A Philosophical Analysis of Bayesian model selection

Romeijn, J.-W.; Schoot, R. van de

Published in:
EPRINTS-BOOK-TITLE

Publication date:
2008

A Philosophical Analysis of Bayesian Model Selection for Inequality Constrained Models

Jan-Willem Romeijn[1] and Rens van de Schoot[2]

[1] Department of Theoretical Philosophy, Oude Boteringestraat 52, 9712 GL, Groningen, the Netherlands j.w.romeijn@rug.nl
[2] Department of Methodology and Statistics, PO Box 80140, 3508 TC Utrecht, the Netherlands a.g.j.vandeschoot@uu.nl

1 Bayesian Model Selection

This chapter provides an answer to the question what it is, philosophically speaking, to choose a model in a statistical procedure, and what this amounts to in the context of a Bayesian inference. Special attention is given to Bayesian model selection, specifically the choice between inequality-constrained and unconstrained models based on their Bayes factors and posterior model probabilities .

Many of the foregoing chapters have provided examples of model selection by means of Bayes factors, and chapter 4 has provided a thorough introduction to the subject. For the sake of completeness, and in order to introduce some terminology that will be used in this chapter, we will briefly rehearse Bayesian model selection here. Say that we have some data E, and that we think these data are sampled from a distribution $p_{\mu_p\mu_s}(E)$, characterized by two parameters $\mu_p \in [0,1]$ and $\mu_s \in [0,1]$. We say that each pair of values for μ_p and μ_s presents a specific hypothesis $H_{\mu_p\mu_s}$ concerning the data. By contrast, a statistical model consists of a set of hypotheses. One possible statistical model for the data allows for all possible values of both parameters, that is, $\langle \mu_p, \mu_s \rangle \in [0,1]^2$. Call this model $\mathcal{M}_0$; it consists of the entire range of hypotheses $H_{\mu_p\mu_s}$. Another possible model, $\mathcal{M}_1$, imposes the restriction that $\mu_p > \mu_s$; this model is restricted by an inequality constraint . Note that both models consist of a particular set of statistical hypotheses $H_{\mu_p\mu_s}$, each of which fixes a fully specified distribution for the data, $p(E|H_{\mu_p\mu_s}) = p_{\mu_p\mu_s}(E)$. Their difference is that the latter restrict the possible values for the parameters μ_p and μ_s. How can we compare these two models?

The Bayesian model selection procedure, as discussed in this book, presents an answer to the latter question. This answer employs the so-called marginal likelihoods of the models:

$$p(E|\mathcal{M}_j) = \int_0^1 \int_0^1 p(E|H_{\mu_p\mu_s}) p_j(H_{\mu_p\mu_s})\, d\mu_p d\mu_s$$
$$= \int_0^1 \int_0^1 p_{\mu_p\mu_s}(E) p_j(H_{\mu_p\mu_s})\, d\mu_p d\mu_s, \quad (1)$$

where $j = 0, 1$ indexes the models. Notice that for both models, the integration runs over the whole domain $[0, 1]^2$. However, the prior for the two models is different: $p_0(H_{\mu_p\mu_s})d\mu_p\mu_s = 1$ while $p_1(H_{\mu_p\mu_s})d\mu_p\mu_s = 2$ if $\mu_p > \mu_s$ and $p_1(H_{\mu_p\mu_s})d\mu_p\mu_s = 0$ otherwise. Both these priors integrate to one, but the prior for $\mathcal{M}_1$ is such that only the distributions for which $\mu_p > \mu_s$ are included in the computation of the marginal likelihood. Finally, it must be emphasized that the marginal likelihood for a model is different from the ordinary likelihood of a hypothesis, although both are probabilities of the data E. The likelihood of a hypothesis is the well-known expression $p(E|H_{\mu_p\mu_s})$. The marginal likelihood of a model is essentially a mixture of the likelihoods of hypotheses that are included in the model, weighted with the probability of the hypotheses.

We may now use these expressions of the marginal likelihood to compute the Bayes factor for the models, $\mathcal{M}_0$ and $\mathcal{M}_1$:

$$BF_{01} = \frac{p(E|\mathcal{M}_0)}{p(E|\mathcal{M}_1)}. \quad (2)$$

It may then turn out that $BF_{01} \ll 1$, in which case the unrestricted model $\mathcal{M}_0$ seems strongly favored over the restricted model $\mathcal{M}_1$. But here we may wonder: what support exactly is provided by the high value of the Bayes factor? It must be emphasized that the comparison of two models , for example $\mathcal{M}_0$ versus $\mathcal{M}_1$, is not the same as a comparison between two rival hypotheses, for example $H_{1/2 1/2}$ versus $H_{1/3 2/3}$, because models are not themselves hypotheses, rather they are sets of hypotheses. In the case of hypotheses, a comparison by means of a Bayes factor makes perfect sense. But a Bayes factor may not be suitable for the comparison between models. This point is particularly pressing for comparisons of inequality-constrained models, because they contain partially overlapping sets of hypotheses.

In this chapter we set out to investigate this latter question from a foundational perspective. We discuss what statistics was supposed to deliver in the first place, and in what way Bayesian statistics delivers this. After we have got clear on Bayesian statistics in its ordinary application, we can discuss the application of Bayesian statistics in the context of model selection, and in particular in the context of comparing models with different inequality constraints. It will be seen that this leads to a challenging question on the exact use, or function, of Bayes factors for models.

The chapter is set up as follows. In Section 2 we spell out the philosophical setting for statistical inference, dealing with the problem of induction and with the answers to this problem provided by Popper and Carnap. Section 3 presents a parallel between statistical inference and another system of

reasoning, deductive logic. In Section 4, based on this parallel, we describe the role of a statistical model in a Bayesian statistical inference as a specific type of premise in an inductive. We can thereby identify elements of the views of both Popper and Carnap in Bayesian statistical inference, and extend Bayesian inference to model selection, in particular the selection by means of Bayes factors. This leads to a discussion of some problematic aspects of Bayesian model selection procedures in Section 5. We will address two specific worries. Firstly, a comparison of models in terms of their posterior model probabilities does not seem to make sense if the models overlap. We will remedy this by organizing the space on which the models are defined a bit differently. Secondly, and in view of this reorganization, we ask how we can interpret the probability assignments to hypotheses.

2 Statistics and the Problem of Induction

This section deals with statistics, its relation to the problem of induction, and the solutions that Popper and Carnap provided for this problem, drawing on standard textbooks in the philosophy of science such as Bird [2] and Curd and Cover [6]. We will see that these solutions, in this context termed inductivism and rationalism, are endpoints in a spectrum of positions, and that as such they both miss out on an important aspect of statistical reasoning.

2.1 The Problem of Induction

Induction is a mode of inference that allows us to move from observed data to as yet unknown data elements and empirical generalizations. A typical example of an inductive inference is presented in statements 1 and 2:

1. The sun has risen every morning up until now.
2. So, the sun will also rise tomorrow.
3. Even stronger, it will rise on all future days.
4. Alternatively, it will probably rise on all future days.
5. Or at least it will probably rise tomorrow.

Here the observed data is expressed in 1, namely that the sun has always risen up until now. This observed data may be viewed as the sole premise of the inference. On the basis of it we may want to affirm several other statements, labeled 2 to 5, all of which can be viewed as conclusions of an inductive inference.

Premises and conclusions are statements, and as such they may be true or false. Of an inference, however, we cannot say that it is true or false. Rather we say that it is valid or invalid, where validity means that the inference provides a certain kind of guarantee: if the premisses are true and the inference from these premisses to a conclusion is valid, then we have the guarantee that the conclusion is true. When applied to the inductive inferences above,

validity means the following: if the sun has indeed always risen up until now, and if the inductive inference is valid, then we can rely on the truth of the conclusion, namely that the sun will also rise tomorrow. The trouble with inductive inference, as presented above, is that its validity is very hard to establish. Nobody will seriously doubt the truth of the statement that the sun will rise tomorrow. But when asked whether on the basis of all its past risings we can validly infer, and hence whether we are justified to believe, that the sun will rise tomorrow, we are met with embarrassing difficulties.

Let us examine these difficulties in some more detail. David Hume asked himself the question how we can derive new observations from observations that we have done in the past. He argued in An Enquiry Concerning Human Understanding [15]: "But why this experience should be extended to future times, and to other objects, which for aught we know, may be only in appearance similar; this is the main question on which I would insist" (pp. 33–34). In other words, inductive inferences seem to presuppose that a sequence of observations in the future will occur as it always has in the past. However, even very long series of the same observation are perfectly consistent with the next observation being quite different. The problem of induction is that no further basis can be found in the observations themselves for this presupposed constancy of the observations. To illustrate again with the example, we might conclude from the observations that the sun has risen every morning up until now that the next morning the sun will also rise. But what justification can there be for presupposing this constancy? In the next subsection we will try to provide some possible answers to this question, and we will show how these answers fail.

2.2 Uniformity Assumptions

A first possible answer is to justify the presupposition of constancy, and hence inductive inference, by using induction itself. That is, we could say that an inductive inference will work in the future because it has worked in the past. For example, we made many inductive inferences about many different topics, for example that the sun has always risen, and until now all these inferences led to true conclusions. Or closer to scientific practice, we have often used a T-test successfully in the past, and so we may conclude that it will be a valuable method in the future as well. However, if we justify induction on the grounds that it has worked in the past, then we enter a vicious circle. The argument fails to prove anything, because it takes for granted what it is supposed to prove. We can therefore run the exact same criticism of induction again, this time on the level of the inferences: there is, again, no logical necessity that the previous success of the inferences guarantees future successes.

A second possible answer is to justify the constancy of observations by assuming an overall uniformity of nature. For example, we might say that the sun has always risen in the past and that since nature is uniform, this pattern will continue into the future. Note, however, that this is quite a strong

assumption. It is not just questionable whether nature really is uniform, it is also dubitable whether we can apply this assumption to the natural and the human sciences alike. Uniformity could hold for the natural sciences, like the sunrise example, but whether it is also applicable to human science remains discussable. Is it for example true that the positive correlation between social isolation and aggression, as it might be established in psychology, continues to hold after the introduction of internet? To infer by induction, we have to assume a rather strong uniformity in nature.

Moreover, even while the assumption of uniformity must be very strong, one might argue that it is still not strong enough. Just stating that nature is uniform does not yet determine the exact patterns that will continue in future times. This problem is nicely brought out by Goodman's [11] so-called new riddle of induction, which we will present briefly. Say that the predicate Green belongs to emeralds which appear to have the color green at any time. Suppose that up until the year 2008 we have observed many emeralds to be Green. We thus have evidence statements that emerald 1 is Green, emerald 2 is Green, etc. The standard inductive inference then is that all emeralds examined before the year 2008 were Green, so emeralds after that year will be Green as well. In this case we call the predicate Green projectable: findings of the past can be projected unto the future. But now consider a somewhat different predicate: an object is Grue if either it has been observed before 2008 and it appeared green, or it has been observed after 2008 and appeared blue. Similarly, something is Bleen if observed before 2008 appearing blue, or after 2008 appearing green. We may redescribe what we observed until now as emerald 1 is Grue, emerald 2 is Grue, etc. So with the very same inductive inference just used on Green, but now taking Grue to be the projectable predicate, we might conclude that emeralds observed after 2010 will be Grue, so that we predict emeralds observed after 2010 will appear blue to us! It thus seems that simply assuming the uniformity of nature is not specific enough. If we are to apply the uniformity assumption, we must stipulate the exact predicates with respect to which nature is uniform.

In reaction to Goodman's riddle, we might argue that we can make a principled distinction between candidate predicates on grounds of their simplicity, defending induction by saying nature is uniform and simple. It seems that a model where emerald are Green before time 2010 and are also Green after 2010 is simpler. However, we might also describe this model in a complicated way, saying that emeralds are Grue before time 2010 and are Bleen after 2010: in both cases the result is that emeralds appear green throughout. Goodman points out [11] (pp. 74–75) that predicates such as Grue and Bleen only appear to be more complex than the predicate Green or Blue. This is because we have defined Grue in terms of blue and green, whereas the predicate Green is only defined in term of the color green. In other words, the model we favor depends on which predicates are established in our language, leaving inductive inference relative to the language in which they are formulated. The ultimate question is therefore what predicates are considered the natural ones.

Hence we cannot salvage inductive inference by imposing further simplicity constraints. We need a decision on the projectability of certain patterns or predicates.

This last remark concludes our discussion on the philosophical problem of induction. The fact that in inductive inference we must always make a choice for a specific projectable predicate will reappear in later Sections.

2.3 Induction in Science

We now explain the problem of induction and its relevance to scientific practice, by identifying inductive inference within a more scientific example. The first thing to note here is that, especially in the social sciences, scientists make probabilistic inferences. In terms of the example of subsection 2.1, from the data expressed in statement 1, they generally derive statements like 4 and 5. This is because in the social sciences, the data often show patterns that are not completely stable. However, we can still say that such probabilistic inferences are inductive.

To illustrate induction, we will use a rather simplified version of the example provided in Chapter 1 about amnesia in Dissociative Identity Disorder (DID) . The research question of Huntjens et al. [16] is to determine whether DID is a genuine disorder or rather a iatrogenic disorder, that is, a pseudo-disorder caused by the influence of the therapist on suggestible individuals. The design allowed the authors to compare the overall memory performance, called the Recognition Scores, between true DID-patients, controls, DID-simulators, and true amnesiacs. Let us say we are now only interested in the question whether the memory performance of DID-patients differs from the performance of DID-simulators. If the performance of DID-patients is higher than that of DID-simulators, we conclude that DID is not a iatrogenic disorder. To investigate this difference, the researchers selected a sample from a population of people diagnosed with DID, and also a sample of 'normal' people who are asked to simulate DID. The memory performance of the two groups was observed in a number of trials and, based on the difference in the memory performance, a generalized statement was made about the existence of DID.

With this scientific example of Dissociative Identity Disorder in place, we can restate the problem of induction. Suppose that the observations until now show that the entire group of DID-patients is better in memory performance than the group DID-simulators. By induction we might then infer that all DID-patients are better in memory performance than DID-simulators, and hence that DID is a real disorder. Or alternatively, suppose that on average the DID-patients are better in memory performance than the DID-simulators. In that case we might infer, again by induction, that this average difference holds for the entire populations of DID-patients and DID-simulators, and hence that a randomly chosen DID-patient can be expected to have better

memory performance than a randomly chosen DID-simulator. This expectation is typically spelled out in terms of a probability assignment; in the social sciences, such probabilistic conclusions are much more common than strict universal generalisations.

Because such general or predictive conclusions concerning DID-patients and DID-simulators are arrived at by induction, they are subject to the problem, sketched in the foregoing, that they are very hard to justify. More in particular, as the discussion of Goodman's riddle suggested, justifying such conclusions involves the explicit choice for predicates that are projectable. We will argue in the next two Sections that the statistical justification of conclusions in the DID example requires such a choice. The predicates at issue are the test scores of the DID-patients and DID-simulators respectively: if we want these test scores to be indicative of what is going on in the populations at large, we must somehow assume that they are based on, or refer to, some stable properties of the individuals in that population. As already announced, we return to this in later sections. In order to properly discuss the assumption, we first turn to two well-known responses to the problem of induction, by Carnap and Popper respectively.

2.4 Carnap on the Problem of Induction

The philosophical discussion on the justification of induction is rich and multifaceted. In the following we will not provide an overview of this discussion, but rather we will present a specific take on it in order to portray statistics as a particular solution. For this we will first visit two important figures in the debate on induction, Karl Popper and Rudolf Carnap.

Carnap was one of the central figures of logical empiricism, a philosophical movement that dominated the philosophy of science in the first half of the twentieth century. In this movement, two discussions took center stage: one concerned the nature of science and its demarcation from pseudo-science, and the other concerned the justification of science, which was intimately connected to the justification of conclusions arrived at by inductive inference. For the logical empiricists, as the name suggests, the main features of science were its firm foundation in primitive empirical fact, and the further feature that more general scientific claims can be derived from these empirical facts by logical means. Hence the logical empiricists faced a double challenge: to establish the firm foundations of science in primitive empirical fact, and to provide a logical system that would allow us to derive more advanced scientific claims from these primitives.

Carnap's contribution to the second part of the logical empiricist programme is also the salient part of the programme for present purposes [4] [5]. Carnap tried to find the degree of confirmation that a given set of empirical evidence gives to some scientific hypothesis. To this aim he used both logic and probability theory. Both evidence and hypotheses were expressed in terms of a formal logical language, and the degree of confirmation was subsequently

expressed in terms of a probability function over this language, the so-called confirmation function $c(H, E)$. The function $c(H, E)$ is the degree to which hypothesis H is supported by evidence E, or in other words, c is the degree to which someone is rationally entitled to believe in the hypothesis H on the basis of full belief in the evidence E. The crucial ingredient in the determination of this function is Carnap's notion of logical probability: the probability assignment over the language in which H and E are sentences is fully determined by the structure of the language itself and symmetry requirements on the probability function with respect to the language. The confirmation function $c(H, E)$ can therefore be determined by a priori arguments from the language.

The main achievement of Carnap was that he managed to derive a general inductive rule on the basis of his concept of logical probability . This inductive rule allowed him to make justified predictions of future observations on the basis of a record of past observations. Say, for example, that we are given a record of the memory performance of n individuals, either DID-simulators or patients, in which a number of n_0 people scored below guessing level and n_1 people scored above guessing level. We may denote each individual test result with Q_i^q where $q \in \{0, 1\}$ and 0 means scoring below, 1 means scoring above guessing level. The record of all n results is $E_n = \bigcap_{i=1}^{n} Q_i$. Carnap's c-function then gives the degree of confirmation for the next person passing the test, the event denoted with Q_{n+1}^1:

$$c(Q_{n+1}^1, E_n) = \frac{n_1 + \gamma\lambda}{n_0 + n_1 + \lambda}, \tag{3}$$

where γ is the initial probability for passing the test and λ the firmness of that initial estimate. This degree of confirmation for Q_{n+1}^1 is the best guess we can make for the performance of the next individual; depending on the data we may thus be able to conclude that the predictions for DID-patients and DID-simulators differ. Carnap maintained that in this way he solved the problem of induction. By casting the problem in a formal framework, defining a function that made explicit the degree to which we are rationally entitled to believe hypotheses on the basis of evidence, and by grounding this degree in the structure of the logical framework, he provided a logical system that allows us to derive predictions, albeit probabilistic ones, from the primitive empirical facts.

One of the weaknesses of Carnap's system is that it is fairly abstract, and that it does not readily connect to the methods and statistical techniques used by scientists. For the purpose of this chapter, however, we would like to point to another set of related worries, to do with language as a determining factor in the Carnapian system. Recall that the justification of the Carnapian inductive inferences rests on applying symmetry principles, as determined by the notion of logical probability, to some language. Moreover, following Goodman, we are stuck with an assumption on which predicates are projectable once the language is chosen. If the language adopts Grue and Bleen, then those are

the predicates that will accumulate inductive confirmation or disconfirmation. The obvious question is: how do we determine the exact set of predicates to which the notion of logical probability can be applied? First of all, language in the Carnapian system is idealized and highly artificial, whereas most scientific theories are expressed in vague language, usually English. It is unclear how to isolate the salient predicates from the fluid scientific discourse. Related to this, to apply a Carnapian system we must hold this artificial language constant, and refuse new predicates to be introduced, or otherwise we must accept that the degree of confirmation of scientific hypotheses will change whenever new predicates are introduced. But both options sit badly with scientific method as we know it.

And finally, even if we accept the artificiality and the fixity of the language, we encounter a problem with its poverty, because the notion of a statistical or general hypothesis is virtually absent from it. The way Carnap has set up his inductive logic and the confirmation function $c(H, E)$ in it, both the evidence E and the hypothesis H must be finite expressions in a language that only has observations as primitive terms. Typically, the evidence and hypotheses are past and future observations respectively, as in the example provided above. Now it must be admitted that this is largely due to a philosophical predisposition among the logical empiricists, namely to restrict scientific inference to the empirical realm. In principle the formalism allows for extensions to general hypotheses, as attested by the inductive logical systems of Hintikka [12]. However, the inclusion of general hypotheses in Carnapian inductive logic remains very limited, and attempts to remedy that shortcoming have not exactly appealed to the general philosophical public.

We conclude that within Carnapian systems we cannot formulate hypotheses on possible patterns in the data, let alone change or introduce them. In the following sections it will be seen that Bayesian statistics, as well as classical statistics, does better than the Carnapian inductive system on the count of both fixity and poverty.

2.5 Popper on the Problem of Induction

Before turning to statistics, we deal with another important contributor to the debate on inductive inference, Karl Popper [17] . Popper's views on induction can be explained most easily in conjunction with his position in the debate on the demarcation of science from pseudo-science. Popper rejected the view of the logical empiricists, who argued that science is defined by its roots in empirical fact and their logical implications, stating instead that falsifiability is the distinguishing feature of science. According to Popper, the hallmark of good science is that it puts itself at risk of being proven wrong. It generates distinct predictions that can be checked against the empirical facts and can subsequently be proven false. So, for example, the claim that the sun will rise tomorrow is scientific, because tomorrow we may find out that the sun has not risen, thus proving it wrong. The claim, on the other hand, that the sun

will never rise anymore, is not scientific, because at any point in time we must leave open the possibility of a future rising.

Popper's views on inductive inference can be seen as the continuation of this line of thought. From the view that claims are only scientific in virtue of their possible falsification, it is a small step to the view that the only claims that can be considered genuine scientific knowledge are those that result from falsification . So according to Popper, we cannot base any knowledge on inductive inference. In the example, we cannot conclude anything towards future occasions of a rising sun from the fact that up until now the sun has always risen. As Popper would say, the theory that the sun will always rise is not yet disproved. But at best this motivates us to go on checking the claim that the sun will always rise. If, on the other hand, the sun does not rise tomorrow, science has truly advanced, because at that point we can be certain that the claim that the sun will always rise is false, and hence that the claim that on some day the sun will not rise is true. In short, Popper argues that inductive inferences towards general claims cannot provide us with scientific knowledge, but that deductive inferences towards the denial of general claims do provide knowledge. Deductive inference is valid, but inductive inference is not.

In our DID example, the question is how can we generalize towards a conclusion on the existence of DID, on the basis of observations of the memory performance of DID-patients and DID-simulators. Now does Popper allow us to conclude that DID patients are universally better in memory performance than DID-simulators and therefore that DID is a genuine disorder? Bypassing the further difficulty that in the DID example the theory is cast in terms of probabilities, and that probabilistic statements can strictly speaking never be proven false, Popper would argue there is never any positive evidence for such a general statement, let alone for concluding that DID is a genuine disorder. We can only conclude, by means of a single counterexample, that such a general statement is not true. So after our observation of a difference between the two groups of DID-patients and DID-simulators, the theory that DID is a genuine disorder is not disproved by the data and therefore the theory, for the time being, is not rejected. But it is not proven by the data either.

Admittedly, this is a rather critical view on inductive inference. Popper's position has aptly been named critical rationalism. But as the term rationalism suggests, the views of Popper also have a more positive part which is of interest to the present discussion. While Carnap put the starting point of scientific knowledge in primitive empirical facts, as captured in a formal language, Popper put forward the view that science always starts with a hypothesis, some bold claim or general statement, which we may subsequently attempt to falsify. He referred to this as the searchlight theory of knowledge: the realm of empirical fact can provide some kind of knowledge, but the researcher has to provide a searchlight, more specifically a guiding hypothesis via which this realm can make itself known. Put differently, it is not the observations that come to us with their own message, rather we take the initiative

to seek out the observations to meet our own interest. Against the empiricist and inductivist views of Carnap, Popper's views show a marked rationalist tendency, in the fact that the mind rather than the world is the first cause in the production of knowledge.

Summing up, we have dealt with two very different views on the problem of induction in the foregoing. In the next two sections, we shall argue that statistical inference occupies a middling position between the two views, and that both Popper and Carnap fail to capture an important aspect of the solution inherent to statistical inference. On the one hand statistical inference is inductivist, because it allows us to learn from the data. And on the other hand it is rationalist, because what is learnt from the data is entirely determined by the statistical model that we choose.

3 Bayesian Inference as Deduction

The discussion on Carnap and Popper makes clear that the opinions on how to justify inductive inference diverge widely. Because Bayesian statistical inference is a way of dealing with inductive inference as well, the question arises how it might be positioned relative to these diverging opinions. In the next two sections we will argue that Bayesian statistical inference contains both falsificationist and inductivist elements. More in detail, in this section we show that the methodology of Bayesian statistical inference can be spelled out by framing these inferences in a probabilistic logic , following ideas of Howson [13] [14] and Romeijn [18] [19]. It will become apparent that Bayesian inference is similar to deductive inferences . This will lead to a discussion of model selection procedures in the next section, which will reveal the position of Bayesian statistical inference in the spectrum between Carnap and Popper.

3.1 Deductive and Inductive Inference

Let us briefly compare deductive and inductive logic. Recall that in deductive logic, an argument is valid if the truth of its premises guarantees the truth of the conclusion. So a perfectly valid argument might lead to a false conclusion, on the ground that one of its premises is false. Take for example the premises that all apples are fruit, and that all fruit grows on bulldozers. By deductive inference, we therefore validly conclude that all apples grow on bulldozers, even while this is most certainly not true. Deduction serves to explain and rearrange our knowledge without adding to its content. Inductive inference, by contrast, seems to add to the content of our knowledge. We obtain observations, and then amplify and generalize them to arrive at general conclusions. So an important difference between deduction and induction seems to be that while deduction is conceptually closed and only brings out the conclusions already present in the premises, induction adds to the content of the

premises. As a result of this, conclusions obtained with inductive inferences do not necessarily have the same degree of certainty as the initial premises.

Nevertheless we will investigate in this section the parallel between deductive and inductive inference. To do so, we will first study a specific deductive argument, and after that we will introduce an argument in Bayesian logic that can be seen as the inductive counterpart to the deductive argument. The example of deductive inference that we will study is the so-called proof by contraposition:

If H, then E (premise 1).
E is false (premise 2).
Therefore, H is false (conclusion).

To examine this inference in more detail, we will make use of the DID-example we discussed earlier. The analogy between deductive and Bayesian inference suggests that just like the deductive inference, Bayesian inference is valid.

3.2 Deduction in the DID-example

The full design of the study of Huntjes *et al.* allowed the authors to compare estimations of the overall memory capacities of DID-simulators (μ_{sim}), true DID-patients (μ_{pat}), true amnesiacs (μ_{amn}), and controls (μ_{con}). We can formulate many different general models concerning the latent memory capacities of these groups, for example:

- $M_0 : \mu_{sim} < \mu_{pat} = \mu_{amn} < \mu_{con}$
- $M_1 : \mu_{sim} = \mu_{pat} < \mu_{amn} < \mu_{con}$
- $M_2 : \mu_{pat} = \mu_{con} = \mu_{sim} = \mu_{amn}$
- $M_3 : \mu_{pat} > \mu_{con} > \mu_{sim} < \mu_{amn}$
- ...

Note that this is a list of models, and not of general hypotheses. The statement that $\mu_{sim} < \mu_{pat} = \mu_{amn} < \mu_{con}$, for example, is consistent with a large number of different valuations of these parameters, and each of these valuations presents a separate hypothesis. So the statement concerns a set of hypotheses, or a model for short.

For convenience we will make the example of the present section a bit easier. First of all, we will abstract away from the parameters μ_{amn} concerning amnesiacs, and μ_{con} concerning people from the control group. Second, in this section we will not deal with models but rather with specific hypotheses, that is, specific valuations for the parameters μ_{pat} and μ_{sim}. Third, we are restricting attention to two hypotheses in particular, H_0 and H_1. For H_0 we choose particular values of the parameters such that $\mu_{pat} > \mu_{sim}$, while for H_1 we choose them such that $\mu_{pat} = \mu_{sim}$. Moreover, we assume for the time being that one of these two hypotheses is true and thus that all the other hypotheses are false, or in logical terms: $H_0 \vee H_1$, where the symbol $\vee$ can be read as 'or'. This expression is the first major premise in the deductive

argument below. Note also that from their definitions, the hypotheses H_0 and H_1 are mutually exclusive, so that $\neg(H_0 \wedge H_1)$, where $\neg$ means 'not' and $\wedge$ can be read as 'and'. This will turn out to be convenient in the representation of the hypotheses below, but we will not use this premise in the argument.

So the inference concerns the two rival hypotheses H_0 and H_1. The empirical evidence, as for instance provided in the study of Huntjes *et al.*, is now used to adjudicate between these two hypotheses. First, we concentrate on a specific empirical difference between these two hypotheses, namely that according to H_0 DID-simulators have a worse memory performance than true DID-patients, while according to H_1 the DID-simulators and true DID-patients have equal capacities. Accordingly, the relevant observations are the scores of members of the two groups, patients and simulators, on some memory test. We might for example find that the difference of the scores of the two groups exceeds a certain threshhold, denoted E, or otherwise we might find that it does not exceed the threshhold, denoted $\neg E$. For the purpose of this example, we suppose that the test scores can tell apart the hypotheses unequivocally: if H_0 is true, then we are certain that the difference in scores on the memory test exceeds a certain threshhold, or in logical parlance, $H_0 \rightarrow E$.

We can specify so called truth values for each combination of hypotheses and evidence, based on the premises of the above. It will be convenient and insightful to represent these premises as truth valuations over all the logical expressions that we can conceive; see the squares of Figure 1. As further explained in the caption, the truth values in the quadrants indicate whether the corresponding logical possibilities, or cells in the grid, are consistent with the premises. More specifically, given some truth valuation over the logical possibilities, we say that a proposition is true if and only if it is true in each of the cells that is assigned a 1. The premises $H_0 \vee H_1$ and $\neg(H_0 \wedge H_1)$ are worked out in the first two squares of Figure 1. They are in a sense implicit to the presentation of the truth valuations in the rightmost square of Figure 1, in which H_0 and H_1 are put side by side as mutually exclusive and jointly exhaustive possibilities.

The latter square also expresses how the hypotheses H_0 and H_1 relate to the data E. According to deductive logic, all the entailment $H_0 \rightarrow E$ says is that we cannot have the combination of H_0 being true yet E being false, so $H_0 \rightarrow E$ is equivalent to $\neg(H_0 \wedge \neg E)$. In sum, the three quadrants of the rightmost square that contain a 1 are the only logical possibilities consistent with the premises.

With this graphical representation of the premises in place, we can bring in the further premise presented by the observations. Say that we observed that the scores of the two groups on the memory test are slightly different, but that the difference does not exceed the given threshhold, so $\neg E$ receives a truth value of 1. In Figure 2, the corresponding truth values can be seen in the middle square. The observation itself does not involve the hypotheses, and therefore $H_0 \wedge \neg E$ and $H_1 \wedge \neg E$ receive the truth value 1, and $H_0 \wedge E$ and $H_1 \wedge E$ receive the truth value 0. So the square on the left and in the middle of Figure 2

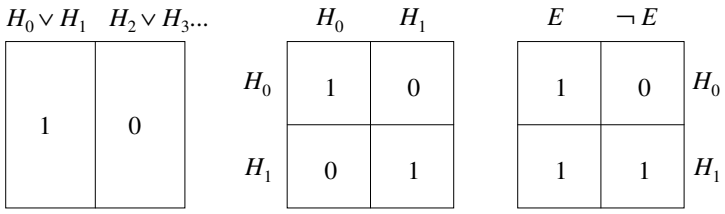

Fig. 1. These three squares indicate summarize the premises of the logical argument. The leftmost square indicates that $H_0 \vee H_1$, and thus that all other hypotheses H_i for $i > 1$ are deemed false. The middle square indicates that $\neg(H_0 \wedge H_1)$, by setting the quadrants in which H_0 and H_1 overlap to 0. Finally, the rightmost square indicates that $H_0 \rightarrow E$, which is equivalent to $\neg(H_0 \wedge \neg E)$. The three quadrants labelled 1 in the rightmost square are the only logical possibilities consistent with the premises.

express the two main premises, one concerning the hypotheses, stemming from Figure 1, and one concerning the observations. The beauty of the graphical representation is that combining these premises is a straightforward operation on the truth valuations: we simply multiply the truth values of the two input premises, as expressed in the square on the right of Figure 2.

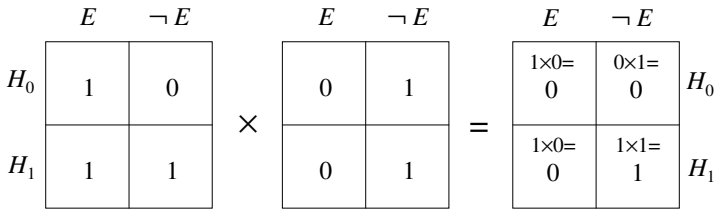

Fig. 2. This calculation with squares summarises the logical argument that runs from the premises given previously, and the additional premise that $\neg E$, to the conclusion of H_1. The leftmost square is equivalent to the rightmost square of Figure 1. The middle square expresses the premise $\neg E$. The truth values in the rightmost square are obtained from the values in the other two squares by multiplying the values in each of the quadrants.

After combining the premises, we see that only $H_1 \wedge \neg E$ receives a truth value of 1. All the other cells have a truth value 0. We can therefore conclude all propositions that include the specific cell $H_1 \wedge \neg E$. Of course we may conclude $\neg E$, but this is hardly surprising, because it was also one of the premises. However, we may also conclude H_1. Via $\neg E$ and $H_0 \rightarrow E$ we learn that H_0 cannot be true, so $\neg E$ falsifies H_0, and by $H_0 \vee H_1$ we can derive that H_1 must be true. We can conclude that the DID-simulators and true DID-patients have equal capacities on memory performance.

3.3 Choosing a Model

In the previous subsection we used deductive inference to derive a conclusion from the premise concerning a finite set of hypotheses, and the premise on how the hypotheses relate to evidence of the observed memory performance, and finally a premise expressing what evidence we received. In this subsection and the next, we will use essentially the same premises, with minor revision as will be explained later, to derive a conclusion by means of Bayesian inference . The conversion has two aspects, namely the use of probabilistic valuations and of Bayes' theorem. In this subsection we will deal with the former.

Apart from providing us with a convenient way of representing the operation of combining premises, the graphical representation of Figures 1 and 2 can be used to illustrate the parallel between deductive and Bayesian logic, which we consider very telling. First consider the graphical representation itself. As in the case of deductive inference, we take the logical possibilities provided by the hypotheses and the evidence as starting point. We distinguish between E and $\neg E$, and similarly we consider hypotheses H_j with $j = 0, 1, 2, \ldots$. Now we want to connect these logical possibilities to probability theory, which is according to the standard axiomatization a function over sets, and hence we are taking the logical possibilities as sets as well. The logical possibility H_0 is the set of all those imaginable or possible worlds in which the hypothesis H_0 is true, and similarly, E is the set of all those possible worlds in which the observation E occurred. Accordingly, instead of $H_0 \wedge E$ we will write $H_0 \cap E$. That is, instead of working with the logical operation $\wedge$, from now on we use the set-theoretical operation of intersection. Similarly, we will write $\neg E$ as $\bar{E}$, the set-theoretical complement of E.

Next consider the inference concerning the logical possibilities. Recall that the idea of deductive inference was to find a truth valuation of certain proposition, based on the truth valuations of a combination of premises. Again, Bayesian inference does roughly the same. The key difference between deductive and Bayesian logic is that Bayesian logic does not use truth values of 0 and 1, as does deductive logic. Rather it uses probabilistic valuations p, that is, valuations of logical possibilities within the interval $[0, 1]$ and satisfying the axioms of probability theory. So the cell $H_0 \cap E$ in the space of logical possibility receives some probability, $p(H_0 \cap E) = 2/5$ for instance. The probability values of all the cells must sum to 1. But apart from that difference in valuation function, the workings of Bayesian logic will turn out to be very similar to the workings of deductive logic. Just like deductive logic, Bayesian logic computes probabilistic conclusions on the basis of probability assignments over logical possibilities.

Let us have a look at the above deductive inference to make the above claims precise. The first premise in the foregoing is that we restrict ourselves to two hypotheses, H_0 and H_1. We assigned a truth value of 1 to $H_0 \vee H_1$, so that we ruled out all the H_j for $j > 1$. In Bayesian logic, we can do the same by assigning all probability to the hypotheses H_0 and H_1, $p(H_0 \cap H_1) = 1$.

That is, only these two hypotheses receive a probability and the remaining hypotheses $H_2, H_3, \ldots$ receive a probability of 0. But note that this probability assignment is not yet specific enough: we still have many ways of allocating the probability among the two hypotheses H_0 and H_1. On the basis of a symmetry argument, we might distribute the probability evenly: $p(H_0) = p(H_1) = 1/2$.

We have now chosen the hypotheses, but we have not determined the probability assignment on the level of logical possibilities. Both the hypotheses H_0 and H_1 might allow for the occurrence of the observations E and $\neg E$, and we need to specify the probability valuations of these cells. Recall that in the deductive case we said that $H_0 \wedge \neg E$ was impossible. This was admittedly a rather strong assumption: normally test results cannot outright falsify any hypothesis, rather they make hypotheses more or less likely. By using the probability valuations, we can make such weak relations between observations and hypotheses precise. If H_0 is true we think it is far more probable than not that the difference between the DID-groups on memory performance exceeds the threshold , but this need not be strictly implied. So we might specify that conditional on H_0 being true, E is 4 times more likely than $\bar{E}$, so that $p(E|H_0) = 4/5$ and $p(\bar{E}|H_0) = 1/5$. Similarly, if H_1 is true we might consider it somewhat less probable than not that the difference between the DID-groups on memory performance exceeds the threshold, so we might specify $p(E|H_1) = 2/5$ and $p(\bar{E}|H_1) = 3/5$.

Together with the probability assignment over H_0 and H_1, we have thereby fixed the probability assignment for all the logical possibilities. We can compute $p(H_j \cap E) = p(H_j)p(E|H_j)$ and similarly $p(H_j \cap \bar{E}) = p(H_j)p(\bar{E}|H_j)$. This leads to the probability assignment over the logical possibility presented in the left square of Figure 3.

	E	$\neg E$
H_0	2/5	1/10
H_1	1/5	3/10

E	$\neg E$	
4	1	H_0
2	3	H_1

Fig. 3. The square on the left represents the probability assignment over the logical possibilities in terms of probability mass. The square on the right provides the same information in terms of odds.

The square on the right side of this figure effectively depicts the same probability assignment, but written down in terms of odds. The difference is that the odds do not have to add up to 1. Only their ratios matter. In the following we will only make use of the odds.

Finally, we want to point to the relation of the above with Bayesian statistics as we know it. In the foregoing we chose two hypotheses, defined the probabilities of the observations conditional on them, and we chose the probabilities of the hypotheses themselves. In Bayesian statistics, this comes down to the choice of a model, or a set of possible statistical hypotheses, then the definition of a likelihood function for each of the hypotheses in the model, and the determination of so-called prior probabilities. Of course, statistical models are normally much more complicated and elaborate, but the general idea remains the same.

3.4 Bayesian Inference

As indicated, we are drawing an analogy between deductive inference and Bayesian inference. It will be clear that the determination of probabilities, or odds, over the logical possibilities in Figure 3 runs parallel to the first of the two main premises in the logical argument, as summarized in Figure 1. Now the second premise of the Bayesian inference is almost the same as the one we used for deductive inference. We observe $\bar{E}$, and in the deductive example, $\neg E$ therefore receives a truth valuation of 1. In Bayesian inference, as will be seen, we will say that the adapted probability for $\bar{E}$ must be 1. The question is how the addition of this premise reflects on the probability assignment over the logical possibilities, as given in Figure 3. In particular, how is the adapted probability distributed between the hypotheses H_0 and H_1?

Note first that the new premise is, strictly speaking, in contradiction with the probability assignment already given. We have $p(\bar{E}) = p(H_0 \cap \bar{E}) + p(H_1 \cap \bar{E})$ and hence $p(\bar{E}) = 2/5$. To express the probabilities after we observed $\bar{E}$, we must therefore make use of a so-called posterior probability assignment, which we will denote with $p_{\bar{E}}$. This is a new probability assignment, that is consistent with assigning $\bar{E}$ unit probability. To obtain the posterior probability assignment from the prior one, we can use the combination of Bayes' rule and Bayes' theorem:

$$p_{\bar{E}}(\cdot) = p(\cdot|\bar{E}) = p(\cdot)\frac{p(\bar{E}|\cdot)}{p(\bar{E})}. \tag{4}$$

Bayes' theorem is given by the second equality. It is a theorem of probability theory, and as such very hard to argue with. The interesting and contentious equality is the first one, which we might call Bayes' rule. Note that it is not a theorem of probability theory. Rather it relates two different probability functions, the prior distribution p and the posterior distribution $p_{\bar{E}}$, and thus expresses how we must adapt the probabilities if we add the further premise $\bar{E}$. In other words, Bayes' rule expresses how we can construct a new probability assignment $p_{\bar{E}}$ which incorporates the fact that we assign a probability 1 to the data $\bar{E}$, based on the old probability assignment p, in which the data $\bar{E}$ had a probability smaller than 1.

Now let us compute some posterior probabilities, based on the fact that we have $p_{\bar{E}}(\bar{E}) = 1$. By Bayes' rule, we can compute the posterior probability for the hypotheses H_0 and H_1 on the basis of the prior probability and the likelihoods. For H_1 we find

$$p_{\bar{E}}(H_1) = p(H_1|\bar{E}) = p(H_1)\frac{p(\bar{E}|H_1)}{p(\bar{E})} = 1/2 \times \frac{3/5}{2/5} = \frac{3}{4}. \tag{5}$$

In words, the observation that $\bar{E}$ leads to a posterior probability for H_1 that is higher than the prior probability. In this sense at least, Bayesian inference mimics the deductive inference, where $\bar{E}$ also favored H_1. But why are we to believe the posterior probabilities arrived at by means of Bayesian inference?

We will now argue that there is a much more genuine sense in which the Bayesian inference resembles the deductive inference. This resemblance provides us with a reason to believe that the posterior probabilities are in a sense the correct probabilities for the hypotheses after the observation of $\bar{E}$. As Figure 4 illustrates, if we represent the probability valuations as odds, we can combine the two main premises of the Bayesian inference in exactly the same way as in deductive inference.

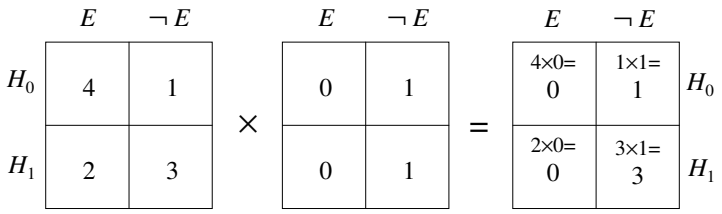

Fig. 4. This calculation with squares summarises the Bayesian statistical inference. The leftmost square is equivalent to the square on the righthand side of of Figure 3. The middle square expresses the premise $\neg E$. The odds in the rightmost square are obtained from the values in the other two squares by multiplying the values in each of the quadrants.

It is not a coincidence that the results of this operation are the odds that correspond to the posterior probabilities arrived at by Bayesian inference. Changing the probability assignment in accordance with the observation $\bar{E}$, as laid down in Equation (4), is nothing but the rescaling of the probabilities to the proportions of the probabilities within $\bar{E}$. This is exactly what the formula does. Bayes' rule allows us to "zoom in" on the probability assignment over the hypotheses within $\bar{E}$.

Thus Bayesian inference is like deductive inference in two important respects. Firstly, they both make use of a valuation function over a set of elementary logical possibilities, although there are also differences here. As for these possibilities, in the case of deductive inference they are maximally specific

propositions, and in the case of Bayesian inference they are sets of possible worlds. And as for the valuations, in the case of deductive inference they are truth valuations, and in the case of Bayesian inference they are probabilities. Secondly, and most notably, the operation for combining a valuation with a further premise, in particular with an observation such as $\bar{E}$, is exactly the same. Rather suggestively, we might say that Bayesian inference is therefore valid in exactly the same way as that deductive inference is.

Although it has been noticed before, we want to emphasize again that the above example is nothing like a serious Bayesian statistical inference: usually the model contains many more statistical hypotheses, and there are normally many more possible observations, or elements in the sample space. However, the inferential steps are exactly the same. In Bayesian statistics we choose a model, fix the likelihoods of the hypotheses in the model, and finally determine a prior. Then we collect data and incorporate these data in the so-called posterior probability assignment over the model by means of a Bayesian update. We therefore maintain that the above example tells us something about Bayesian statistical inference in general.

3.5 Summing up

We have discussed how to derive a conclusion based on a set of premises, first by using deductive inference, and then by using Bayesian inference. We have shown that Bayesian inference follows roughly the same procedure as deductive inference. This suggests that Bayesian inference, like deductive inference, is valid. That is, if the premises are true, then so is the conclusion. In the following we will elaborate how these ideas may be used to position Bayesian statistics in the philosophical debate over statistics, and in particular how they can be applied to the Bayesian model selection described in Section 1.

4 Model Selection

We have seen that Bayesian statistics can be provided with a philosophical underpinning by portraying it as a logic. Against this backdrop we will now explain how statistics, and Bayesian statistics in particular, unites the views of Carnap and Popper on induction. This may well raise some eyebrows: in what sense do we do justice to Popper's views when we redistribute probability over a number of hypotheses in the light of data? Recall that an important aspect of Popper's view is falsificationism ,which states that we can only learn from data if the data rule out some hypothesis. Bayesian statistical inference goes much further than that, because it allows us to learn positive facts from the data. Nevertheless, in the following we will argue that in some important respect Bayesian statistical inference retains the rationalist spirit.

4.1 Models as Uniformity Assumptions

The foregoing already indicated that in a Bayesian inference, the choice for a model can be understood as the choice of a certain kind of premise. We drew a parallel between, on the one hand, the choice for a prior restricted to $p(H_0) = p(H_1) = 1/2$ together with the likelihood functions of both hypotheses, and, on the other hand, the choice for $H_0 \vee H_1$ together with $H_0 \rightarrow E$. In this subsection we will investigate this parallel further. In particular, we will relate the choice of a certain model, as elaborated in Section 3, with the choice of a certain set of projectable predicates, as discussed in Section 2.

Let us return to the nature of inductive inference as it was illustrated in the example of Section 2.1. It can be noted that the inference from statement 1 to statement 2, by itself, seems to miss a component. It is more or less implicit in the inference that what has happened in the past can be expected to happen in the future. As we discussed, one possible take on the problem of induction is that this component must be added to the inductive inference as an explicit premise. At first glance this premise might simply be that the world is a boring place, and that the same events will keep repeating themselves. But it was easily seen that simply adding this premise cannot solve the problem: we ran into predicates like Grue. As exhibited clearly in the inductive logical systems devised by Carnap, if we want to infer anything inductively, we must choose the exact set of predicates with respect to which the world is boring, that is, the predicates that are supposed to stay constant. In philosophical parlance, we must select the projectable predicates.

There is a rather nice formal relation between the Carnapian systems and Bayesian statistical inference, which has an immediate bearing on this point. Note first that the c-function of Equation (3) only depends on the number of earlier results, n_0 and n_1, and not on the exact order in which these results were observed. Inductive logical systems with this property are called exchangeable. Famously, De Finetti [8] proved that any exchangeable inductive logical system can be represented as a Bayesian inference over a particular model, namely the model of binomial hypotheses, and furthermore that every prior over this model singles out a unique exchangeable system. As in the foregoing, we write Q^1_{n+1} for the result of person $n+1$ scoring above chance level in a memory test, meaning that this person scored better than the expected score of filling in the test randomly. We denote the binomial hypotheses with H_θ. These hypotheses have the following likelihoods:

$$p(Q^1_{n+1}|E_n \cap H_\theta) = \theta. \tag{6}$$

This means that all test results are independent and identically distributed. The model of binomial hypotheses, which features in De Finetti's representation theorem, includes all these hypotheses: $\{H_\theta : \theta \in [0,1]\}$. It can be proved that prior probability functions of the form $p(H_\theta) \sim \theta^{\gamma\lambda-1}(1-\theta)^{(1-\gamma)\lambda-1}$ lead to the Carnapian inductive systems of Equation 3). That is,

$$c(Q^1_{n+1}, E_n) = \int_0^1 p(Q^1_{n+1}|H_\theta \cap E_n)p(H_\theta|E_n)\, d\theta, \tag{7}$$

in which $c(Q^1_{n+1}, E_n)$ is the expected value of the response of subject $n+1$ given earlier responses E_n, $p(Q^1_{n+1}|H_\theta \cap E_n)$ is the likelihood of the hypothesis H_θ for the event of this subject scoring above chance level, and $p(H_\theta|E_n)$ is the posterior probability over all hypotheses H_θ in the model of binomial hypotheses, given the earlier responses E_n. The interested reader may consult Festa [9] for further details on this. For present purposes it is only important to remember that Carnapian inductive systems can be replicated in a Bayesian inference.

This mathematical fact provides us with crucial insight into the nature of choosing a model. Recall that in the Carnapian system, the choice of the predicates, in this case scoring above, on, or below chance level in the memory test, effectively determined the projectable or stable pattern in the data: the observed relative frequencies of scoring were supposed to be indicative of the scoring of future subjects. But we can identify exactly these projectable patterns in the statistical model that, according to the representation theorem, underpins the Carnapian system. For each of the binomial hypotheses the probability of scoring above chance level is stable and constant over time. The choice for this specific set of hypotheses, or this statistical model for short, is effectively the choice for a set of projectable predicates, namely the chance for scoring over a certain level is stable and constant over time. In our view this is exactly the function of choosing a model as part of a Bayesian statistical inference: to fix the starting point, namely the set of hypotheses and the associated probabilistic patterns, so that the data are allowed to select the most fitting one.

The choice for a specific model, or for specific hypotheses to be part of the model, reflects the interest and often the background knowledge of the researcher. But this also means that a researcher can help herself to more informative conclusions by choosing her hypotheses well, and similarly that she can ruin it by choosing her model badly. For instance, she might choose for the gruesome variants of the binomial hypotheses introduced in the above:

$$p(Q^1_{n+1} = 1|E_n \cap G_{N\theta}) = \begin{cases} \theta & \text{if } n < N, \\ (1-\theta) & \text{if } n \geq N. \end{cases} \tag{8}$$

In words, the hypotheses $G_{N\theta}$ dictate that up until the N-th observation Q^q_n for $n < N$, the probability for $q = 1$ is θ, but that for $n \geq N$ the probability for $q = 1$ is $(1-\theta)$. We might take the model $\{G_{N\theta} : \theta \in [0,1]\}$ for some large N, choose a uniform prior $p(G_{N\theta})d\theta = 1$, and then start updating with observations of subjects doing the memory test. For values of $n+1 < N$ the choice of this model leads straightforwardly to the Carnapian prediction rule of Equation (3), with $\gamma = 1/2$ and $\lambda = 2$. Now say that by far most subjects $i < N$ pass the test, so that $n_1 >> n_0$. Using the Carnapian system and

assuming that $n < N$, we have $p(Q^1_{n+1}|E_n) \gg p(Q^0_{n+1}|E_n)$. But what can we predict for the subject indexed i with $i > N$ on the basis of E_n? Because of the sudden reversal in the likelihood functions of the hypotheses, we effectively swap the places of scoring on or above chance level in the prediction, so on the basis of a large majority of people exceeding chance level in E_n, we predict that subjects $i > N$ will most likely fail! Or in mathematical words, $p(Q^1_i|E_n) \ll p(Q^0_i|E_n)$ for $i \geq N$.

We saw in the example on gruesome predicates of Section 2.2 that the wrong predicate choice may lead to useless predictions, and we have here seen that the same holds for the choice of models, thus indicating how the choice of a certain model resembles the choice of a projectable predicate. Bayesian statistical inference therefore has, at least, this one distinct Carnapian streak: it allows for inductive inference on the basis of a specific uniformity assumption.

4.2 Models as Searchlight

In the foregoing we claimed that Bayesian statistical inference occupies a middle position between Carnap and Popper. Partly the link with Carnap has now been made clear, and so we turn to the relation with Popper, in particular with his searchlight theory of knowledge alluded to in Section 2.5 .

We first identify this searchlight theory in Bayesian statistical inference, which will point us to an important difference between Carnapian inductive systems and Bayesian statistical inference. We have already seen how both make use of specific uniformity assumptions. However, in the case of Carnapian systems there seems to be very little by way of actively choosing, let alone comparing the assumptions. In the views of Carnap, the choice for a language, and thus the uniformity assumptions inherent to it, is a precondition for dealing with the problem of induction in terms of a logic. In fact, according to Carnap [3], it is a precondition for dealing with philosophical problems in general. So it seems that for Carnapian systems, the choice for a specific uniformity assumption is beyond the reach of logical analysis. By contrast, in Bayesian statistical inference the choice for a uniformity assumption, by choosing a model, is an explicit part of the logical account. As also argued in the foregoing, the choice of a model determines the type of probabilistic pattern that we can identify in the data. In other words, it provides us with a searchlight looking at the data. The explicit choice for a model signals a rationalist tendency in Bayesian statistical inference. The origin of empirical knowledge is not naked observation, but observation within the context of a theoretical starting point, namely a model.

We may wonder whether we can extend the parallel between the Popperian view on induction and Bayesian statistics, in particular whether Bayesian statistics presents us with a notion of falsification . To answer this, consider the probabilistic inference in the example of Section 3.4. We might argue that this Bayesian inference already exhibits a weak form of falsification: the hypothesis H_0 is proved unlikely by the data, and so we may decide to discard that

hypothesis, or at least not use it in predictions or decision making. However, apart from the fact that low probability is not the same as logical impossibility, the use of the specific model $\{H_0, H_1\}$ determines that either one of them will accumulate most probability in the light of the data. So by discarding H_0 we can infer H_1. Therefore, discarding H_0 is not a falsification of the starting point of the inference, namely the model $\{H_0, H_1\}$.

In the DID-example, it may happen that we find further data E' for which $p(E'|H_1)$ is very small, so that H_1 fits badly with the data as well. In such a case the whole model fits badly with the data. Similarly, in the example concerning the hypotheses $G_{N\theta}$ of the preceding Section, we may observe further subjects $i \geq N$ doing the memory test. On the basis of our model choice and the fact that subjects $i < N$ performed very well, we expect these new subjects to perform very badly. But it may certainly happen that the subjects $i \geq N$ perform very well. We then want to conclude that something was amiss with the model choice, e.g. that the true hypothesis is not to be found among the hypotheses in the model. Note also that such cases do not allow us to draw any positive conclusions: we just conclude that none of the hypotheses in the model is any good, and that some unspecified other hypothesis would have been better. Such cases of bad model fit come a bit closer to the idea of falsification in Popper. Now we want to emphasize immediately that finding a bad model fit is not the same as definitively falsifying the model, in the same way as that finding a low probability for H_0 is nothing like logically deriving the falsehood of H_0. Low probability, or even zero probability for that matter, is entailed by, but does not entail logical impossibility. Still, the closest we can get within Bayesian statistical inference to the idea of falsification is the idea of bad model fit.

However, the falsification of a model is not an integral part of the Bayesian inference machinery. The model can be chosen explicitly in the Bayesian inference, by distributing prior probability over a restricted set of hypotheses. But the tools of Bayesian inference do not allow for changes to that initial choice for a model. In the words of Dawid [7], the Bayesian is "well-calibrated": inherent to the choice of set of hypotheses, i.e. a model, is the assumption that the true hypothesis is among them. It is impossible to change this assumption without after the fact changing the prior probability, which is a non-Bayesian move. Of course we can change a statistical model in a controlled and rational way, by turning to model selection techniques [1]. There are various criteria for model fit, and various ways of off-setting model fit against the complexity of models. But with the exception of the Bayesian information criterion, the standard model selection techniques do not take the explicit form of a Bayesian inference. And even the Bayesian information criterion only employs an approximation of posterior model probabilities.

4.3 Bayesian Model Selection

We are now ready to present Bayesian model selection , as it was presented in Section 1, against the philosophical background of Bayesian statistics . Concerning this philosophical background, we argued that it combines the inductivist view of Carnap with the falsificationist view of Popper. As in the work of Carnap, Bayesian statistics allows us to reason inductively from the data by assuming that certain data patterns, summarized in a model, are invariant. But this is only possible once we have made a specific selection of hypotheses to begin with, and in this sense Bayesian statistics also have a marked Popperian component. In the same line, the assessment of a model against the data runs parallel to falsification in the view of Popper.

How does Bayesian model selection fit into this background? It is important to keep clear on the roles of models and hypotheses here. Bayesian model selection deals with the assessment of model fit, that is, with the fit of a collection of statistical hypotheses. It therefore extends the reach of standard Bayesian statistical inference, which concerns the fit of specific statistical hypotheses once the model is given. On the other hand, in Bayesian model selection the rival models are understood as statistical hypotheses themselves. That is, they are somehow understood as claims about patterns in the data, as expressed in a likelihood function. These likelihood functions are not straightforwardly defined, as they are in the case of a normal Bayesian statistical inference. They are so-called marginal likelihoods , because they involve the likelihoods of the hypotheses inside the rival models. Bayesian model selection is thus similar to standard Bayesian statistical inference, in the sense that rival models are treated as if they were normal statistical hypotheses. This makes Bayesian model selection very attractive: it benefits from all the arguments standardly given to support Bayesian statistical inference. However, the key difference also leads to some problematic aspects, to which we will now turn.

5 A challenge for Bayesian Model Selection

This section discusses some problematic aspects of applying Bayesian inference to models. These aspects relate directly to the philosophical background for Bayesian statistical inference, as provided in the preceding Sections. Firstly we have a closer look at the fact that in Bayesian model selection, models are conceived as hypotheses. Secondly, we ask how to understand the probability assignments over models. First we provide a tentative solution, but it will be seen that this solution puts more weight on the second problem. The section ends with a challenge to the proponents of Bayesian model selection.

5.1 Models as Hypotheses?

To illustrate the first of our two concerns, it is useful to recollect a well-known finding from the psychology of reasoning, concerning the so-called conjunction

fallacy . In an experiment done by Tversky and Kahneman [20], subjects were presented with the following story:

> Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations. Which is more likely?
> 1. Linda is a bank teller.
> 2. Linda is a bank teller and is active in the feminist movement

Rather surprisingly, a majority of normal subjects think the second fact to be the more likely one. This is odd, because the axioms of probability do not allow a conjunction to be more probable than either of its conjuncts: it is a theorem that $p(E \wedge E') \leq p(E)$ for any pair of events or facts E and E'. Clearly, people do not follow the axioms of probability in their intuitive judgements of likeliness.

Next, consider the example of Bayesian model selection in Section 1, and in particular the two models that are being compared, $\mathcal{M}_0$ and $\mathcal{M}_1$. Recall that both models consisted of the same hypotheses $H_{\mu_{pat}\mu_{sim}}$, that $\mathcal{M}_0$ contained all these hypotheses, and that the model $\mathcal{M}_1$ was subject to the further constraints that $\mu_{pat} < \mu_{sim}$. At first sight, this situation is completely identical to the situation with Linda the bank teller. We may write the model $\mathcal{M}_1$ as a conjunction of facts, namely the model $\mathcal{M}_0$ and the further fact that $\mu_{pat} < \mu_{sim}$. This fits well with the fact that the set of hypotheses associated with $\mathcal{M}_1$ is strictly included in the set of hypotheses associated with $\mathcal{M}_0$. It is, under closer scrutiny, truly remarkable that a set that is strictly included in another set can nevertheless have a larger probability. Is Bayesian model selection implicitly violating the axioms of probability ?

The reader will be relieved to find that the answer to this question is negative. To explain this, we simply need to cast the comparison of both models and hypotheses in a different set-theoretical framework, as illustrated in Figure 5. As we have conceptualized the two models $\mathcal{M}_0$ and $\mathcal{M}_1$ in the above, they are overlapping sets. Even stronger, all elements $H_{\mu_{pat}\mu_{sim}}$ in $\mathcal{M}_1$ are also a member of $\mathcal{M}_0$. However, nothing prevents us from using two distinct sets of hypotheses, labeled $H_{0\mu_{pat}\mu_{sim}}$ and $H_{1\mu_{pat}\mu_{sim}}$, which are different from a set-theoretical point of view by virtue of being labeled differently, even while they have exactly the same likelihood functions over the data. The model $\mathcal{M}_0$ consists of the hypotheses $H_{0\mu_{pat}\mu_{sim}}$, while the model $\mathcal{M}_1$ consists of the different hypotheses $H_{1\mu_{pat}\mu_{sim}}$. The model $\mathcal{M}_1$ is further restricted by the fact that $p(H_{1\mu_{pat}\mu_{sim}}) = 0$ if $\mu_{pat} \geq \mu_{sim}$. In this framework, Bayesian model selection is not presenting a blatant violation of the axioms of probability. However, we may now argue that something else is wrong.

The empirical content of ordinary statistical hypotheses is in their likelihood function. That is, statistical hypotheses can in a sense be told apart by the data, even though they are distinguishable only in the limit. Consider, for example, the hypotheses of Section 4.1, as defined in Equation (6). It is

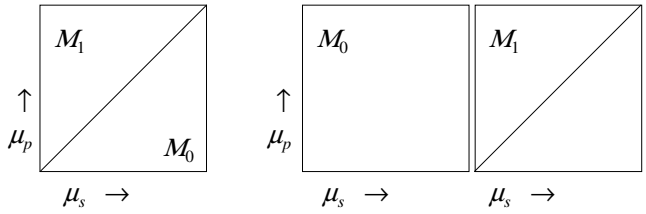

Fig. 5. The leftmost square shows the two models as nested sets of statistical hypotheses. On the right side, the two models are disjunct sets of statistical hypotheses, but these hypotheses have identical likelihood functions.

logically possible that the hypothesis H_θ with $\theta = 1/2$ is true, and that nevertheless the limiting relative frequency of passings in an infinitely long sequence of test results is equal to some other fraction, such as $3/4$; the interested reader may consult Gaifman and Snir [10]. But the probability of this happening is 0. In close connection to this, there are the so-called convergence theorems of Bayesian statistical inference, which show in general that if the hypothesis H_θ is true, the posterior probability $p(H_\theta|E_n)$ will tend to 1 in the limit of larger and larger data sets E_n. In this particular sense we can say that ordinary statistical hypotheses can be told apart by the data.

With this notion of empirical content in place, consider the two statistical models $\mathcal{M}_0$ and $\mathcal{M}_1$ of the DID example, which consist in part of statistical hypotheses that have identical likelihood functions. Can they be told apart by the data in the limit? Of course, if the true hypothesis does not satisfy the restriction imposed by the model $\mathcal{M}_1$, namely that $\mu_{sim} < \mu_{pat}$, then given sufficient data the posterior probability of model $\mathcal{M}_0$ will tend to 1. However, if the true hypothesis does satisfy the restriction imposed by the model $\mathcal{M}_1$, then there is no such limiting behavior. In that case there are two hypotheses with correct values for μ_{pat} and μ_{sim}, namely $H_{0\mu_{pat}\mu_{sim}}$ and $H_{1\mu_{pat}\mu_{sim}}$. And these two hypotheses have exactly the same likelihood function, hence there can never be any piece of data that tells against the one and in favor of the other. Admittedly, within the two models $\mathcal{M}_0$ and $\mathcal{M}_1$ separately, the convergence theorems alluded to in the foregoing take care that the hypotheses $H_{0\mu_{pat}\mu_{sim}}$ and $H_{1\mu_{pat}\mu_{sim}}$ will both attract all the probability. But exactly because $H_{0\mu_{pat}\mu_{sim}}$ and $H_{1\mu_{pat}\mu_{sim}}$ will in the limit attract all probability within their respective models, the initial probability ratio between the two hypotheses $H_{0\mu_{pat}\mu_{sim}}$ and $H_{1\mu_{pat}\mu_{sim}}$ will be retained. To be precise, we have $p(H_{0\mu_{pat}\mu_{sim}})d\mu_{pat}\mu_{sim} = 1/2$ and $p(H_{1\mu_{pat}\mu_{sim}})d\mu_{pat}\mu_{sim} = 1$, because the prior over models is $p(\mathcal{M}_0) = p(\mathcal{M}_1) = 1/2$ while within the two models the prior is uniform, and thus $p(\mathcal{M}_0|E_n) = 1/3$ and $p(\mathcal{M}_1|E_n) = 2/3$ for $n \to \infty$. For a more detailed treatment of this effect in the DID-example, we refer to Chapter 4.

Summing up, it seems that we can avoid a violation of the axioms of probability in Bayesian model selection. We can do so by reconceptualising the models involved in the selection. However, understanding models in this way may leave us with an identifiability problem: if the true parameter values satisfy the restriction at issue, the data do not single out a unique statistical hypothesis, or a single model model for that matter. Instead we retain the difference between the hypotheses and models that we have ourselves imposed at the onset. We may argue that this is not a big deal. After all, once we have gained access to true parameter values, the distinction between $H_{0\mu_{pat}\mu_{sim}}$ and $H_{1\mu_{pat}\mu_{sim}}$, or between $\mathcal{M}_0$ and $\mathcal{M}_1$, may be inessential. This reaction leads us to consider the following question: how can we interpret the intermittent probability assignments over the two models, as long as we do not have the true parameter values? What, if we are eventually interested in the true parameter values, are these probability assignments about?

5.2 The Probability of a Model

Unfortunately these questions are not a cliffhanger, or some other rhetorical device. By way of an answer we only have some suggestions to offer. However, we do feel that these suggestions invite further research, and we are confident that such research will not be in vain.

One rather natural answer to the above questions is that the probability of the model presents us with a specific trade-off between two different aspects of model selection. On the one hand, the probability of the models measures model fit : the better the hypotheses within a model fit the data, the higher the marginal likelihood of the model, and hence the higher the posterior model probability. On the other hand, the probability of the model reflects the simplicity of the model. The number of inequality restrictions in a model is directly related to the value of the probability density function within the model. For example, as indicated in the foregoing, the hypotheses in $\mathcal{M}_1$ have a probability that is twice as large as that of their empirically equivalent counterparts in $\mathcal{M}_0$, because in an intuitive sense the space occupied by $\mathcal{M}_1$ is half of that occupied by $\mathcal{M}_0$. The probability density over the restricted model is therefore twice as large as the probability density over the unrestricted model. Hypotheses in a restricted and hence simpler model are thus given a head start via the prior. This is reminiscent of the standard situation in model selection, in which typically the more complex model has more parameters and hence occupies a larger space as well.

This view on Bayesian model selection invites a host of further questions. One question is whether we have any reason for choosing this specific trade-off between simplicity and model fit. It is as yet unclear whether the bonus for simplicity that is implicit in Bayesian model selection always latches onto our intuitive or independently motivated criteria for the model selection at hand. If this is not the case, we may tweak the priors over the models, as they can be used as an independent component in Bayesian model selection.

Another question is how the trade-off between simplicity and fit fares in cases in which the two models are of different dimensionality, for example if we compare the model $\mathcal{M}_0$ to a third model, $\mathcal{M}_2$, which has the restriction that $\mu_{pat} = \mu_{sim}$. In such cases of differing dimensionality, we may also ask how Bayesian model selection relates to other ways of trading off simplicity and fit, e.g. Aikaike's criterion which concerns differing dimensionality as well. These are all legitimate research questions. We expect that a study into the relation between Bayesian model selection and complexity will therefore be very fruitful.

Apart from weighing simplicity and fit against each other, we can conceive of another function for comparing models in a Bayesian model selection procedure. It may be that eventually the interest of a Bayesian statistical inference lies in determining the values of the parameters in a statistical model. The employment of several models in a Bayesian model selection procedure may be a way of finding the best estimate for some parameter efficiently. This view on the use of several models leads us to consider an interpretation of the posterior model probabilities of an entirely different nature, namely as a clever means to enhance the convergence properties of the Bayesian inference. But before we wholeheartedly adopt this view, it will be wise to investigate the convergence properties of Bayesian statistical inference using multiple models in more detail.

Whatever the exact results of either of the two research lines suggested in the foregoing, we feel that we have already taken one step forward. By describing Bayesian model selection as the continuation of Bayesian statistical inference, and by describing the latter as the continuation of deductive inference, we have provided a context for understanding Bayesian model selection in a philosophical way. We hope that the ground work is laid, and that any further investigations into understanding PMP's posterior model probabilities and Bayes' factors do not have to start at square one.

References

[1] Barnet, V.: Comparative Statistical Inference. Wiley, New York (1999)
[2] Bird, A.: Philosophy of Science. McGill-Queen's University Press, Montreal (1998)
[3] Carnap, R.: Scheinprobleme in der Philosophie. Weltkreis-Verlag, Berlin (1928)
[4] Carnap, R.: The Foundations of Probability. University of Chicago Press, Chicago (1950)
[5] Carnap, R.: The Continuum of Inductive Methods. University of Chicago Press, Chicago (1952)
[6] Cover, J. A., Curd, M.: Philosophy of Science: The Central Issues. Norton and Co., New York (1998)
[7] Dawid, A.P.: The Well–Calibrated Bayesian. Journal of the American Statistical Association, **77**, 605–613 (1982)

[8] De Finetti, B.: Probability, Induction and Statistics. Wiley, New York (1972)
[9] Festa, R.: Optimum Inductive Methods. Kluwer, Dordrecht (1993)
[10] Gaifman, H., Snir, M.: Probabilities over Rich Languages. Journal of Symbolic Logic **47**, 495–548 (1982)
[11] Goodman, N.: Fact, Fiction, and Forecast. Harvard University Press, Cambridge (MA) (1955)
[12] Hintikka, J.: A Two–dimensional Continuum of Inductive Methods. In: Hintikka, J., Suppes, P. (ed.) Aspects of Inductive Logic, North Holland, Amsterdam (1966)
[13] Howson, C.: A Logic of Induction. Philosophy of Science, **64**, 268–90 (1997)
[14] Howson, C.: Hume's Problem. Clarendon Press, Oxford (2000)
[15] Hume, D.: An Enquiry Concerning Human Understanding (1748). Tom Beauchamp (ed.), Oxford University Press, Oxford (1999)
[16] Huntjens, R.J.C., Peters, M.L., Woertman, L., Bovenschen, L.M., Martin, R.C., Postma, A.: Inter-identity amnesia in dissociative identity disorder: A simulated memory impairment? Psychological Medicine, **36**, 857–863 (2006)
[17] Popper, K.: The Logic of Scientific Discovery. Hutchinson, London (1959)
[18] Romeijn, J.W.: Hypotheses and Inductive Predictions. Synthese, **141**, 333–64 (2004)
[19] Romeijn, J.W.: Bayesian Inductive Logic. PhD thesis, University of Groningen, Groningen (2005)
[20] Tversky, A., Kahneman, D.: Extension versus intuitive reasoning: The conjunction fallacy in probability judgment. Psychological Review, **90**, 293-315 (1983)

Index