British Journal of Mathematical and Statistical Psychology (2013), 66, 8–38

The British Psychological Society

www.wileyonlinelibrary.com

Philosophy and the practice of Bayesian statistics

Andrew Gelman[1]* and Cosma Rohilla Shalizi[2]

[1]Department of Statistics and Department of Political Science, Columbia University, New York, USA

[2]Statistics Department, Carnegie Mellon University, Santa Fe Institute, Pittsburgh, USA

A substantial school in the philosophy of science identifies Bayesian inference with inductive inference and even rationality as such, and seems to be strengthened by the rise and practical success of Bayesian statistics. We argue that the most successful forms of Bayesian statistics do not actually support that particular philosophy but rather accord much better with sophisticated forms of hypothetico-deductivism. We examine the actual role played by prior distributions in Bayesian models, and the crucial aspects of model checking and model revision, which fall outside the scope of Bayesian confirmation theory. We draw on the literature on the consistency of Bayesian updating and also on our experience of applied work in social science. Clarity about these matters should benefit not just philosophy of science, but also statistical practice. At best, the inductivist view has encouraged researchers to fit and compare models without checking them; at worst, theorists have actively discouraged practitioners from performing model checking because it does not fit into their framework.

1. The usual story – which we don't like

> In so far as I have a coherent philosophy of statistics, I hope it is 'robust' enough to cope in principle with the whole of statistics, and sufficiently undogmatic not to imply that all those who may think rather differently from me are necessarily stupid. If at times I do seem dogmatic, it is because it is convenient to give my own views as unequivocally as possible. (Bartlett, 1967, p. 458)

Schools of statistical inference are sometimes linked to approaches to the philosophy of science. 'Classical' statistics – as exemplified by Fisher's *p*-values, Neyman–Pearson hypothesis tests, and Neyman's confidence intervals – is associated with the hypothetico-deductive and falsificationist view of science. Scientists devise hypotheses, deduce implications for observations from them, and test those implications. Scientific hypotheses

*Correspondence should be addressed to Andrew Gelman, Department of Statistics and Department of Political Science, 1016 Social Work Bldg, Columbia University, New York, NY 10027 USA (e-mail: gelman@stat.columbia.edu).

DOI:10.1111/j.2044-8317.2011.02037.x

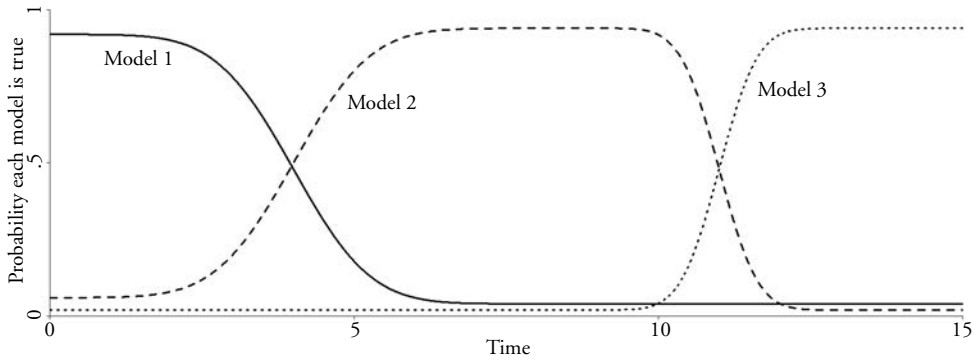

Figure 1. Hypothetical picture of idealized Bayesian inference under the conventional inductive philosophy. The posterior probability of different models changes over time with the expansion of the likelihood as more data are entered into the analysis. Depending on the context of the problem, the time scale on the *x*-axis might be hours, years, or decades, in any case long enough for information to be gathered and analysed that first knocks out hypothesis 1 in favour of hypothesis 2, which in turn is dethroned in favour of the current champion, model 3.

can be rejected (i.e., falsified), but never really established or accepted in the same way. Mayo (1996) presents the leading contemporary statement of this view.

In contrast, Bayesian statistics or 'inverse probability' – starting with a prior distribution, getting data, and moving to the posterior distribution – is associated with an inductive approach of learning about the general from particulars. Rather than employing tests and attempted falsification, learning proceeds more smoothly: an accretion of evidence is summarized by a posterior distribution, and scientific process is associated with the rise and fall in the posterior probabilities of various models; see Figure 1 for a schematic illustration. In this view, the expression $p(\theta|y)$ says it all, and the central goal of Bayesian inference is computing the posterior probabilities of hypotheses. Anything not contained in the posterior distribution $p(\theta|y)$ is simply irrelevant, and it would be irrational (or incoherent) to attempt falsification, unless that somehow shows up in the posterior. The goal is to learn about general laws, as expressed in the probability that one model or another is correct. This view, strongly influenced by Savage (1954), is widespread and influential in the philosophy of science (especially in the form of Bayesian confirmation theory – see Howson & Urbach, 1989; Earman, 1992) and among Bayesian statisticians (Bernardo & Smith, 1994). Many people see support for this view in the rising use of Bayesian methods in applied statistical work over the last few decades.[1]

[1] Consider the current (9 June 2010) state of the Wikipedia article on Bayesian inference, which begins as follows:

> Bayesian inference is statistical inference in which evidence or observations are used to update or to newly infer the probability that a hypothesis may be true.

It then continues:

> Bayesian inference uses aspects of the scientific method, which involves collecting evidence that is meant to be consistent or inconsistent with a given hypothesis. As evidence accumulates, the degree of belief in a hypothesis ought to change. With enough evidence, it should become very high or very low. ...Bayesian inference uses a numerical estimate of the degree of belief in a hypothesis before evidence has been observed and calculates a numerical estimate of the degree of belief in the hypothesis after evidence has been observed. ...Bayesian inference usually relies on degrees of belief, or subjective probabilities, in the induction process and does not necessarily claim to provide an objective method of induction.

We think most of this received view of Bayesian inference is wrong.[2] Bayesian methods are no more inductive than any other mode of statistical inference. Bayesian data analysis is much better understood from a hypothetico-deductive perspective.[3] Implicit in the best Bayesian practice is a stance that has much in common with the error-statistical approach of Mayo (1996), despite the latter's frequentist orientation. Indeed, crucial parts of Bayesian data analysis, such as model checking, can be understood as 'error probes' in Mayo's sense.

We proceed by a combination of examining concrete cases of Bayesian data analysis in empirical social science research, and theoretical results on the consistency and convergence of Bayesian updating. Social-scientific data analysis is especially salient for our purposes because there is general agreement that, in this domain, all models in use are wrong – not merely falsifiable, but actually false. With enough data – and often only a fairly moderate amount – any analyst could reject any model now in use to any desired level of confidence. Model fitting is nonetheless a valuable activity, and indeed the crux of data analysis. To understand why this is so, we need to examine how models are built, fitted, used and checked, and the effects of misspecification on models.

Our perspective is not new; in methods and also in philosophy we follow statisticians such as Box (1980, 1983, 1990), Good and Crook (1974), Good (1983), Morris (1986), Hill (1990) and Jaynes (2003). All these writers emphasized the value of model checking and frequency evaluation as guidelines for Bayesian inference (or, to look at it another way, the value of Bayesian inference as an approach for obtaining statistical methods with good frequency properties; see Rubin, 1984). Despite this literature, and despite the strong thread of model checking in applied statistics, this philosophy of Box and others remains a minority view that is much less popular than the idea of Bayes being used to update the probabilities of different candidate models being true (as can be seen, for example, by the Wikipedia snippets given in footnote 1).

A puzzle then arises. The evidently successful methods of modelling and model checking (associated with Box, Rubin and others) seem out of step with the accepted view of Bayesian inference as inductive reasoning (what we call here 'the usual story'). How can we understand this disjunction? One possibility (perhaps held by the authors of the Wikipedia article) is that the inductive Bayes philosophy is correct and that the model-building approach of Box and others can, with care, be interpreted in that way. Another possibility is that the approach characterized by Bayesian model checking and continuous model expansion could be improved by moving to a fully Bayesian approach centring on the posterior probabilities of competing models. A third possibility, which we advocate, is that Box, Rubin and others are correct and that the usual philosophical story of Bayes as inductive inference is faulty.

Nonetheless, some Bayesian statisticians believe probabilities can have an objective value and therefore Bayesian inference can provide an objective method of induction.

These views differ from those of, for example, Bernardo and Smith (1994) or Howson and Urbach (1989) only in the omission of technical details.

[2] We are claiming that most of the standard philosophy of Bayes is wrong, *not* that most of Bayesian inference itself is wrong. A statistical method can be useful even if its common philosophical justification is in error. It is precisely because we believe in the importance and utility of Bayesian inference that we are interested in clarifying its foundations.

[3] We are not interested in the hypothetico-deductive 'confirmation theory' prominent in philosophy of science from the 1950s to the 1970s, and linked to the name of Hempel (1965). The hypothetico-deductive account of scientific method to which we appeal is distinct from, and much older than, this particular sub-branch of confirmation theory.

We are interested in philosophy and think it is important for statistical practice – if nothing else, we believe that strictures derived from philosophy can inhibit research progress.[4] That said, we are statisticians, not philosophers, and we recognize that our coverage of the philosophical literature will be incomplete. In this presentation, we focus on the classical ideas of Popper and Kuhn, partly because of their influence in the general scientific culture and partly because they represent certain attitudes which we believe are important in understanding the dynamic process of statistical modelling. We also emphasize the work of Mayo (1996) and Mayo and Spanos (2006) because of its relevance to our discussion of model checking. We hope and anticipate that others can expand the links to other modern strands of philosophy of science such as Giere (1988), Haack (1993), Kitcher (1993) and Laudan (1996) which are relevant to the freewheeling world of practical statistics; our goal here is to demonstrate a possible Bayesian philosophy that goes beyond the usual inductivism and can better match Bayesian practice as we know it.

2. The data-analysis cycle

We begin with a very brief reminder of how statistical models are built and used in data analysis, following Gelman, Carlin, Stern, and Rubin (2004), or, from a frequentist perspective, Guttorp (1995).

The statistician begins with a model that stochastically generates all the data y, whose joint distribution is specified as a function of a vector of parameters θ from a space Θ (which may, in the case of some so-called non-parametric models, be infinite-dimensional). This joint distribution is the likelihood function. The stochastic model may involve other (unmeasured but potentially observable) variables $\tilde{y}$ – that is, missing or latent data – and more or less fixed aspects of the data-generating process as covariates. For both Bayesians and frequentists, the joint distribution of $(y, \tilde{y})$ depends on θ. Bayesians insist on a full joint distribution, embracing observables, latent variables and parameters, so that the likelihood function becomes a conditional probability density, $p(y|\theta)$. In designing the stochastic process for $(y, \tilde{y})$, the goal is to represent the systematic relationships between the variables and between the variables and the parameters, and as well as to represent the noisy (contingent, accidental, irreproducible) aspects of the data stochastically. Against the desire for accurate representation one must balance conceptual, mathematical and computational tractability. Some parameters thus have fairly concrete real-world referents, such as the famous (in statistics) survey of the rat population of Baltimore (Brown, Sallow, Davis, & Cochran, 1955). Others, however, will reflect the specification as a mathematical object more than the reality being modelled – t-distributions are sometimes used to model heavy-tailed observational noise, with the number of degrees of freedom for the t representing the shape of the distribution; few statisticians would take this as realistically as the number of rats.

Bayesian modelling, as mentioned, requires a joint distribution for $(y, \tilde{y}, \theta)$, which is conveniently factored (without loss of generality) into a prior distribution for the parameters, $p(\theta)$, and the complete-data likelihood, $p(y, \tilde{y}|\theta)$, so that $p(y|\theta) = \int p(y, \tilde{y}|\theta) d\tilde{y}$. The prior distribution is, as we will see, really part of the model. In practice, the various parts of the model have functional forms picked by a mix of substantive knowledge,

[4] For example, we have more than once encountered Bayesian statisticians who had no interest in assessing the fit of their models to data because they felt that Bayesian models were by definition subjective, and thus neither could nor should be tested.

scientific conjectures, statistical properties, analytical convenience, disciplinary tradition and computational tractability.

Having completed the specification, the Bayesian analyst calculates the posterior distribution $p(\theta|y)$; it is so that this quantity makes sense that the observed y and the parameters θ must have a joint distribution. The rise of Bayesian methods in applications has rested on finding new ways to actually carry through this calculation, even if only approximately, notably by adopting Markov chain Monte Carlo methods, originally developed in statistical physics to evaluate high-dimensional integrals (Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953; Newman & Barkema, 1999), to sample from the posterior distribution. The natural counterpart of this stage for non-Bayesian analyses are various forms of point and interval estimation to identify the set of values of θ that are consistent with the data y.

According to the view sketched in Section 1 above, data analysis basically ends with the calculation of the posterior $p(\theta|y)$. At most, this might be elaborated by partitioning Θ into a set of models or hypotheses, $\Theta_1, \ldots, \Theta_K$, each with a prior probability $p(\Theta_k)$ and its own set of parameters θ_k. One would then compute the posterior parameter distribution within each model, $p(\theta_k|y, \Theta_k)$, and the posterior probabilities of the models,

$$p(\Theta_k|y) = \frac{p(\Theta_k)p(y|\Theta_k)}{\sum_{k'}(p(\Theta_{k'})p(y|\Theta_{k'}))} = \frac{p(\Theta_k)\int p(y, \theta_k|\Theta_k)d\theta_k}{\sum_{k'}(p(\Theta_{k'})\int p(y, \theta_k|\Theta_{k'})d\theta_{k'})}.$$

These posterior probabilities of hypotheses can be used for Bayesian model selection or Bayesian model averaging (topics to which we return below). Scientific progress, in this view, consists of gathering data – perhaps through well-designed experiments, designed to distinguish among interesting competing scientific hypotheses (cf. Atkinson & Donev, 1992; Paninski, 2005) – and then plotting the $p(\Theta_k|y)$ over time and watching the system learn (as sketched in Figure 1).

In our view, the account of the last paragraph is crucially mistaken. The data-analysis process – Bayesian or otherwise – does not end with calculating parameter estimates or posterior distributions. Rather, the model can then be *checked*, by comparing the implications of the fitted model to the empirical evidence. One asks questions such as whether simulations from the fitted model resemble the original data, whether the fitted model is consistent with other data not used in the fitting of the model, and whether variables that the model says are noise ('error terms') in fact display readily-detectable patterns. Discrepancies between the model and data can be used to learn about the ways in which the model is inadequate for the scientific purposes at hand, and thus to motivate expansions and changes to the model (Section 4.).

2.1. Example: Estimating voting patterns in subsets of the population

We demonstrate the hypothetico-deductive Bayesian modelling process with an example from our recent applied research (Gelman, Lee, & Ghitza, 2010). In recent years, American political scientists have been increasingly interested in the connections between politics and income inequality (see, for example, McCarty, Poole, & Rosenthal 2006). In our own contribution to this literature, we estimated the attitudes of rich, middle-income and poor voters in each of the 50 states (Gelman, Park, Shor, Bafumi, & Cortina, 2008). As we described in our paper on the topic (Gelman, Shor, Park, & Bafumi, 2008), we began by fitting a varying-intercept logistic regression: modelling votes (coded as $y = 1$ for votes for the Republican presidential candidate and $y = 0$

for Democratic votes) given family income (coded in five categories from low to high as $x = -2, -1, 0, 1, 2$), using a model of the form $\Pr(y = 1) = \text{logit}^{-1}(a_s + bx)$, where s indexes state of residence – the model is fitted to survey responses – and the varying intercepts a_s correspond to some states being more Republican-leaning than others. Thus, for example, a_s has a positive value in a conservative state such as Utah and a negative value in a liberal state such as California. The coefficient b represents the 'slope' of income, and its positive value indicates that, within any state, richer voters are more likely to vote Republican.

It turned out that this varying-intercept model did not fit our data, as we learned by making graphs of the average survey response and fitted curves for the different income categories within each state. We had to expand to a varying-intercept, varying-slope model, $\Pr(y = 1) = \text{logit}^{-1}(a_s + b_s x)$, in which the slopes b_s varied by state as well. This model expansion led to a corresponding expansion in our understanding: we learned that the gap in voting between rich and poor is much greater in poor states such as Mississippi than in rich states such as Connecticut. Thus, the polarization between rich and poor voters varied in important ways geographically.

We found this not through any process of Bayesian induction but rather through model checking. Bayesian inference was crucial, not for computing the posterior probability that any particular model was true – we never actually did that – but in allowing us to fit rich enough models in the first place that we could study state-to-state variation, incorporating in our analysis relatively small states such as Mississippi and Connecticut that did not have large samples in our survey.[5]

Life continues, though, and so do our statistical struggles. After the 2008 election, we wanted to make similar plots, but this time we found that even our more complicated logistic regression model did not fit the data – especially when we wanted to expand our model to estimate voting patterns for different ethnic groups. Comparison of data to fit led to further model expansions, leading to our current specification, which uses a varying-intercept, varying-slope logistic regression as a baseline but allows for non-linear and even non-monotonic patterns on top of that. Figure 2 shows some of our inferences in map form, while Figure 3 shows one of our diagnostics of data and model fit.

The power of Bayesian inference here is *deductive*: given the data and some model assumptions, it allows us to make lots of inferences, many of which can be checked and potentially falsified. For example, look at New York state (in the bottom row of Figure 3): apparently, voters in the second income category supported John McCain much more than did voters in neighbouring income groups in that state. This pattern is theoretically possible but it arouses suspicion. A careful look at the graph reveals that this is a pattern in the raw data which was moderated but not entirely smoothed away by our model. The natural next step would be to examine data from other surveys. We may have exhausted what we can learn from this particular data set, and Bayesian inference was a key tool in allowing us to do so.

3. The Bayesian principal–agent problem

Before returning to discussions of induction and falsification, we briefly discuss some findings relating to Bayesian inference under misspecified models. The key idea is that

[5] Gelman and Hill (2006) review the hierarchical models that allow such partial pooling.

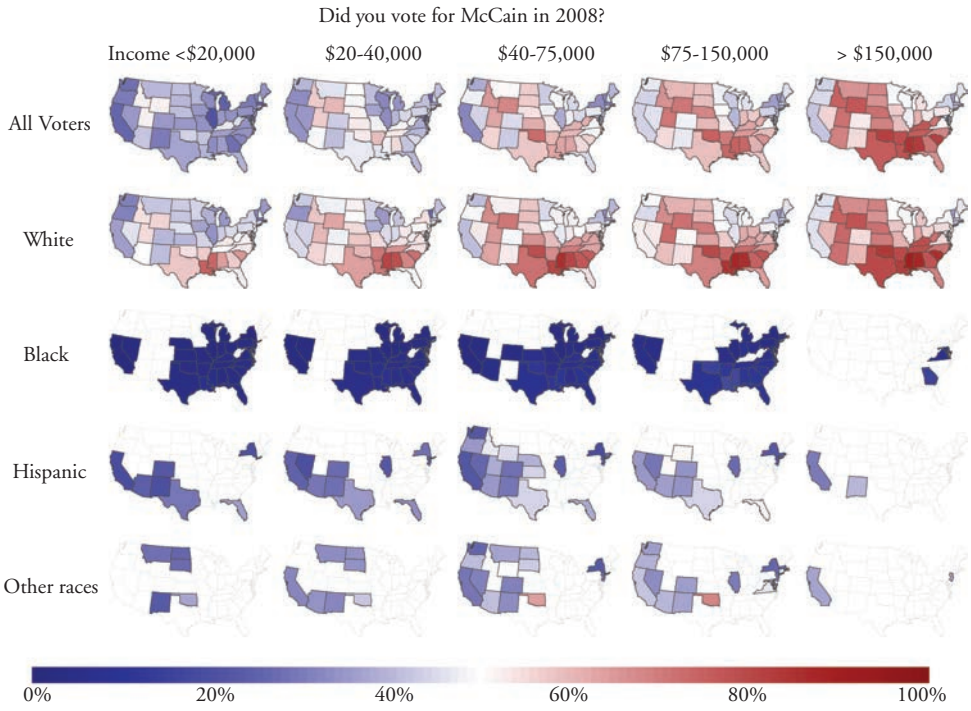

Figure 2. [Colour online]. States won by John McCain and Barack Obama among different ethnic and income categories, based on a model fitted to survey data. States coloured deep red and deep blue indicate clear McCain and Obama wins; pink and light blue represent wins by narrower margins, with a continuous range of shades going to grey for states estimated at exactly 50–50. The estimates shown here represent the culmination of months of effort, in which we fitted increasingly complex models, at each stage checking the fit by comparing to data and then modifying aspects of the prior distribution and likelihood as appropriate. This figure is reproduced from Ghitza and Gelman (2012) with the permission of the authors.

Bayesian inference for model selection – statements about the posterior probabilities of candidate models – does not solve the problem of learning from data about problems with existing models.

In economics, the 'principal–agent problem' refers to the difficulty of designing contracts or institutions which ensure that one selfish actor, the 'agent', will act in the interests of another, the 'principal', who cannot monitor and sanction their agent without cost or error. The problem is one of aligning incentives, so that the agent serves itself by serving the principal (Eggertsson, 1990). There is, as it were, a Bayesian principal–agent problem as well. The Bayesian agent is the methodological fiction (now often approximated in software) of a creature with a prior distribution over a well-defined hypothesis space Θ, a likelihood function $p(y|\theta)$, and conditioning as its sole mechanism of learning and belief revision. The principal is the actual statistician or scientist.

The ideas of the Bayesian agent are much more precise than those of the actual scientist; in particular, the Bayesian (in this formulation, with which we disagree) is

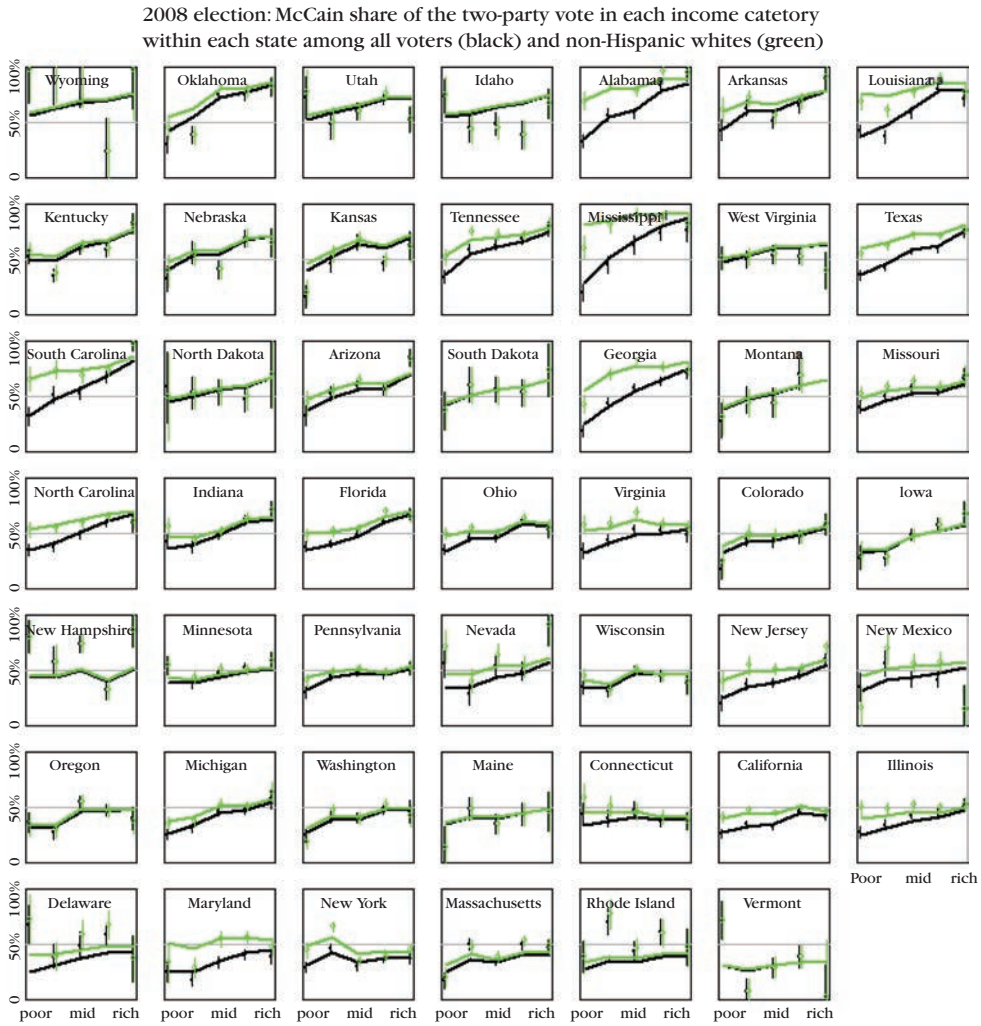

Figure 3. [Colour online]. Some of the data and fitted model used to make the maps shown in Figure 2. Dots are weighted averages from pooled June–November Pew surveys; error bars show ± 1 standard error bounds. Curves are estimated using multilevel models and have a standard error of about 3% at each point. States are ordered in decreasing order of McCain vote (Alaska, Hawaii and the District of Columbia excluded). We fitted a series of models to these data; only this last model fitted the data well enough that we were satisfied. In working with larger data sets and studying more complex questions, we encounter increasing opportunities to check model fit and thus falsify in a way that is helpful for our research goals. This figure is reproduced from Ghitza and Gelman (2012) with the permission of the authors.

certain that *some* θ is the exact and complete truth, whereas the scientist is not.[6] At some point in history, a statistician may well write down a model which he or she

[6] In claiming that 'the Bayesian' is certain that some θ is the exact and complete truth, we are not claiming that actual Bayesian scientists or statisticians hold this view. Rather, we are saying that this is implied by the philosophy we are attacking here. All statisticians, Bayesian and otherwise, recognize that the philosophical position which ignores this approximation is problematic.

believes contains all the systematic influences among properly defined variables for the system of interest, with correct functional forms and distributions of noise terms. This could happen, but we have never seen it, and in social science we have never seen anything that comes close. If nothing else, our own experience suggests that however many different specifications we thought of, there are always others which did not occur to us, but cannot be immediately dismissed *a priori*, if only because they can be seen as alternative approximations to the ones we made. Yet the Bayesian agent is required to start with a prior distribution whose support covers *all* alternatives that could be considered.[7]

This is not a small technical problem to be handled by adding a special value of θ, say θ^{∞} standing for 'none of the above'; even if one could calculate $p(y|\theta^{\infty})$, the likelihood of the data under this catch-all hypothesis, this in general would *not* lead to just a small correction to the posterior, but rather would have substantial effects (Fitelson & Thomason, 2008). Fundamentally, the Bayesian agent is limited by the fact that its beliefs always remain within the support of its prior. For the Bayesian agent the truth must, so to speak, be always already partially believed before it can become known. This point is less than clear in the usual treatments of Bayesian convergence, and so worth some attention.

Classical results (Doob, 1949; Schervish, 1995; Lijoi, Prünster, & Walker, 2007) show that the Bayesian agent's posterior distribution will concentrate on the truth with *prior* probability 1, provided some regularity conditions are met. Without diving into the measure-theoretic technicalities, the conditions amount to: (i) the truth is in the support of the prior; and (ii) the information set is rich enough that some consistent estimator exists (see the discussion in Schervish, 1995, Section 7.4.1). When the truth is *not* in the support of the prior, the Bayesian agent still thinks that Doob's theorem applies and assigns zero prior probability to the set of data under which it does not converge on the truth.

The convergence behaviour of Bayesian updating with a misspecified model can be understood as follows (Berk, 1966, 1970; Kleijn & van der Vaart, 2006; Shalizi, 2009). If the data are actually coming from a distribution q, then the Kullback–Leibler divergence rate, or relative entropy rate, of the parameter value θ is

$$d(\theta) = \lim_{n\to\infty} \frac{1}{n} \mathrm{E}\left[\log \frac{p(y_1, y_2, \ldots, y_n|\theta)}{q(y_1, y_2, \ldots, y_n)}\right],$$

with the expectation being taken under q. (For details on when the limit exists, see Gray, 1990.) Then, under not-too-onerous regularity conditions, one can show (Shalizi, 2009) that

$$p(\theta|y_1, y_2, \ldots, y_n) \approx p(\theta) \exp\left\{-n(d(\theta) - d^*)\right\},$$

with d^* being the essential infimum of the divergence rate. More exactly,

$$-\frac{1}{n} \log p(\theta|y_1, y_2, \ldots, y_n) \to d(\theta) - d^*,$$

[7] It is also not at all clear that Savage and other founders of Bayesian decision theory ever thought that this principle should apply outside of the small worlds of artificially simplified and stylized problems – see Binmore (2007). But as scientists we care about the real, large world.

q-almost surely. Thus the posterior distribution comes to concentrate on the parts of the prior support which have the lowest values of $d(\theta)$ and the highest expected likelihood.[8] There is a geometric sense in which these parts of the parameter space are closest approaches to the truth within the support of the prior (Kass & Vos, 1997), but they may or may not be close to the truth in the sense of giving accurate values for parameters of scientific interest. They may not even be the parameter values which give the best predictions (Grünwald & Langford, 2007; Müller, 2011). In fact, one cannot even guarantee that the posterior will concentrate on a single value of θ at all; if $d(\theta)$ has multiple global minima, the posterior can alternate between (concentrating around) them forever (Berk, 1966).

To sum up, what Bayesian updating does when the model is false (i.e., in reality, always) is to try to concentrate the posterior on the best attainable approximations to the distribution of the data, 'best' being measured by likelihood. But depending on *how* the model is misspecified, and how θ represents the parameters of scientific interest, the impact of misspecification on inferring the latter can range from non-existent to profound.[9] Since we are quite sure our models are wrong, we need to check whether the misspecification is so bad that inferences regarding the scientific parameters are in trouble. It is by this non-Bayesian checking of Bayesian models that we solve our principal–agent problem.

4. Model checking

In our view, a key part of Bayesian data analysis is model checking, which is where there are links to falsificationism. In particular, we emphasize the role of posterior predictive checks, creating simulations and comparing the simulated and actual data. Again, we are following the lead of Box (1980), Rubin (1984) and others, also mixing in a bit of Tukey (1977) in that we generally focus on visual comparisons (Gelman *et al.*, 2004, Chapter 6).

Here is how this works. A Bayesian model gives us a joint distribution for the parameters θ and the observables y. This implies a marginal distribution for the data,

$$p(y) = \int p(y|\theta)p(\theta)d\theta.$$

If we have observed data y, the prior distribution $p(\theta)$ shifts to the posterior distribution $p(\theta|y)$, and so a different distribution of observables,

$$p(y^{\text{rep}}|y) = \int p(y^{\text{rep}}|\theta)p(\theta|y)d\theta,$$

where we use y^{rep} to denote hypothetical alternative or future data, a replicated data set of the same size and shape as the original y, generated under the assumption that

[8] More precisely, regions of Θ where $d(\theta) > d^*$ tend to have exponentially small posterior probability; this statement covers situations such as $d(\theta)$ only approaching its essential infimum as $\|\theta\| \to \infty$. See Shalizi (2009) for details.

[9] White (1994) gives examples of econometric models where the influence of misspecification on the parameters of interest runs through this whole range, though only considering maximum likelihood and maximum quasi-likelihood estimation.

the fitted model, prior and likelihood both, is true. By simulating from the posterior distribution of y^{rep}, we see what typical realizations of the fitted model are like, and in particular whether the observed data set is the kind of thing that the fitted model produces with reasonably high probability.[10]

If we summarize the data with a test statistic $T(y)$, we can perform graphical comparisons with replicated data. In practice, we recommend graphical comparisons (as illustrated by our example above), but for continuity with much of the statistical literature, we focus here on p-values,

$$\Pr\left(T(y^{\text{rep}}) > T(y)|y\right),$$

which can be approximated to arbitrary accuracy as soon as we can simulate y^{rep}. (This is a valid posterior probability in the model, and its interpretation is no more problematic than that of any other probability in a Bayesian model.) In practice, we find graphical test summaries more illuminating than p-values, but in considering ideas of (probabilistic) falsification, it can be helpful to think about numerical test statistics.[11]

Under the usual understanding that T is chosen so that large values indicate poor fits, these p-values work rather like classical ones (Mayo, 1996; Mayo & Cox, 2006) – they are in fact generalizations of classical p-values, merely replacing point estimates of parameters θ with averages over the posterior distribution – and their basic logic is one of falsification. A very low p-value says that it is very improbable, under the model, to get data as extreme along the T-dimension as the actual y; we are seeing something which would be very improbable if the model were true. On the other hand, a high p-value merely indicates that $T(y)$ is an aspect of the data which would be unsurprising if the model is true. Whether this is evidence *for* the usefulness of the model depends how likely it is to get such a high p-value when the model is false: the 'severity' of the test, in the terminology of Mayo (1996) and Mayo and Cox (2006).

Put a little more abstractly, the hypothesized model makes certain probabilistic assumptions, from which other probabilistic implications follow deductively. Simulation works out what those implications are, and tests check whether the data conform to them. Extreme p-values indicate that the data violate regularities implied by the model, or approach doing so. If these were strict violations of deterministic implications, we could just apply *modus tollens* to conclude that the model was wrong; as it is, we nonetheless have evidence and probabilities. Our view of model checking, then, is firmly in the long hypothetico-deductive tradition, running from Popper (1934/1959) back through Bernard (1865/1927) and beyond (Laudan, 1981). A more direct influence on our thinking about these matters is the work of Jaynes (2003), who illustrated how

[10] For notational simplicity, we leave out the possibility of generating new values of the hidden variables $\tilde{y}$ and set aside choices of which parameters to vary and which to hold fixed in the replications; see Gelman, Meng, and Stern (1996).

[11] There is some controversy in the literature about whether posterior predictive checks have too little power to be useful statistical tools (Bayarri & Berger, 2000, 2004), how they might be modified to increase their power (Robins, van der Vaart, & Ventura, 2000; Fraser & Rousseau, 2008), whether some form of empirical prior predictive check might not be better (Bayarri & Castellanos, 2007), etc. This is not the place to rehash this debate over the interpretation or calculation of various Bayesian tail-area probabilities (Gelman, 2007). Rather, the salient fact is that all participants in the debate agree on *why* the tail-area probabilities are relevant: they make it possible to reject a Bayesian model without recourse to a specific alternative. All participants thus *disagree* with the standard inductive view, which reduces inference to the probability that a hypothesis is true, and are simply trying to find the most convenient and informative way to check Bayesian models.

we may learn the most when we find that our model does not fit the data – that is, when it is falsified – because then we have found a problem with our model's assumptions.[12] And the better our probability model encodes our *scientific* or *substantive* assumptions, the more we learn from specific falsification.

In this connection, the prior distribution $p(\theta)$ is one of the assumptions of the model and does not need to represent the statistician's personal degree of belief in alternative parameter values. The prior is connected to the data, and so is potentially testable, via the posterior predictive distribution of future data y^{rep}:

$$p(y^{\text{rep}}|y) = \int p(y^{\text{rep}}|\theta)p(\theta|y)d\theta = \int p(y^{\text{rep}}|\theta)\frac{p(y|\theta)p(\theta)}{\int p(y|\theta')p(\theta')d\theta'}d\theta.$$

The prior distribution thus has implications for the distribution of replicated data, and so can be checked using the type of tests we have described and illustrated above.[13] When it makes sense to think of further data coming from the same source, as in certain kinds of sampling, time-series or longitudinal problems, the prior also has implications for these new data (through the same formula as above, changing the interpretation of y^{rep}), and so becomes testable in a second way. There is thus a connection between the model-checking aspect of Bayesian data analysis and 'prequentialism' (Dawid & Vovk, 1999; Grünwald, 2007), but exploring that would take us too far afield.

One advantage of recognizing that the prior distribution is a testable part of a Bayesian model is that it clarifies the role of the prior in inference, and where it comes from. To reiterate, it is hard to claim that the prior distributions used in applied work represent statisticians' states of knowledge and belief before examining their data, if only because most statisticians do not believe their models are true, so their prior degree of belief in all of Θ is not 1 but 0. The prior distribution is more like a regularization device, akin to the penalization terms added to the sum of squared errors when doing ridge regression and the lasso (Hastie, Tibshirani, & Friedman, 2009) or spline smoothing (Wahba, 1990). All such devices exploit a sensitivity–stability trade-off: they stabilize estimates and predictions by making fitted models less sensitive to certain details of the data. Using an informative prior distribution (even if only weakly informative, as in Gelman, Jakulin, Pittau, & Su, 2008) makes our estimates less sensitive to the data than, say, maximum-likelihood estimates would be, which can be a net gain.

Because we see the prior distribution as a testable part of the Bayesian model, we do not need to follow Jaynes in trying to devise a unique, objectively correct prior distribution for each situation – an enterprise with an uninspiring track record (Kass & Wasserman, 1996), even leaving aside doubts about Jaynes's specific proposal (Seidenfeld, 1979, 1987; Csiszár, 1995; Uffink, 1995, 1996). To put it even more succinctly, 'the model', for a Bayesian, is the combination of the prior distribution and

---

[12] A similar point was expressed by the sociologist and social historian Charles Tilly (2004, p. 597), writing from a very different disciplinary background: 'Most social researchers learn more from being wrong than from being right – provided they then recognize that they were wrong, see why they were wrong, and go on to improve their arguments. Post hoc interpretation of data minimizes the opportunity to recognize contradictions between arguments and evidence, while adoption of formalisms increases that opportunity. Formalisms blindly followed induce blindness. Intelligently adopted, however, they improve vision. Being obliged to spell out the argument, check its logical implications, and examine whether the evidence conforms to the argument promotes both visual acuity and intellectual responsibility.'

[13] Admittedly, the prior only has observable implications in conjunction with the likelihood, but for a Bayesian the reverse is also true.

the likelihood, each of which represents some compromise among scientific knowledge, mathematical convenience and computational tractability.

This gives us a lot of flexibility in modelling. We do not have to worry about making our prior distributions match our subjective beliefs, still less about our model containing all possible truths. Instead we make some assumptions, state them clearly, see what they imply, and check the implications. This applies just much to the prior distribution as it does to the parts of the model showing up in the likelihood function.

4.1. Testing to reveal problems with a model

We are not interested in falsifying our model for its own sake – among other things, having built it ourselves, we know all the shortcuts taken in doing so, and can already be morally certain it is false. With enough data, we can certainly detect departures from the model – this is why, for example, statistical folklore says that the chi-squared statistic is ultimately a measure of sample size (cf. Lindsay & Liu, 2009). As writers such as Giere (1988, Chapter 3) explain, the hypothesis linking mathematical models to empirical data is not that the data-generating process is exactly isomorphic to the model, but that the data source resembles the model closely enough, in the respects which matter to us, that reasoning based on the model will be reliable. Such reliability does not require complete fidelity to the model.

The goal of model checking, then, is not to demonstrate the foregone conclusion of falsity as such, but rather to learn how, in particular, this model fails (Gelman, 2003).[14] When we find such particular failures, they tell us how the model must be improved; when severe tests cannot find them, the inferences we draw about those aspects of the real world from our fitted model become more credible. In designing a *good* test for model checking, we are interested in finding particular errors which, if present, would mess up particular inferences, and devise a test statistic which is sensitive to this sort of misspecification. This process of examining, and ruling out, possible errors or misspecifications is of course very much in line with the 'eliminative induction' advocated by Kitcher (1993, Chapter 7).[15]

All models will have errors of approximation. Statistical models, however, typically assert that their errors of approximation will be unsystematic and patternless – 'noise' (Spanos, 2007). Testing this can be valuable in revising the model. In looking at the red-state/blue-state example, for instance, we concluded that the varying slopes mattered not just because of the magnitudes of departures from the equal-slope assumption, but also because there was a pattern, with richer states tending to have shallower slopes.

What we are advocating, then, is what Cox and Hinkley (1974) call 'pure significance testing', in which certain of the model's implications are compared directly to the data, rather than entering into a contest with some alternative model. This is, we think, more in line with what actually happens in science, where it can become clear that even

[14] In addition, no model is safe from criticism, even if it 'passes' all possible checks. Modern Bayesian models in particular are full of unobserved, latent and unobservable variables, and non-identifiability is an inevitable concern in assessing such models; see, for example, Gustafson (2005), Vansteelandt, Goetghebeur, Kenward, & Molenberghs (2006) and Greenland (2009). We find it somewhat dubious to claim that simply putting a prior distribution on non-identified quantities somehow resolves the problem; the 'bounds' or 'partial identification' approach, pioneered by Manski (2007), seems to be in better accord with scientific norms of explicitly acknowledging uncertainty (see also Vansteelandt *et al.*, 2006; Greenland, 2009).

[15] Despite the name, this is, as Kitcher notes, actually a deductive argument.

large-scale theories are in serious trouble and cannot be accepted unmodified even if there is no alternative available yet. A classical instance is the status of Newtonian physics at the beginning of the twentieth century, where there were enough difficulties – the Michaelson–Morley effect, anomalies in the orbit of Mercury, the photoelectric effect, the black-body paradox, the stability of charged matter, etc. – that it was clear, even before relativity and quantum mechanics, that something would have to give. Even today, our current best theories of fundamental physics, namely general relativity and the standard model of particle physics, an instance of quantum field theory, are universally agreed to be ultimately wrong, not least because they are mutually incompatible, and recognizing this does not require that one have a replacement theory (Weinberg, 1999).

4.2. Connection to non-Bayesian model checking

Many of these ideas about model checking are not unique to Bayesian data analysis and are used more or less explicitly by many communities of practitioners working with complex stochastic models (Ripley, 1988; Guttorp, 1995). The reasoning is the same: a model is a story of how the data could have been generated; the fitted model should therefore be able to generate synthetic data that look like the real data; failures to do so in important ways indicate faults in the model.

For instance, simulation-based model checking is now widely accepted for assessing the goodness of fit of statistical models of social networks (Hunter, Goodreau, & Handcock, 2008). That community was pushed toward predictive model checking by the observation that many model specifications were 'degenerate' in various ways (Handcock, 2003). For example, under certain exponential-family network models, the maximum likelihood estimate gave a distribution over networks which was bimodal, with both modes being very different from observed networks, but located so that the expected value of the sufficient statistics matched observations. It was thus clear that these specifications could not be right even before more adequate specifications were developed (Snijders, Pattison, Robins, & Handcock, 2006).

At a more philosophical level, the idea that a central task of statistical analysis is the search for specific, consequential errors has been forcefully advocated by Mayo (1996), Mayo and Cox (2006), Mayo and Spanos (2004), and Mayo and Spanos (2006). Mayo has placed a special emphasis on the idea of *severe* testing – a model being severely tested if it passes a probe which had a high probability of detecting an error if it is present. (The exact definition of a test's severity is related to, but not quite, that of its power; see Mayo, 1996, or Mayo & Spanos, 2006, for extensive discussions.) Something like this is implicit in discussions about the relative merits of particular posterior predictive checks (which can also be framed in a non-Bayesian manner as graphical hypothesis tests based on the parametric bootstrap).

Our contribution here is to connect this hypothetico-deductive philosophy to Bayesian data analysis, going beyond the evaluation of Bayesian methods based on their frequency properties – as recommended by Rubin (1984) and Wasserman (2006), among others – to emphasize the learning that comes from the discovery of systematic differences between model and data. At the very least, we hope this paper will motivate philosophers of hypothetico-deductive inference to take a more serious look at Bayesian data analysis (as distinct from Bayesian theory) and, conversely, motivate philosophically minded Bayesian statisticians to consider alternatives to the inductive interpretation of Bayesian learning.

4.3. Why not just compare the posterior probabilities of different models?

As mentioned above, the standard view of scientific learning in the Bayesian community is, roughly, that posterior odds of the models under consideration are compared, given the current data.[16] When Bayesian data analysis is understood as simply getting the posterior distribution, it is held that 'pure significance tests have no role to play in the Bayesian framework' (Schervish, 1995, p. 218). The dismissal rests on the idea that the prior distribution can accurately reflect our actual knowledge and beliefs.[17] At the risk of boring the reader by repetition, there is just no way we can ever have any hope of making Θ include all the probability distributions which might be correct, let alone getting $p(\theta|y)$ if we did so, so this is deeply unhelpful advice. The main point where we disagree with many Bayesians is that we do not see Bayesian methods as generally useful for giving the posterior probability that a model is true, or the probability for preferring model A over model B, or whatever.[18] Beyond the philosophical difficulties, there are technical problems with methods that purport to determine the posterior probability of models, most notably that in models with continuous parameters, aspects of the model that have essentially no effect on posterior inferences *within* a model can have huge effects on the comparison of posterior probability *among* models.[19] Bayesian inference is good for deductive inference within a model we prefer to evaluate a model by comparing it to data.

In rehashing the well-known problems with computing Bayesian posterior probabilities of models, we are not claiming that classical p-values are the answer. As is indicated by the literature on the Jeffreys–Lindley paradox (notably Berger & Sellke, 1987), p-values can drastically overstate the evidence against a null hypothesis. From our model-building Bayesian perspective, the purpose of p-values (and model checking more generally) is not to reject a null hypothesis but rather to explore aspects of a model's misfit to data.

In practice, if we are in a setting where model A or model B might be true, we are inclined not to do *model selection* among these specified options, or even to perform *model averaging* over them (perhaps with a statement such as 'we assign 40% of our

---

[16] Some would prefer to compare the modification of those odds called the Bayes factor (Kass & Raftery, 1995). Everything we have to say about posterior odds carries over to Bayes factors with few changes.

[17] As Schervish (1995) continues: 'If the [parameter space Θ] describes all of the probability distributions one is willing to entertain, then one cannot reject [Θ] without rejecting probability models altogether. If one is willing to entertain models not in [Θ], then one needs to take them into account' by enlarging Θ, and computing the posterior distribution over the enlarged space.

[18] There is a vast literature on Bayes factors, model comparison, model averaging, and the evaluation of posterior probabilities of models, and although we believe most of this work to be philosophically unsound (to the extent that it is designed to be a direct vehicle for scientific learning), we recognize that these can be useful techniques. Like all statistical methods, Bayesian and otherwise, these methods are summaries of available information that can be important data-analytic tools. Even if none of a class of models is plausible as truth, and even if we are not comfortable accepting posterior model probabilities as degrees of belief in alternative models, these probabilities can still be useful as tools for prediction and for understanding structure in data, as long as these probabilities are not taken too seriously. See Raftery (1995) for a discussion of the value of posterior model probabilities in social science research and Gelman and Rubin (1995) for a discussion of their limitations, and Claeskens and Hjort (2008) for a general review of model selection. (Some of the work on 'model-selection tests' in econometrics (e.g., Vuong, 1989; Rivers & Vuong, 2002) is exempt from our strictures, as it tries to find which model is *closest* to the data-generating process, while allowing that all of the models may be misspecified, but it would take us too far afield to discuss this work in detail.)

[19] This problem has been called the Jeffreys–Lindley paradox and is the subject of a large literature. Unfortunately (from our perspective) the problem has usually been studied by Bayesians with an eye on 'solving' it – that is, coming up with reasonable definitions that allow the computation of non-degenerate posterior probabilities for continuously parameterized models – but we think that this is really a problem without a solution; see Gelman *et al.* (2004, Section 6.7).

posterior belief to A and 60% to B') but rather to do *continuous model expansion* by forming a larger model that includes both A and B as special cases. For example, Merrill (1994) used electoral and survey data from Norway and Sweden to compare two models of political ideology and voting: the 'proximity model' (in which you prefer the political party that is closest to you in some space of issues and ideology) and the 'directional model' (in which you like the parties that are in the same direction as you in issue space, but with a stronger preference to parties further from the centre). Rather than using the data to pick one model or the other, we would prefer to think of a model in which voters consider both proximity and directionality in forming their preferences (Gelman, 1994).

In the social sciences, it is rare for there to be an underlying theory that can provide meaningful constraints on the functional form of the expected relationships among variables, let alone the distribution of noise terms.[20] Taken to its limit, then, the idea of continuous model expansion counsels social scientists pretty much to give up using parametric statistical models in favour of non-parametric, infinite-dimensional models, advice which the ongoing rapid development of Bayesian non-parametrics (Ghosh & Ramamoorthi, 2003; Hjort, Holmes, Müller, & Walker, 2010) makes increasingly practical. While we are certainly sympathetic to this, and believe a greater use of nonparametric models in empirical research is desirable on its own merits (cf. Li & Racine, 2007), it is worth sounding a few notes of caution.

A technical, but important, point concerns the representation of uncertainty in Bayesian non-parametrics. In finite-dimensional problems, the use of the posterior distribution to represent uncertainty is in part supported by the Bernstein–von Mises phenomenon, which ensures that large-sample credible regions are also confidence regions. This simply fails in infinite-dimensional situations (Cox, 1993; Freedman, 1999), so that a naive use of the posterior distribution becomes unwise.[21] (Since we regard the prior and posterior distributions as regularization devices, this is not especially troublesome for us.) Relatedly, the prior distribution in a Bayesian non-parametric model is a stochastic process, always chosen for tractability (Ghosh & Ramamoorthi, 2003; Hjort *et al.*, 2010), and any pretense of representing an actual inquirer's beliefs abandoned.

Most fundamentally, switching to non-parametric models does not really resolve the issue of needing to make approximations and check their adequacy. All non-parametric models themselves embody assumptions such as conditional independence which are hard to defend except as approximations. Expanding our prior distribution to embrace *all* the models which are actually compatible with our prior knowledge would result in a mess we simply could not work with, nor interpret if we could. This being the case, we feel there is no contradiction between our preference for continuous model expansion and our use of *adequately checked* parametric models.[22]

---

[20] See Manski (2007) for a critique of the econometric practice of making modelling assumptions (such as linearity) with no support in economic theory, simply to get identifiability.

[21] Even in parametric problems, Müller (2011) shows that misspecification can lead credible intervals to have sub-optimal coverage properties – which, however, can be fixed by a modification to their usual calculation.

[22] A different perspective – common in econometrics (e.g., Wooldridge, 2002) and machine learning (e.g., Hastie *et al.*, 2009) – reduces the importance of models of the data source, either by using robust procedures that are valid under departures from modelling assumptions, or by focusing on prediction and external validation. We recognize the theoretical and practical appeal of both these approaches, which can be relevant to Bayesian inference. (For example, Rubin, 1978, justifies random assignment from a Bayesian perspective as a tool for obtaining robust inferences.) But it is not possible to work with *all* possible models when considering

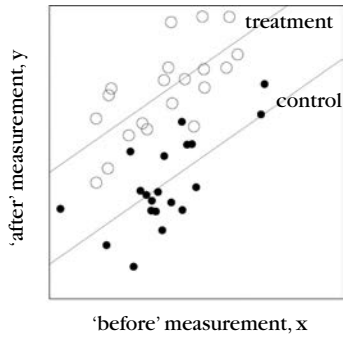

Figure 4. Sketch of the usual statistical model for before-after data. The difference between the fitted lines for the two groups is the estimated treatment effect. The default is to regress the 'after' measurement on the treatment indicator and the 'before' measurement, thus implicitly assuming parallel lines.

4.4. Example: Estimating the effects of legislative redistricting

We use one of our own experiences (Gelman & King, 1994) to illustrate scientific progress through model rejection. We began by fitting a model comparing treated and control units – state legislatures, immediately after redistricting or not – following the usual practice of assuming a constant treatment effect (parallel regression lines in 'before–after' plots, with the treatment effect representing the difference between the lines). In this example, the outcome was a measure of partisan bias, with positive values representing state legislatures where the Democrats were overrepresented (compared to how we estimated the Republicans would have done with comparable vote shares) and negative values in states where the Republicans were overrepresented. A positive treatment effect here would correspond to a redrawing of the district lines that favoured the Democrats.

Figure 4 shows the default model that we (and others) typically use for estimating causal effects in before–after data. We fitted such a no-interaction model in our example too, but then we made some graphs and realized that the model did not fit the data. The line for the control units actually had a much steeper slope than the treated units. We fitted a new model, and it had a completely different story about what the treatment effects meant.

The graph for the new model with interactions is shown in Figure 5. The largest effect of the treatment was not to benefit the Democrats or Republicans (i.e., to change the intercept in the regression, shifting the fitted line up or down) but rather to change the slope of the line, to reduce partisan bias.

Rejecting the constant-treatment-effect model and replacing it with the interaction model was, in retrospect, a crucial step in this research project. This pattern of higher before–after correlation in the control group than in the treated group is

fully probabilistic methods – that is, Bayesian inferences that are summarized by joint posterior distributions rather than point estimates or predictions. This difficulty may well be a motivation for shifting the foundations of statistics away from probability and scientific inference, and towards developing a technology of robust prediction. (Even when prediction is the only goal, with limited data bias–variance considerations can make even misspecified parametric models superior to non-parametric models.) This, however, goes far beyond the scope of the present paper, which aims merely to explicate the implicit philosophy guiding current practice.

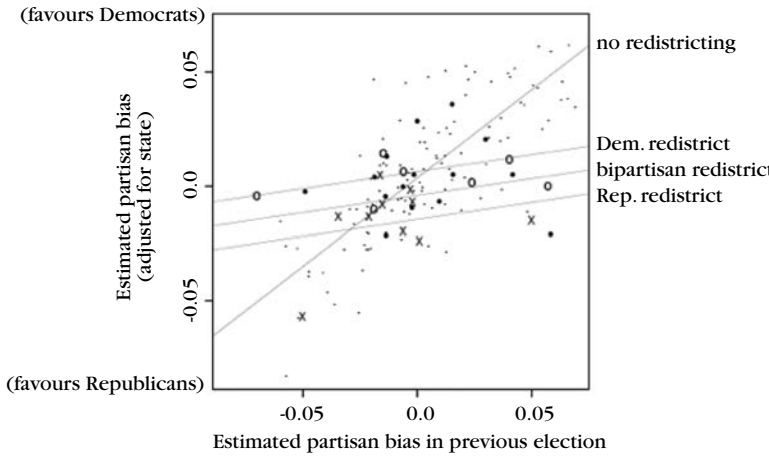

Figure 5. Effect of redistricting on partisan bias. Each symbol represents a state election year, with dots indicating controls (years with no redistricting) and the other symbols corresponding to different types of redistricting. As indicated by the fitted lines, the 'before' value is much more predictive of the 'after' value for the control cases than for the treated (redistricting) cases. The dominant effect of the treatment is to bring the expected value of partisan bias towards zero, and this effect would not be discovered with the usual approach (pictured in Figure 4), which is to fit a model assuming parallel regression lines for treated and control cases. This figure is re-drawn after Gelman and King (1994), with the permission of the authors.

quite general (Gelman, 2004), but at the time we did this study we discovered it only through the graph of model and data, which falsified the original model and motivated us to think of something better. In our experience, falsification is about plots and predictive checks, not about Bayes factors or posterior probabilities of candidate models.

The relevance of this example to the philosophy of statistics is that we began by fitting the usual regression model with no interactions. Only after visually checking the model fit – and thus falsifying it in a useful way without the specification of any alternative – did we take the crucial next step of including an interaction, which changed the whole direction of our research. The shift was induced by a falsification – a bit of deductive inference from the data and the earlier version of our model. In this case the falsification came from a graph rather than a *p*-value, which in one way is just a technical issue, but in a larger sense is important in that the graph revealed not just a lack of fit but also a sense of the direction of the misfit, a refutation that sent us usefully in a direction of substantive model improvement.

5. The question of induction

As we mentioned at the beginning, Bayesian inference is often held to be inductive in a way that classical statistics (following the Fisher or Neyman–Pearson traditions) is not. We need to address this, as we are arguing that all these forms of statistical reasoning are better seen as hypothetico-deductive.

The common core of various conceptions of induction is some form of inference from particulars to the general – in the statistical context, presumably, inference from

the observations y to parameters θ describing the data-generating process. But if *that* were all that was meant, then not only is 'frequentist statistics a theory of inductive inference' (Mayo & Cox, 2006), but the whole range of guess-and-test behaviors engaged in by animals (Holland, Holyoak, Nisbett, & Thagard, 1986), including those formalized in the hypothetico-deductive method, are also inductive. Even the unpromising-sounding procedure, 'pick a model at random and keep it until its accumulated error gets too big, then pick another model completely at random', would qualify (and could work surprisingly well under some circumstances – cf. Ashby, 1960; Foster & Young, 2003). So would utterly irrational procedures ('pick a new random θ when the sum of the least significant digits in y is 13'). Clearly something more is required, or at least implied, by those claiming that Bayesian updating is inductive.

One possibility for that 'something more' is to generalize the truth-preserving property of valid deductive inferences: just as valid deductions from true premises are themselves true, good inductions from true observations should also be true, at least in the limit of increasing evidence.[23] This, however, is just the requirement that our inferential procedures be consistent. As discussed above, using Bayes's rule is not sufficient to ensure consistency, nor is it necessary. In fact, every proof of Bayesian consistency known to us either posits that there is a consistent non-Bayesian procedure for the same problem, or makes other assumptions which entail the existence of such a procedure. In any case, theorems establishing consistency of statistical procedures make *deductively valid* guarantees about these procedures – they are theorems, after all – but do so on the basis of probabilistic assumptions linking future events to past data.

It is also no good to say that what makes Bayesian updating inductive is its conformity to some axiomatization of rationality. If one accepts the Kolmogorov axioms for probability, and the Savage axioms (or something like them) for decision-making,[24] then updating by conditioning follows, and a prior belief state $p(\theta)$ plus data y *deductively* entail that the new belief state is $p(\theta|y)$. In any case, lots of learning procedures can be axiomatized (all those which can be implemented algorithmically, to start with). To pick *this* system, we would need to know that it produces good results (cf. Manski, 2011), and this returns us to previous problems. To know that this axiom system leads us to approach the truth rather than become convinced of falsehoods, for instance, is just the question of consistency again.

Karl Popper, the leading advocate of hypothetico-deductivism in the last century, denied that induction was even possible; his attitude is well paraphrased by Greenland (1998) as: 'we never use any argument based on observed repetition of instances that does not also involve a hypothesis that predicts both those repetitions and the unobserved instances of interest'. This is a recent instantiation of a tradition of anti-inductive arguments that goes back to Hume, but also beyond him to al Ghazali (1100/1997) in the Middle Ages, and indeed to the ancient Sceptics (Kolakowski, 1968). As forcefully put by Stove (1982, 1986), many apparent arguments against this view of induction can be viewed as statements of abstract premises linking both the observed data and unobserved instances – various versions of the 'uniformity of nature' thesis have been popular, sometimes resolved into a set of more detailed postulates, as in

[23] We owe this suggestion to conversation with Kevin Kelly; cf. Kelly (1996, especially Chapter 13).

[24] Despite his ideas on testing, Jaynes (2003) was a prominent and emphatic advocate of the claim that Bayesian inference is the logic of inductive inference as such, but preferred to follow Cox (1946, 1961) rather than Savage. See Halpern (1999) on the formal invalidity of Cox's proofs.

Russell (1948, Part VI, Chapter 9), though Stove rather maliciously crafted a parallel argument for the existence of 'angels, or something very much like them'.[25] As Norton (2003) argues, these highly abstract premises are both dubious and often superfluous for supporting the sort of actual inferences scientists make – 'inductions' are supported not by their matching certain formal criteria (as deductions are), but rather by material facts. To generalize about the melting point of bismuth (to use one of Norton's examples) requires very few samples, provided we accept certain facts about the homogeneity of the physical properties of elemental substances; whether nature in general is uniform is not really at issue.[26]

Simply put, we think the anti-inductivist view is pretty much right, but that statistical models are tools that let us draw inductive inferences on a deductive background. Most directly, random sampling allows us to learn about unsampled people (unobserved balls in an urn, as it were), but such inference, however inductive it may appear, relies not any axiom of induction but rather on deductions from the statistical properties of random samples, and the ability to actually conduct such sampling. The appropriate design depends on many contingent material facts about the system we are studying, exactly as Norton argues.

Some results in statistical learning theory establish that certain procedures are 'probably approximately correct' in what is called a 'distribution-free' manner (Bousquet, Boucheron, & Lugosi, 2004, Vidyasagar 2003); some of these results embrace Bayesian updating (McAllister, 1999). But here 'distribution-free' just means 'holding uniformly over all distributions in a very large class', for example requiring the data to be independent and identically distributed, or from a stationary, mixing stochastic process. Another branch of learning theory does avoid making any probabilistic assumptions, getting results which hold universally across all possible data sets, and again these results apply to Bayesian updating, at least over some parameter spaces (Cesa-Bianchi & Lugosi, 2006). However, these results are all of the form 'in retrospect, the posterior predictive distribution will have predicted almost as well as the best individual model could have done', speaking entirely about performance on the past training data and revealing nothing about extrapolation to hitherto unobserved cases.

To sum up, one is free to describe statistical inference as a theory of inductive logic, but these would be inductions which are deductively guaranteed by the probabilistic assumptions of stochastic models. We can see no interesting and correct sense in which Bayesian statistics is a logic of induction which does not equally imply that frequentist statistics is also a theory of inductive inference (cf. Mayo & Cox, 2006), which is to say, not very inductive at all.

---

[25] Stove (1986) further argues that induction by simple enumeration is reliable *without* making such assumptions, at least sometimes. However, his calculations make no sense unless his data are independent and identically distributed.

[26] Within environments where such premises hold, it may of course be adaptive for organisms to develop inductive propensities, whose scope would be more or less tied to the domain of the relevant material premises. Barkow, Cosmides, and Tooby (1992) develop this theme with reference to the evolution of domain-specific mechanisms of learning and induction; Gigerenzer (2000) and Gigerenzer, Todd, and ABC Research Group (1999) consider proximate mechanisms and ecological aspects, and Holland *et al.* (1986) propose a unified framework for modelling such inductive propensities in terms of generate-and-test processes. All of this, however, is more within the field of psychology than either statistics or philosophy, as (to paraphrase the philosopher Ian Hacking, 2001) it does not so much solve the problem of induction as evade it.

6. What about Popper and Kuhn?

The two most famous modern philosophers of science are undoubtedly Karl Popper (1934/1959) and Thomas Kuhn (1970), and if statisticians (like other non-philosophers) know about philosophy of science at all, it is generally some version of their ideas. It may therefore help readers to see how our ideas relate to theirs. We do not pretend that our sketch fully portrays these figures, let alone the literatures of exegesis and controversy they inspired, or even how the philosophy of science has moved on since 1970.

Popper's key idea was that of 'falsification' or 'conjectures and refutations'. The inspiring example, for Popper, was the replacement of classical physics, after several centuries as the core of the best-established science, by modern physics, especially the replacement of Newtonian gravitation by Einstein's general relativity. Science, for Popper, advances by scientists advancing theories which make strong, wide-ranging predictions capable of being refuted by observations. A good experiment or observational study is one which tests a specific theory (or theories) by confronting their predictions with data in such a way that a match is not automatically assured; good studies are designed with theories in mind, to give them a chance to fail. Theories which conflict with any evidence must be rejected, since a single counter-example implies that a generalization is false. Theories which are not falsifiable by any conceivable evidence are, for Popper, simply not scientific, though they may have other virtues.[27] Even those falsifiable theories which have survived contact with data so far must be regarded as more or less provisional, since no finite amount of data can ever establish a generalization, nor is there any non-circular principle of induction which could let us regard theories which are compatible with lots of evidence as probably true.[28] Since people are fallible, and often obstinate and overly fond of their own ideas, the objectivity of the process which tests conjectures lies not in the emotional detachment and impartiality of individual scientists, but rather in the scientific community being organized in certain ways, with certain institutions, norms and traditions, so that individuals' prejudices more or less wash out (Popper, 1945, Chapters 23–24).

Clearly, we find much here to agree with, especially the general hypothetico-deductive view of scientific method and the anti-inductivist stance. On the other hand, Popper's specific ideas about testing require, at the least, substantial modification. His idea of a test comes down to the rule of deduction which says that if p implies q, and q is false, then p must be false, with the roles of p and q being played by hypotheses and data, respectively. This is plainly inadequate for statistical hypotheses, yet, as critics have noted since Braithwaite (1953) at least, he oddly ignored the theory of statistical hypothesis testing.[29] It is possible to do better, both through standard hypothesis tests and the kind of predictive checks we have described. In particular, as Mayo (1996) has emphasized, it is vital to consider the *severity* of tests, their capacity to detect violations of hypotheses when they are present.

Popper tried to say how science *ought* to work, supplemented by arguments that his ideals could at least be approximated and often had been. Kuhn's work, in contrast,

---

[27] This 'demarcation criterion' has received a lot of criticism, much of it justified. The question of what makes something 'scientific' is fortunately not one we have to answer; cf. Laudan (1996, Chapters 11–12) and Ziman (2000).

[28] Popper tried to work out notions of 'corroboration' and increasing truth content, or 'verisimilitude', to fit with these stances, but these are generally regarded as failures.

[29] We have generally found Popper's ideas on probability and statistics to be of little use and will not discuss them here.

was much more an attempt to describe how science had, in point of historical fact, developed, supported by arguments that alternatives were infeasible, from which some morals might be drawn. His central idea was that of a 'paradigm', a scientific problem and its solution which served as a model or exemplar, so that solutions to other problems could be developed in imitation of it.[30] Paradigms come along with presuppositions about the terms available for describing problems and their solutions, what counts as a valid problem, what counts as a solution, background assumptions which can be taken as a matter of course, etc. Once a scientific community accepts a paradigm and all that goes with it, its members can communicate with one another and get on with the business of solving puzzles, rather than arguing about what they should be doing. Such 'normal science' includes a certain amount of developing and testing of hypotheses but leaves the central presuppositions of the paradigm unquestioned.

During periods of normal science, according to Kuhn, there will always be some 'anomalies' – things within the domain of the paradigm which it currently cannot explain, or which even seem to refute its assumptions. These are generally ignored, or at most regarded as problems which somebody ought to investigate eventually. (Is a special adjustment for odd local circumstances called for? Might there be some clever calculational trick which fixes things? How sound are those anomalous observations?) More formally, Kuhn invokes the 'Quine–Duhem thesis' (Quine, 1961; Duhem, 1914/1954). A paradigm only makes predictions about observations in conjunction with 'auxiliary' hypotheses about specific circumstances, measurement procedures, etc. If the predictions are wrong, Quine and Duhem claimed that one is always free to fix the blame on the auxiliary hypotheses, and preserve belief in the core assumptions of the paradigm 'come what may'.[31] The Quine–Duhem thesis was also used by Lakatos (1978) as part of his 'methodology of scientific research programmes', a falsificationism more historically oriented than Popper's distinguishing between progressive development of auxiliary hypotheses and degenerate research programmes where auxiliaries become *ad hoc* devices for saving core assumptions from data.

According to Kuhn, however, anomalies can accumulate, becoming so serious as to create a crisis for the paradigm, beginning a period of 'revolutionary science'. It is then that a new paradigm can form, one which is generally 'incommensurable' with the old: it makes different presuppositions, takes a different problem and its solution as exemplars, redefines the meaning of terms. Kuhn insisted that scientists who retain the old paradigm are not being irrational, because (by the Quine–Duhem thesis) they can always explain away the anomalies *somehow*; but neither are the scientists who embrace and develop the new paradigm being irrational. Switching to the new paradigm is more like a bistable illusion flipping (the apparent duck becomes an obvious rabbit) than any process of ratiocination governed by sound rules of method.[32]

---

[30] Examples are Newton's deduction of Kepler's laws of planetary motion and other facts of astronomy from the inverse square law of gravitation, and Planck's derivation of the black-body radiation distribution from Boltzmann's statistical mechanics and the quantization of the electromagnetic field. An internal example for statistics might be the way the Neyman–Pearson lemma inspired the search for uniformly most powerful tests in a variety of complicated situations.

[31] This thesis can be attacked from many directions, perhaps the most vulnerable being that one can often find multiple lines of evidence which bear on either the main principles or the auxiliary hypotheses *separately*, thereby localizing the problems (Glymour, 1980; Kitcher, 1993; Laudan, 1996; Mayo, 1996).

[32] Salmon (1990) proposed a connection between Kuhn and Bayesian reasoning, suggesting that the choice between paradigms could be made rationally by using Bayes's rule to compute their posterior probabilities, with the prior probabilities for the paradigms encoding such things as preferences for parsimony. This has

In some way, Kuhn's distinction between normal and revolutionary science is analogous to the distinction between learning within a Bayesian model, and checking the model in preparation to discarding or expanding it. Just as the work of normal science proceeds within the presuppositions of the paradigm, updating a posterior distribution by conditioning on new data takes the assumptions embodied in the prior distribution and the likelihood function as unchallengeable truths. Model checking, on the other hand, corresponds to the identification of anomalies, with a switch to a new model when they become intolerable. Even the problems with translations between paradigms have something of a counterpart in statistical practice; for example, the intercept coefficients in a varying-intercept, constant-slope regression model have a somewhat different meaning than do the intercepts in a varying-slope model. We do not want to push the analogy too far, however, since most model checking and model reformulation would by Kuhn have been regarded as puzzle-solving within a single paradigm, and his views of how people switch between paradigms are, as we just saw, rather different.

Kuhn's ideas about scientific revolutions are famous because they raise so many disturbing questions about the scientific enterprise. For instance, there has been considerable controversy over whether Kuhn believed in any notion of scientific progress, and over whether or not he should have, given his theory. Yet detailed historical case studies (Donovan, Laudan, & Laudan, 1988) have shown that Kuhn's picture of sharp breaks between normal and revolutionary science is hard to sustain.[33] The leads to a tendency, already remarked by Toulmin (1972, pp. 112–117), either to expand paradigms or to shrink them. Expanding paradigms into persistent and all-embracing, because abstract and vague, bodies of ideas lets one preserve the idea of abrupt breaks in thought, but makes them rare and leaves almost everything to puzzle-solving normal science. (In the limit, there has only been one paradigm in astronomy since the Mesopotamians, something like 'many lights in the night sky are objects which are very large but very far away, and they move in interrelated, mathematically describable, discernible patterns'.) This corresponds, we might say, to relentlessly enlarging the support of the prior. The other alternative is to shrink paradigms into increasingly concrete, specific theories and even models, making the standard for a 'revolutionary' change very small indeed, in the limit reaching any kind of conceptual change whatsoever.

We suggest that there is actually some validity to both moves, that there is a sort of (weak) self-similarity involved in scientific change. Every scale of size and complexity, from local problem-solving to big-picture science, features progress of the 'normal science' type, punctuated by occasional revolutions. For example, in working on an applied research or consulting problem, one typically will start in a certain direction, then suddenly realize one was thinking about it incorrectly, then move forward, and so forth. In a consulting setting, this re-evaluation can happen several times in a couple of

---

at least three big problems. First, all our earlier objections to using posterior probabilities to chose between theories apply, with all the more force because every paradigm is compatible with a broad range of specific theories. Second, devising priors encoding those methodological preferences – particularly a non-vacuous preference for parsimony – is hard or impossible in practice (Kelly, 2010). Third, it implies a truly remarkable form of Platonism: for scientists to give a paradigm positive posterior probability, they must, by Bayes's rule, have always given it strictly positive prior probability, *even before having encountered a statement of the paradigm*.

[33] Arguably this is true even of Kuhn (1957).

hours. At a slightly longer time scale, we commonly reassess any approach to an applied problem after a few months, realizing there was some key feature of the problem we were misunderstanding, and so forth. There is a link between the size and the typical time scales of these changes, with small revolutions occurring fairly frequently (every few minutes for an exam-type problem), up to every few decades for a major scientific consensus. (This is related to but somewhat different from the recursive subject-matter divisions discussed by Abbott, 2001.) The big changes are more exciting, even glamorous, but they rest on the hard work of extending the implications of theories far enough that they can be decisively refuted.

To sum up, our views are much closer to Popper's than to Kuhn's. The latter encouraged a close attention to the history of science and to explaining the process of scientific change, as well as putting on the agenda many genuinely deep questions, such as when and how scientific fields achieve consensus. There are even analogies between Kuhn's ideas and what happens in good data-analytic practice. Fundamentally, however, we feel that deductive model checking is central to statistical and scientific progress, and that it is the threat of such checks that motivates us to perform inferences within complex models that we know ahead of time to be false.

7. Why does this matter?

Philosophy matters to practitioners because they use it to guide their practice; even those who believe themselves quite exempt from any philosophical influences are usually the slaves of some defunct methodologist. The idea of Bayesian inference as inductive, culminating in the computation of the posterior probability of scientific hypotheses, has had malign effects on statistical practice. At best, the inductivist view has encouraged researchers to fit and compare models without checking them; at worst, theorists have actively discouraged practitioners from performing model checking because it does not fit into their framework.

In our hypothetico-deductive view of data analysis, we build a statistical model out of available parts and drive it as far as it can take us, and then a little farther. When the model breaks down, we dissect it and figure out what went wrong. For Bayesian models, the most useful way of figuring out how the model breaks down is through posterior predictive checks, creating simulations of the data and comparing them to the actual data. The comparison can often be done visually; see Gelman *et al.* (2004, Chapter 6) for a range of examples. Once we have an idea about where the problem lies, we can tinker with the model, or perhaps try a radically new design. Either way, we are using deductive reasoning as a tool to get the most out of a model, and we test the model – it is falsifiable, and when it is consequentially falsified, we alter or abandon it. None of this is especially subjective, or at least no more so than any other kind of scientific inquiry, which likewise requires choices as to the problem to study, the data to use, the models to employ, etc. – but these choices are by no means arbitrary whims, uncontrolled by objective conditions.

Conversely, a problem with the inductive philosophy of Bayesian statistics – in which science 'learns' by updating the probabilities that various competing models are true – is that it assumes that the true model (or, at least, the models among which we will choose or over which we will average) is one of the possibilities being considered. This does

not fit our own experiences of learning by finding that a model does not fit and needing to expand beyond the existing class of models to fix the problem.

Our methodological suggestions are to construct large models that are capable of incorporating diverse sources of data, to use Bayesian inference to summarize uncertainty about parameters in the models, to use graphical model checks to understand the limitations of the models, and to move forward via continuous model expansion rather than model selection or discrete model averaging. Again, we do not claim any novelty in these ideas, which we and others have presented in many publications and which reflect decades of statistical practice, expressed particularly forcefully in recent times by Box (1980) and Jaynes (2003). These ideas, important as they are, are hardly ground-breaking advances in statistical methodology. Rather, the point of this paper is to demonstrate that our commonplace (if not universally accepted) approach to the practice of Bayesian statistics is compatible with a hypothetico-deductive framework for the philosophy of science.

We fear that a philosophy of Bayesian statistics as subjective, inductive inference can encourage a complacency about picking or averaging over existing models rather than trying to falsify and go further.[34] Likelihood and Bayesian inference are powerful, and with great power comes great responsibility. Complex models can and should be checked and falsified. This is how we can learn from our mistakes.

Acknowledgements

We thank the National Security Agency for grant H98230-10-1-0184, the Department of Energy for grant DE-SC0002099, the Institute of Education Sciences for grants ED-GRANTS-032309-005 and R305D090006-09A, and the National Science Foundation for grants ATM-0934516, SES-1023176 and SES-1023189. We thank Wolfgang Beirl, Chris Genovese, Clark Glymour, Mark Handcock, Jay Kadane, Rob Kass, Kevin Kelly, Kristina Klinkner, Deborah Mayo, Martina Morris, Scott Page, Aris Spanos, Erik van Nimwegen, Larry Wasserman, Chris Wiggins, and two anonymous reviewers for helpful conversations and suggestions.

References

Abbott, A. (2001). *Chaos of disciplines*. Chicago: University of Chicago Press.

al Ghazali, Abu Hamid Muhammad ibn Muhammad at-Tusi (1100/1997). *The incoherence of the philosophers = Tahafut al-falasifah: A parallel English-Arabic text*, trans. M. E. Marmura. Provo, UT: Brigham Young University Press.

Ashby, W. R. (1960). *Design for a brain: The origin of adaptive behaviour* (2nd ed.). London: Chapman & Hall.

Atkinson, A. C., & Donev, A. N. (1992). *Optimum experimental designs*. Oxford: Clarendon Press.

Barkow, J. H., Cosmides, L., & Tooby, J. (Eds.) (1992). *The adapted mind: Evolutionary psychology and the generation of culture*. Oxford: Oxford University Press.

Bartlett, M. S. (1967). Inference and stochastic processes. *Journal of the Royal Statistical Society, Series A*, *130*, 457–478.

[34] Ghosh and Ramamoorthi (2003, p. 112) see a similar attitude as discouraging inquiries into consistency: 'the prior and the posterior given by Bayes theorem [*sic*] are imperatives arising out of axioms of rational behavior – and since we are already rational why worry about one more' criterion, namely convergence to the truth?

Bayarri, M. J., & Berger, J. O. (2000). *P* values for composite null models. *Journal of the American Statistical Association*, *95*, 1127–1142.

Bayarri, M. J., & Berger, J. O. (2004). The interplay of Bayesian and frequentist analysis. *Statistical Science*, *19*, 58–80. doi:10.1214/088342304000000116

Bayarri, M. J., & Castellanos, M. E. (2007). Bayesian checking of the second levels of hierarchical models. *Statistical Science*, *22*, 322–343.

Berger, J. O., & Sellke, T. (1987). Testing a point null hypothesis: Irreconcilability of *p*-values and evidence. *Journal of the American Statistical Association*, *82*, 112–122.

Berk, R. H. (1966). Limiting behavior of posterior distributions when the model is incorrect. *Annals of Mathematical Statistics*, *37*, 51–58. doi:10.1214/aoms/1177699597 Correction: *37* (1966), 745–746.

Berk, R. H. (1970). Consistency a posteriori. *Annals of Mathematical Statistics*, *41*, 894–906. doi:10.1214/aoms/1177696967

Bernard, C. (1865/1927). *Introduction to the study of experimental medicine*, trans. H. C. Greene. New York: Macmillan. First published as *Introduction à l'étude de la médecine experimentale*, Paris: J. B. Baillière. Reprinted New York: Dover, 1957.

Bernardo, J. M., & Smith, A. F. M. (1994). *Bayesian theory*. New York: Wiley.

Binmore, K. (2007). Making decisions in large worlds. Technical Report 266, ESRC Centre for Economic Learning and Social Evolution, University College London. Retrieved from http://else.econ.ucl.ac.uk/papers/uploaded/266.pdf

Bousquet, O., Boucheron, S., & Lugosi, G. (2004). Introduction to statistical learning theory. In O. Bousquet, U. von Luxburg, & G. Rätsch (Eds.), *Advanced lectures in machine learning* (pp. 169–207). Berlin: Springer.

Box, G. E. P. (1980). Sampling and Bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society, Series A*, *143*, 383–430.

Box, G. E. P. (1983). An apology for ecumenism in statistics. In G. E. P. Box, T. Leonard & C.-F. Wu (Eds.), *Scientific inference, data analysis, and robustness* (pp. 51–84). New York: Academic Press.

Box, G. E. P. (1990). Comment on 'The unity and diversity of probability' by Glen Shafer. *Statistical Science*, *5*, 448–449. doi:10.1214/ss/1177012024

Braithwaite, R. B. (1953). *Scientific explanation: A study of the function of theory, probability and law in science*. Cambridge: Cambridge University Press.

Brown, R. Z., Sallow, W., Davis, D. E., & Cochran, W. G. (1955). The rat population of Baltimore, 1952. *American Journal of Epidemiology*, *61*, 89–102.

Cesa-Bianchi, N., & Lugosi, G. (2006). *Prediction, learning, and games*. Cambridge: Cambridge University Press.

Claeskens, G., & Hjort, N. L. (2008). *Model selection and model averaging*. Cambridge: Cambridge University Press.

Cox, D. D. (1993). An analysis of Bayesian inference for nonparametric regression. *Annals of Statistics*, *21*, 903–923. doi:10.1214/aos/1176349157

Cox, D. R., & Hinkley, D. V. (1974). *Theoretical statistics*. London: Chapman & Hall.

Cox, R. T. (1946). Probability, frequency, and reasonable expectation. *American Journal of Physics*, *14*, 1–13.

Cox, R. T. (1961). *The algebra of probable inference*. Baltimore, MD: Johns Hopkins University Press.

Csiszár, I. (1995). Maxent, mathematics, and information theory. In K. M. Hanson& R. N. Silver (Eds.), *Maximum entropy and Bayesian methods: Proceedings of the Fifteenth International Workshop on Maximum Entropy and Bayesian Methods* (pp. 35–50). Dordrecht: Kluwer Academic.

Dawid, A. P., & Vovk, V. G. (1999). Prequential probability: Principles and properties. *Bernoulli*, *5*, 125–162. Retrieved from: http://projecteuclid.org/euclid.bj/1173707098

Donovan, A., Laudan, L., & Laudan, R. (Eds.), (1988). *Scrutinizing science: Empirical studies of scientific change*. Dordrecht: Kluwer Academic. Reprinted 1992 (Baltimore, MD: Johns Hopkins University Press) with a new introduction.

Doob, J. L. (1949). Application of the theory of martingales. In *Colloques internationaux du Centre National de la Recherche Scientifique*, Vol. 13 (pp. 23–27). Paris: Centre National de la Recherche Scientifique.

Duhem, P. (1914/1954). *The aim and structure of physical theory*, trans. P. P. Wiener. Princeton, NJ: Princeton University Press.

Earman, J. (1992). *Bayes or bust? A critical account of Bayesian confirmation theory*. Cambridge, MA: MIT Press.

Eggertsson, T. (1990). *Economic behavior and institutions*. Cambridge: Cambridge University Press.

Fitelson, B., & Thomason, N. (2008). Bayesians sometimes cannot ignore even very implausible theories (even ones that have not yet been thought of). *Australasian Journal of Logic*, *6*, 25–36. Retrieved from: http://philosophy.unimelb.edu.au/ajl/2008/2008_2.pdf

Foster, D. P., & Young, H. P. (2003). Learning, hypothesis testing and Nash equilibrium. *Games and Economic Behavior*, *45*, 73–96. doi:10.1016/S0899-8256(03)00025-3

Fraser, D. A. S., & Rousseau, J. (2008). Studentization and deriving accurate *p*-values. *Biometrika*, *95*, 1–16. doi:10.1093/biomet/asm093

Freedman, D. A. (1999). On the Bernstein-von Mises theorem with infinite-dimensional parameters. *Annals of Statistics*, *27*, 1119–1140. doi:10.1214/aos/1017938917

Gelman, A. (1994). Discussion of 'A probabilistic model for the spatial distribution of party support in multiparty elections' by S. Merrill. *Journal of the American Statistical Association*, *89*, 1198.

Gelman, A. (2003). A Bayesian formulation of exploratory data analysis and goodness-of-fit testing. *International Statistical Review*, *71*, 369–382. doi:10.1111/j.1751-5823.2003.tb00203.x

Gelman, A. (2004). Treatment effects in before-after data. In A. Gelman& X.-L. Meng (Eds.), *Applied Bayesian modeling and causal inference from incomplete-data perspectives* (pp. 191–198). Chichester: Wiley.

Gelman, A. (2007). Comment: 'Bayesian checking of the second levels of hierarchical models'. *Statistical Science*, *22*, 349–352. doi:10.1214/07-STS235A

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis* (2nd ed.). Boca Raton, FL: CRC Press.

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge: Cambridge University Press.

Gelman, A., Jakulin, A., Pittau, M. G., & Su, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *Annals of Applied Statistics*, *2*, 1360–1383. doi:10.1214/08-AOAS191

Gelman, A., & King, G. (1994). Enhancing democracy through legislative redistricting. *American Political Science Review*, *88*, 541–559.

Gelman, A., Lee, D., & Ghitza, Y. (2010). Public opinion on health care reform. *The Forum*, *8*(1). doi:10.2202/1540-8884.1355

Gelman, A., Meng, X.-L., & Stern, H. S. (1996). Posterior predictive assessment of model fitness via realized discrepancies (with discussion). *Statistica Sinica*, *6*, 733–807. Retrieved from: http://www3.stat.sinica.edu.tw/statistica/j6n4/j6n41/j6n41.htm

Gelman, A., Park, D., Shor, B., Bafumi, J., & Cortina, J. (2008). *Red state, blue state, rich state, poor state: Why Americans vote the way they do*. Princeton, NJ: Princeton University Press. doi:10.1561/100.00006026

Gelman, A., & Rubin, D. B. (1995). Avoiding model selection in Bayesian social research. *Sociological Methodology*, *25*, 165–173.

Gelman, A., Shor, B., Park, D., & Bafumi, J. (2008). Rich state, poor state, red state, blue state: What's the matter with Connecticut? *Quarterly Journal of Political Science*, *2*, 345–367.

Ghitza, Y., & Gelman, A. (2012). *Deep interactions with MRP: presidential turnout and voting patterns among small electoral subgroups*. Technical report, Department of Political Science, Columbia University.

Ghosh, J. K., & Ramamoorthi, R. V. (2003). *Bayesian nonparametrics*. New York: Springer.

Giere, R. N. (1988). *Explaining science: A cognitive approach*. Chicago: University of Chicago Press.

Gigerenzer, G. (2000). *Adaptive thinking: Rationality in the real world*. Oxford: Oxford University Press.

Gigerenzer, G., Todd, P. M., & ABC Research Group. (1999). *Simple heuristics that make us smart*. Oxford: Oxford University Press.

Glymour, C. (1980). *Theory and evidence*. Princeton, NJ: Princeton University Press.

Good, I. J. (1983). *Good thinking: The foundations of probability and its applications*. Minneapolis: University of Minnesota Press.

Good, I. J., & Crook, J. F. (1974). The Bayes/non-Bayes compromise and the multinomial distribution. *Journal of the American Statistical Association*, *69*, 711–720.

Gray, R. M. (1990). *Entropy and information theory*. New York: Springer.

Greenland, S. (1998). Induction versus Popper: Substance versus semantics. *International Journal of Epidemiology*, *27*, 543–548. doi:10.1093/ije/27.4.543

Greenland, S. (2009). Relaxation penalties and priors for plausible modeling of nonidentified bias sources. *Statistical Science*, *24*, 195–210. doi:10.1214/09-STS291

Grünwald, P. D. (2007). *The minimum description length principle*. Cambridge, MA: MIT Press.

Grünwald, P. D., & Langford, J. (2007). Suboptimal behavior of Bayes and MDL in classification under misspecification. *Machine Learning*, *66*, 119–149. doi:10.1007/s10994-007-0716-7

Gustafson, P. (2005). On model expansion, model contraction, identifiability and prior information: Two illustrative scenarios involving mismeasured variables. *Statistical Science*, *20*, 111–140. doi:10.1214/088342305000000098

Guttorp, P. (1995). *Stochastic modeling of scientific data*. London: Chapman & Hall.

Haack, S. (1993). *Evidence and inquiry: Towards reconstruction in epistemology*. Oxford: Blackwell.

Hacking, I. (2001). *An introduction to probability and inductive logic*. Cambridge: Cambridge University Press.

Halpern, J. Y. (1999). Cox's theorem revisited. *Journal of Artificial Intelligence Research*, *11*, 429–435. doi:10.1613/jair.644

Handcock, M. S. (2003). Assessing degeneracy in statistical models of social networks. Working Paper no. 39, Center for Statistics and the Social Sciences, University of Washington. Retrieved from http://www.csss.washington.edu/Papers/wp39.pdf

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Berlin: Springer.

Hempel, C. G. (1965). *Aspects of scientific explanation*. Glencoe, IL: Free Press.

Hill, J. R. (1990). A general framework for model-based statistics. *Biometrika*, *77*, 115–126.

Hjort, N. L., Holmes, C., Müller, P., & Walker, S. G. (Eds.), (2010). *Bayesian nonparametrics*. Cambridge: Cambridge University Press.

Holland, J. H., Holyoak, K. J., Nisbett, R. E., & Thagard, P. R. (1986). *Induction: Processes of inference, learning, and discovery*. Cambridge, MA: MIT Press.

Howson, C., & Urbach, P. (1989). *Scientific reasoning: The Bayesian approach*. La Salle, IL: Open Court.

Hunter, D. R., Goodreau, S. M., & Handcock, M. S. (2008). Goodness of fit of social network models. *Journal of the American Statistical Association*, *103*, 248–258. doi:10.1198/016214507000000446

Jaynes, E. T. (2003). *Probability theory: The logic of science*. Cambridge: Cambridge University Press.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795.

Kass, R. E., & Vos, P. W. (1997). *Geometrical foundations of asymptotic inference*. New York: Wiley.

Kass, R. E., & Wasserman, L. (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, *91*, 1343–1370.

Kelly, K. T. (1996). *The logic of reliable inquiry*. Oxford: Oxford University Press.

Kelly, K. T. (2010). Simplicity, truth, and probability. In P. Bandyopadhyay & M. Forster (Eds.), *Handbook on the philosophy of statistics*. Dordrecht: Elsevier.

Kitcher, P. (1993). *The advancement of science: Science without legend, objectivity without illusions*. Oxford: Oxford University Press.

Kleijn, B. J. K., & van der Vaart, A. W. (2006). Misspecification in infinite-dimensional Bayesian statistics. *Annals of Statistics*, *34*, 837–877. doi:10.1214/009053606000000029

Kolakowski, L. (1968). *The alienation of reason: A history of positivist thought*, trans. N. Guterman. Garden City, NY: Doubleday.

Kuhn, T. S. (1957). *The Copernican revolution: Planetary astronomy in the development of western thought*. Cambridge, MA: Harvard University Press.

Kuhn, T. S. (1970). *The structure of scientific revolutions* (2nd ed.). Chicago: University of Chicago Press.

Lakatos, I. (1978). *Philosophical papers*. Cambridge: Cambridge University Press.

Laudan, L. (1996). *Beyond positivism and relativism: Theory, method and evidence*. Boulder, Colorado: Westview Press.

Laudan, L. (1981). *Science and hypothesis*. Dodrecht: D. Reidel.

Li, Q., & Racine, J. S. (2007). *Nonparametric econometrics: Theory and practice*. Princeton, NJ: Princeton University Press.

Lijoi, A., Prünster, I., & Walker, S. G. (2007). Bayesian consistency for stationary models. *Econometric Theory*, *23*, 749–759. doi:10.1017/S0266466607070314

Lindsay, B., & Liu, L. (2009). Model assessment tools for a model false world. *Statistical Science*, *24*, 303–318. doi:10.1214/09-STS302

Manski, C. F. (2007). *Identification for prediction and decision*. Cambridge, MA: Harvard University Press.

Manski, C. F. (2011). Actualist rationality. *Theory and Decision*, 71. doi:10.1007/s11238-009-9182-y

Mayo, D. G. (1996). *Error and the growth of experimental knowledge*. Chicago: University of Chicago Press.

Mayo, D. G., & Cox, D. R. (2006). Frequentist statistics as a theory of inductive inference. In J. Rojo (ed.), *Optimality: The Second Erich L. Lehmann Symposium* (pp. 77–97). Bethesda, MD: Institute of Mathematical Statistics.

Mayo, D. G., & Spanos, A. (2004). Methodology in practice: Statistical misspecification testing. *Philosophy of Science*, *71*, 1007–1025.

Mayo, D. G., & Spanos, A. (2006). Severe testing as a basic concept in a Neyman-Pearson philosophy of induction. *British Journal for the Philosophy of Science*, *57*, 323–357. doi:10.1093/bjps/axl003

McAllister, D. A. (1999). Some PAC-Bayesian theorems. *Machine Learning*, *37*, 355–363. doi:10.1023/A:1007618624809

McCarty, N., Poole, K. T., & Rosenthal, H. (2006). *Polarized America: The dance of ideology and unequal riches*. Cambridge, MA: MIT Press.

Merrill III, S. (1994). A probabilistic model for the spatial distribution of party support in multiparty electorates. *Journal of the American Statistical Association*, *89*, 1190–1197.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, *21*, 1087–1092. doi:10.1063/1.1699114

Morris, C. N. (1986). Comment on 'Why isn't everyone a Bayesian?'. *American Statistician*, *40*, 7–8.

Müller, U. K. (2011). Risk of Bayesian inference in misspecified models, and the sandwich covariance matrix. *Econometrica*, submitted. Retrieved from http://www.princeton.edu/~umueller/sandwich.pdf

Newman, M. E. J., & Barkema, G. T. (1999). *Monte Carlo methods in statistical physics*. Oxford: Clarendon Press.

Norton, J. D. (2003). A material theory of induction. *Philosophy of Science*, *70*, 647–670. doi:10.1086/378858

Paninski, L. (2005). Asymptotic theory of information-theoretic experimental design. *Neural Computation*, *17*, 1480–1507. doi:10.1162/0899766053723032

Popper, K. R. (1934/1959). *The logic of scientific discovery*. London: Hutchinson.

Popper, K. R. (1945). *The open society and its enemies*. London: Routledge.

Quine, W. V. O. (1961). *From a logical point of view: Logico-philosophical essays* (2nd ed.). Cambridge, MA: Harvard University Press.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, *25*, 111–196.

Ripley, B. D. (1988). *Statistical inference for spatial processes*. Cambridge: Cambridge University Press.

Rivers, D., & Vuong, Q. H. (2002). Model selection tests for nonlinear dynamic models. *Econometrics Journal*, *5*, 1–39. doi:10.1111/1368-423X.t01-1-00071

Robins, J. M., van der Vaart, A., & Ventura, V. (2000). Asymptotic distribution of *p* values in composite null models (with discussions and rejoinder). *Journal of the American Statistical Association*, *95*, 1143–1172.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Annals of Statistics*, *6*, 34–58. doi:10.1214/aos/1176344064

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics*, *12*, 1151–1172. doi:10.1214/aos/1176346785

Russell, B. (1948). *Human knowledge: Its scope and limits*. New York: Simon and Schuster.

Salmon, W. C. (1990). The appraisal of theories: Kuhn meets Bayes. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association* (Vol. 2, pp. 325–332). Chicago: University of Chicago Press.

Savage, L. J. (1954). *The foundations of statistics*. New York: Wiley.

Schervish, M. J. (1995). *Theory of statistics*. Berlin: Springer.

Seidenfeld, T. (1979). Why I am not an objective Bayesian: Some reflections prompted by Rosenkrantz. *Theory and Decision*, *11*, 413–440. doi:10.1007/BF00139451

Seidenfeld, T. (1987). Entropy and uncertainty. In I. B. MacNeill& G. J. Umphrey (Eds.), *Foundations of statistical inference* (pp. 259–287). Dordrecht: D. Reidel.

Shalizi, C. R. (2009). Dynamics of Bayesian updating with dependent data and misspecified models. *Electronic Journal of Statistics*, *3*, 1039–1074. doi:10.1214/09-EJS485

Snijders, T. A. B., Pattison, P. E., Robins, G. L., & Handcock, M. S. (2006). New specifications for exponential random graph models. *Sociological Methodology*, *36*, 99–153. doi:10.1111/j.1467-9531.2006.00176.x

Spanos, A. (2007). Curve fitting, the reliability of inductive inference, and the error-statistical approach. *Philosophy of Science*, *74*, 1046–1066. doi:10.1086/525643

Stove, D. C. (1982). *Popper and after: Four modern irrationalists*. Oxford: Pergamon Press.

Stove, D. C. (1986). *The rationality of induction*. Oxford: Clarendon Press.

Tilly, C. (2004). Observations of social processes and their formal representations. *Sociological Theory*, *22*, 595–602. Reprinted in Tilly (2008). doi:10.1111/j.0735-2751.2004.00235.x

Tilly, C. (2008). *Explaining social processes*. Boulder, CO: Paradigm.

Toulmin, S. (1972). *Human understanding: The collective use and evolution of concepts*. Princeton, NJ: Princeton University Press.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

Uffink, J. (1995). Can the maximum entropy principle be explained as a consistency requirement? *Studies in the History and Philosophy of Modern Physics*, *26B*, 223–261. doi:10.1016/1355-2198(95)00015-1

Uffink, J. (1996). The constraint rule of the maximum entropy principle. *Studies in History and Philosophy of Modern Physics*, *27*, 47–79. doi:10.1016/1355-2198(95)00022-4

Vansteelandt, S., Goetghebeur, E., Kenward, M. G., & Molenberghs, G. (2006). Ignorance and uncertainty regions as inferential tools in a sensitivity analysis. *Statistica Sinica*, *16*, 953–980.

Vidyasagar, M. (2003). *Learning and generalization: With applications to neural networks* (2nd ed.). Berlin: Springer.

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, *57*, 307–333.

Wahba, G. (1990). *Spline models for observational data*. Philadelphia: Society for Industrial and Applied Mathematics.

Wasserman, L. (2006). Frequentist Bayes is objective. *Bayesian Analysis*, *1*, 451–456. doi:10.1214/06-BA116H

Weinberg, S. (1999). What is quantum field theory, and what did we think it was? In T. Y. Cao (Ed.), *Conceptual foundations of quantum field theory* (pp. 241–251). Cambridge: Cambridge University Press.

White, H. (1994). *Estimation, inference and specification analysis*. Cambridge: Cambridge University Press.

Wooldridge, J. M. (2002). *Econometric analysis of cross section and panel data*. Cambridge, MA: MIT Press.

Ziman, J. (2000). *Real science: What it is, and what it means*. Cambridge: Cambridge University Press.

Received 28 June 2011; revised version received 6 December 2011