability and Learning at Columbia University. For helpful comments and discussion, thanks to Kenny Easwaran, Jim Joyce, Laura Ruetsche, and three very helpful anonymous referees.

References

Henderson, L., Goodman, N., Tenenbaum, J. & Woodward, J. (2010). The structure and dynamics of scientific theories: A hierarchical Bayesian perspective. Philosophy of Science, 77, 172–200.

Jefferys, W. & Berger, J. (1992). Ockham's razor and Bayesian analysis. American Scientist, 80, 64–72.

Kelly, K. & Glymour, C. (2004). Why Bayesian confirmation does not capture the logic of scientific justification. (In C. Hitchcock (Ed.), Contemporary Debates in Philosophy of Science (pp 94–114). Oxford: Blackwell.)

McKay, D. (2003). Information Theory, Inference, and Learning Algorithms. (Cambridge: Cambridge University Press.)

Rosenkrantz, R. (1983). Why Glymour is a Bayesian. (In J. Earman (Ed.), Testing Scientific Theories (pp. 69–97). Minneapolis: University of Minnesota Press.)

Seidenfeld, T. (1979). Why I am not an objective Bayesian; some reflections prompted by Rosenkrantz. Theory and Decision, 11, 413–440.

White, R. (2005). Why favour simplicity? Analysis, 65, 205–210.

Department of Philosophy, University of Michigan