This is the author's final version of the article published in the International Journal of Approximate Reasoning, 103, 184–194 (2018). The final publication is available at: https://doi.org/10.1016/j.ijar.2018.09.004

An Evaluation of Probabilistic Approaches to Inference to the Best Explanation

David H. Glass
School of Computing, Ulster University, Shore Road,
Newtownabbey, Co. Antrim, BT37 0QB, UK

Abstract

This paper presents results of computer simulations for a number of different probabilistic versions of inference to the best explanation (IBE), which are distinguished by the probabilistic measures used to identify the best explanation. Simulation results are presented which include cases involving ignorance of a catch-all hypothesis, uncertainty regarding the prior probability distribution over the remaining hypotheses, initial elimination of implausible hypotheses, and variations in the number of pieces of evidence taken into consideration. The results show that at least some versions of IBE perform very well in a wide range of cases. In particular, the results for all approaches remain very similar (or improve in some cases) when just the two hypotheses with the highest prior probabilities are retained and the rest are eliminated from consideration.

1 Introduction

Inference to the best explanation (IBE) or abduction is an ampliative mode of reasoning which is often defended as central to scientific reasoning, but also seems to capture aspects of evidential reasoning as it occurs more generally, including in everyday life. IBE proceeds by considering a number of plausible candidate hypotheses in a given evidential context and then comparing these hypotheses in order to make an inference to the one that best explains the relevant evidence. It has been discussed widely in both the philosophy of science and computer science

literature [1–4].

There has been a lot of interest in the relationship between IBE and probability, particularly in debates about the compatibility or otherwise between IBE and Bayesian inference [5–8] and on IBE or abduction in the context of Bayesian networks [9–14]. Here the focus is on using probabilistic approaches to determine which one of a collection of competing hypotheses provides the best explanation of the evidence. These approaches enable two key questions about IBE to be addressed. First, they provide ways of making IBE precise by specifying what is meant by 'best'. Various probabilistic measures have been proposed to this end in the literature [15–19]. Second, they make it possible to investigate whether selecting the best explanation is a good strategy for inferring truth. In this context, some studies have been carried out using computer simulations to evaluate how well various measures perform in hypothesis selection tasks and provide more general defences of versions of IBE based on probabilistic measures [18, 20, 21]. Hence, although the paper does not attempt to address all the philosophical issues surrounding IBE (see for example [2, 3]), the results obtained are very relevant to those debates.

An interesting aspect of previous work is that in cases where there is uncertainty in the prior probability distribution over the hypotheses, inference based on a coherence measure for ranking hypotheses outperformed the approach that simply selects the most probable hypothesis in light of the evidence [18]. The current paper builds on this work by providing a more adequate and realistic account of how probability can be used in IBE and a more systematic evaluation of how IBE so construed performs as a mode of reasoning when ignorance of a catch-all hypothesis, uncertainty regarding the prior distribution, initial elimination of implausible hypotheses, and variations in the number of pieces of evidence available are taken into account.

These are important considerations for IBE in general. For example, whether in the context of scientific reasoning or in everyday, commonsense reasoning, it would be unrealistic to assume complete knowledge of all possible hypotheses. As such, inclusion of a catch-all hypothesis allows for IBE to be modelled and evaluated in a more realistic way. Similarly, even among the known hypotheses, it would be unreasonable to expect that they should all be evaluated in detail since the

number of hypotheses could be large and some hypotheses might be considered very implausible based on background knowledge. Hence, modelling this practical aspect of IBE enables us to see how it affects its performance.

Rather than exploring IBE in the context of particular applications, whether in science or everyday life, the goal of the paper is to evaluate IBE as a general mode of inference. More specifically, the goal is to investigate the performance of various probabilistic models of IBE. Thus, the focus is of a conceptual nature, but by incorporating more realistic aspects of IBE into the simulations the results may also have implications for how IBE could be justified and used in practical applications.

2 Measures for comparing hypotheses

A number of measures have been proposed in the literature to quantify how well a hypothesis h explains evidence e. These include the following measure of explanatory power proposed by Schupbach and Sprenger [16]:

$$\mathcal{E}_{SS}(e,h) = \frac{P(h|e) - P(h|{\sim}e)}{P(h|e) + P(h|{\sim}e)}, \tag{1}$$

an alternative measure of explanatory power proposed by Crupi and Tentori [19]:

$$\mathcal{E}_{CT}(e,h) = \begin{cases} \dfrac{P(e|h) - P(e)}{1 - P(e)} \text{ if } P(e|h) \geq P(e) \\ \dfrac{P(e|h) - P(e)}{P(e)} \text{ if } P(e|h) < P(e), \end{cases} \tag{2}$$

another measure that has been discussed by Good [22] and McGrew [23]:

$$\mathcal{E}_{GM}(e,h) = \ln\left[\frac{P(e|h)}{P(e)}\right]. \tag{3}$$

and the overlap coherence measure (OCM) used to rank explanations by Glass [17, 18]:

$$\mathcal{E}_{OCM}(e,h) = \frac{P(h \wedge e)}{P(h \vee e)}. \tag{4}$$

Criticisms of some of these measures were presented by Glymour [24], while a response has been given by Glass [21]. In order to respond to a criticism based on the fact that advantages of the $\mathcal{E}_{OCM}$ diminished as the sample size (i.e. the number of

samples of evidence) increased, the following alternative product coherence measure (PCM) was proposed and shown to retain the advantages with increasing sample size:

$$\mathcal{E}_{PCM}(e,h) = P(e|h) \times P(h|e). \tag{5}$$

The strategy used in this paper and discussed in detail in section 3 is to consider a set of hypotheses $\{h_i\}$ for evidence e and select the hypothesis which gives the maximum value of a particular measure, $\mathcal{E}$. It would be possible to use all the measures defined above, but it turns out that $\mathcal{E}_{SS}$, $\mathcal{E}_{CT}$ and $\mathcal{E}_{GM}$ all give the same ranking of hypotheses, giving the same result as selecting the hypothesis with the maximum likelihood, $P(e|h_i)$.[1]

Another measure that will be considered is the posterior probability of the hypotheses in light of the evidence, $P(h_i|e)$. The hypothesis that maximizes posterior probability is often referred to in the artificial intelligence literature as the most probable explanation (MPE). Arguably, this is a poor definition of 'best explanation' [15, 17], but it nevertheless provides a standard against which to compare the various explanatory measures.

A final measure which will be considered is the likelihood ratio measure:

$$\mathcal{E}_{LR}(e,h) = \frac{P(e|h)}{P(e|{\sim}h)}. \tag{6}$$

As a popular Bayesian confirmation measure it provides a further alternative to which the other explanatory measures can be compared.

In summary, the following hypothesis selection strategies will be used:

MPE: most probable explanation; selects the hypothesis with the maximum posterior probability, $P(h_i|e)$,

ML: selects the hypothesis with the maximum likelihood, $P(e|h_i)$,

[1] This is straightforward to show for $\mathcal{E}_{CT}$ and $\mathcal{E}_{GM}$. From the definition of $\mathcal{E}_{SS}$ it is easy to show that $\mathcal{E}_{SS}(e,h_1) > \mathcal{E}_{SS}(e,h_2)$ if and only if $P(h_1|e)P(h_2|{\sim}e) > P(h_1|{\sim}e)P(h_2|e)$. Using Bayes' theorem to replace each term and then rearranging, we can see that this expression is true if and only if $P(e|h_1) > P(e|h_2)$, provided that P is a regular probability function. See also theorem 1 in [19].

PCM: selects the hypothesis with the maximum value of $\mathcal{E}_{PCM}$,

OCM: selects the hypothesis with the maximum value of $\mathcal{E}_{OCM}$,

LR: selects the hypothesis with the maximum value of $\mathcal{E}_{LR}$.

Essentially, this means that five different versions of IBE will be implemented and evaluated against each other, one using each of these strategies.[2] However, since PCM and OCM turn out to give almost identical results when only one piece of evidence is considered, results for OCM will only be presented in cases where the sample size is greater than one.

3 Computational implementation of IBE

3.1 Defining IBE

As noted earlier, the general idea is to make an inference to the hypothesis which best explains the evidence. Assuming that an agent has a probability model, then one of five strategies identified in section 2 can be used to make this selection. However, much more detail is needed to make IBE more realistic and to make it feasible to implement and evaluate it. At a high level, it is divided into a two stage process as illustrated in figure 1. In practice there could be a very large set of possible hypotheses and so, following Lipton [2, p.59], a subset of plausible 'live options' will be considered for inference. The second step is then to discriminate between these hypotheses using one of the five strategies in order to make an inference. IBE assumes that the hypotheses are competing and for the purposes of this paper, this will be taken to mean that they are mutually exclusive as is typical for most work on this topic.[3]

Let us now consider the first step in more detail. One aspect of this relates to *ignorance* concerning the set of possible hypotheses; the agent simply may not be

[2]The MPE strategy should really be considered as a standard to compare the others against rather than part of a legitimate account of IBE.

[3]An alternative account of competition is explored in [25] which also explores how inference is affected when the hypotheses are not mutually exclusive, see also [26].

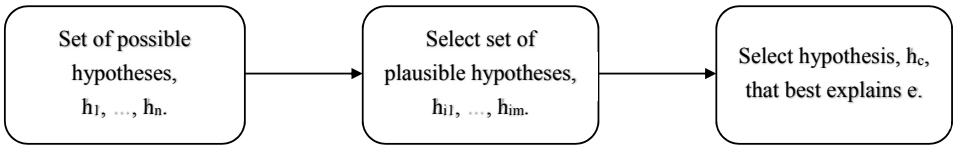


Figure 1: Basic schema for implementing IBE.

aware of some of these hypotheses. This will be modelled via a *catch-all* hypothesis, which could represent the disjunction of multiple mutually exclusive hypotheses. This catch-all hypothesis could be the actual or true hypothesis, in a sense to be discussed shortly, and so could give rise to the evidence, but it cannot be inferred by the agent, who is simply ignorant of this possibility.

Even taking into account the ignorance of the agent, a further aspect to the first step is the *selection* of a subset of *plausible* hypotheses out of all the possible hypotheses of which the agent is aware. This will be modelled by ruling out a number of hypotheses which have low prior probability according to the agent.

But where do these prior probabilities come from? Here we need to distinguish between what we can call the actual or objective probability distribution over the set of all possible hypotheses and the subjective probability of the agent. What is the relationship between these two probabilities? One option is that the subjective probability of the agent is simply a modified version of the objective probability distribution which takes into account ignorance of the catch-all hypothesis. This can be achieved by assigning the values of the objective prior probabilities to the subjective prior probabilities over all the hypotheses excluding the catch-all, and then normalizing to ensure they sum to one. However, we can also incorporate *uncertainty* in the subjective probabilities by assigning the objective prior probabilities with a random error and then normalizing as before. The inclusion of uncertainty in the priors helps to make the model more realistic and hence more relevant to applications since in practice priors can be based on limited knowledge.

So far the focus has been on incorporating ignorance, initial selection and uncertainty in the first step. Let us now consider the second step, i.e. selecting the hypothesis from the set of plausible hypotheses that best explains the evidence. To achieve this the agent's subjective likelihoods will be assigned the values of the ob-

jective likelihoods and then using the five different strategies defined in the previous section the agent will select the plausible hypothesis which maximizes the respective measure.

One final component of the second step is that instead of selecting the best hypothesis based on a single piece of evidence, a sample size of one, we can also look at the selection being made on *multiple samples of evidence*. For example, suppose the agent were trying to select one out of a number of hypotheses concerning the bias of a coin. Instead of just tossing the coin once and making an inference, the agent could toss it multiple times. Again, this helps to make the model more realistic and enables us to see how the performance of the different approaches depends on sample size.

Having identified the relevant factors, let us now consider exactly how IBE will be implemented and evaluated.

3.2 Implementing IBE

Following the approach in [18], but extending it in a number of directions, the first task is to identify an objective probability model, P_O, which is used to identify one of the hypotheses as the actual hypothesis h_A and then simulate whether the evidence e or its negation $\sim e$ occurs. It is important to emphasize that the goal is not to model a particular application of IBE, where prior probabilities and likelihoods might be based on available data or expert opinion. Rather, the goal is to evaluate the various probabilistic approaches for IBE in a much more general way. Hence, the strategy adopted is to sample the entire probability space for the prior probabilities and likelihoods. Each particular simulation can then be viewed as representing a specific instance of IBE with a particular distribution of prior probabilities and likelihoods.

For a task with n possible hypotheses, including a catch-all, the objective prior probabilities, $P_O(h_i)$, where $i \in \{1, ..., n\}$, are obtained by sampling a Dirichlet distribution. A uniform Dirichlet distribution is obtained by setting all the α parameters (corresponding to each of the hypotheses) in the Dirichlet distribution to 1 and this is the distribution that was sampled for all the results presented in the

main paper, although values for all the α parameters of 0.5 and 2 respectively were also used for comparative purposes and are presented in Supplementary Material. Setting $\alpha = 2$ results in the hypotheses being assigned priors that are more similar to each other, while a value of 0.5 results in priors which are more varied so that most tend to be small while a few have higher values.[4] Based on these prior probabilities, one of the hypotheses is randomly selected and designated the actual hypothesis, h_A.

Likelihoods for each hypothesis, $P_O(e|h_i)$, are randomly selected from the uniform distribution over the interval $[0, 1]$. Based on the likelihood of the actual hypothesis, $P_O(e|h_A)$, a random selection is made as to whether the evidence e or its negation $\sim e$ occurs. Or if multiple samples of evidence are required, multiple random selections of e or $\sim e$ are made.

Having identified the actual hypothesis and whether e or $\sim e$ occurs, the next step is to construct the agent's subjective probability model, P_S, which will then be used to make an inference. First of all, one of the possible hypotheses, $h_1, ..., h_n$ is randomly selected and designated the catch-all hypothesis, h_{CA}. The agent's prior probability for this hypothesis is set to zero, $P_S(h_{CA}) = 0$, i.e. it is excluded from the agent's set of possible hypotheses. Clearly, if the catch-all is also the actual hypothesis, $h_{CA} = h_A$, there is no possibility of the agent making a correct inference.

As noted earlier, values for the agent's prior probabilities for the remaining hypotheses are obtained by assigning the respective objective probabilities with a random error obtained from the normal distribution with a specified variance. These values are then normalized to give $P_S(h_i)$. In order to obtain a desired number, ns, of plausible hypotheses, it is simply a matter of selecting the ns hypotheses with the highest values of $P_S(h_i)$.

[4]The author would like to thank an anonymous reviewer for suggesting this approach. In [18, 21], n priors were obtained by selecting $n - 1$ values randomly selected from the uniform distribution over the interval $[0, 1]$. These values together with values 0 and 1 were then put in ascending order and the differences between consecutive numbers were then assigned as the objective prior probabilities, $P_O(h_i)$, where $i \in \{1, ..., n\}$. The results obtained using that approach are identical to those used here, i.e. a Dirichlet distribution with all the α parameters set to 1. The equivalence of the two approaches is demonstrated in Appendix A.

In order to make an inference, the agent's likelihoods are assigned the same values as the objective likelihoods, $P_S(e|h_i) = P_O(e|h_i)$. For each strategy (MPE, ML, PCM, OCM, LR), the agent then selects the plausible hypothesis that maximizes the corresponding measure for the evidence (e or $\sim e$ or a sequence of such values if multiple samples of evidence are required) and if it matches the actual hypothesis it is counted as a success, otherwise it is a failure. This process is then repeated multiple times to get the accuracy for each strategy, which is simply defined as the number of successes divided by the number of trials.[5]

Instead of using simulations, it is possible to obtain analytical results in at least some cases. For example, when there are two hypotheses, a triple integral over the priors and the likelihoods for each hypothesis can be used to determine the expected accuracy for each of the approaches. Letting $x = P_O(h_1) = P_S(h_1)$, $y = P_O(e|h_1) = P_S(e|h_1)$ and $z = P_O(e|h_2) = P_S(e|h_2)$, the result for the MPE approach can be obtained by integrating as follows:

$$2\int_0^1 \int_0^1 \left(\int_0^{\frac{z}{y+z}} (1-x)z \, dx + \int_{\frac{z}{y+z}}^1 xy \, dx \right) dydz \tag{7}$$

which evaluates to $2log(2)/3 + 1/3 \approx 0.7954$. A similar approach yields $pi/4 \approx 0.7854$ for the PCM approach and $2/3$ for the LR and ML approaches. (These results provide a check of corresponding results obtained by simulations that are presented in figure 3a.) However, for larger numbers of hypotheses the multiple integrals become much more complex ($2n-1$ integrals for n hypotheses) and further complexity arises from the selection of a catch-all hypothesis, the agent's restriction of the set of possible hypotheses and the introduction of uncertainty in the priors.

For these reasons, all the results presented in subsequent sections were obtained

[5] Accuracy as defined here makes sense in the current context since it captures how well on average an approach identifies the actual or true hypothesis, which is relevant for IBE. Nevertheless, various other performance metrics could also be explored. For example, a metric could be used that takes into account the agent's overall ranking of hypotheses so that an actual hypothesis that is ranked highly by the agent (though not ranked highest) would contribute to the score. Although IBE is not used to update probabilities in the current work, it is worth noting that in the context of debates about IBE and Bayesianism, there has been discussion about the relevance of epistemic and decision-theoretic principles in determining whether update rules are coherent [7, 8].

from simulations implemented in C++, with the process described earlier being repeated 10^7 times to get average accuracies. Algorithm 1 summarizes the process. The variables $Ph_o(i)$, $Peh_o(i)$, $Ph_s(i)$, and $Peh_s(i)$ represent the probabilities $P_O(h_i)$, $P_O(e|h_i)$, $P_S(h_i)$, and $P_S(e|h_i)$ respectively. Results obtained using this algorithm and slight modifications of it, which will be noted in due course, will now be presented.

4 Experimental results

As noted earlier, prior probabilities for the hypotheses were sampled from a Dirichlet distribution. All the results presented in this section were obtained by using a uniform Dirichlet distribution where all the α parameters were set to 1, but results obtained with $\alpha = 0.5$ and $\alpha = 2$ are presented in Supplementary Material. Qualitatively, the results are very similar. The main difference is that the MPE and PCM approaches in particular perform better for small α and hence their advantage over the ML and LR approaches becomes more pronounced at $\alpha = 0.5$ and less pronounced at $\alpha = 2$. This is due to the fact that there are greater differences between the priors for small α and MPE and PCM are able to take advantage of this.

4.1 Catch-all hypothesis

Results presented in figure 2 are for cases where there is a catch-all hypothesis, but there is no uncertainty in the agent's prior probability, the plausible hypotheses consist of all the possible hypotheses (i.e. there is no initial restriction of the hypotheses apart from excluding the catch-all) and only a single piece of evidence is taken into account. Figure 2a presents results obtained when one hypothesis is randomly selected as the catch-all hypothesis, while figure 2b presents results when the hypothesis with the lowest likelihood is selected as the catch-all hypothesis. This latter option would represent a scenario where the agent is ignorant of a hypothesis with low likelihood, but is aware of all the hypotheses with higher likelihoods. In both cases the results for the MPE approach when there is no catch-all, and so the agent is aware of all the hypotheses, are presented for comparative purposes.

Algorithm 1 Calculate accuracy for explanatory measure $\mathcal{E}$

input: $nrepeat$: no. of repetitions; n : number of hypotheses;
input: ns : no. of plausible hypotheses; nev : no. of evidence samples
ouput: $accuracy$
$success \leftarrow 0$
for j $= 1$ to $nrepeat$ **do**
    for $i \in \{1, ..., n\}$ **do**
        $Ph_o(i) \leftarrow$ randomly select objective prior probabilities
        $Peh_o(i) \leftarrow$ randomly select objective likelihoods
    $h_A \leftarrow$ Randomly select actual hypothesis based on priors
    $h_{CA} \leftarrow$ Randomly select catch-all based on uniform distribution
    $Ph_s(CA) \leftarrow 0$
    for $i \in \{1, ..., CA-1, CA+1, ..., n\}$ **do**
        $Ph_s(i) \leftarrow Ph_o(i)+$ random error from normal distribution
        $Peh_s(i) \leftarrow Peh_o(i)$
    Normalize Ph_s
    Select ns plausible hypotheses with highest values of Ph_s
    $e_{total} \leftarrow \emptyset$
    for $k \in \{1, ..., nev\}$ **do**
        $e_k \leftarrow e$ or $\sim e$ based on likelihood of h_A
        $e_{total} \leftarrow e_{total} \cup e_k$
    Select h_C from plausible hypotheses which maximizes $\mathcal{E}(e_{total}, h_i)$
    if $h_C = h_A$ **then**
        $success \leftarrow success + 1$
$accuracy \leftarrow success/nrepeat$
return accuracy

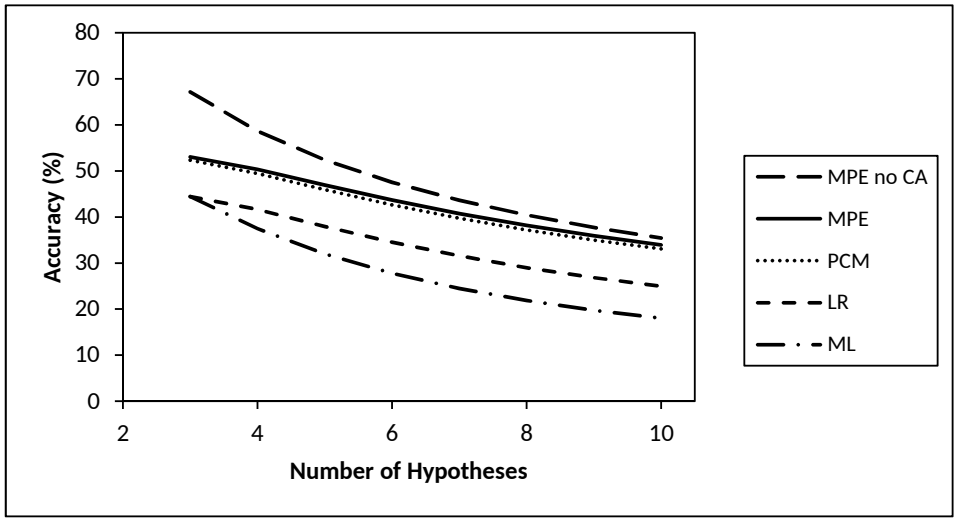
It is clear from both sets of results that there is almost no difference between the results for MPE and PCM when a catch-all hypothesis is included for both approaches. This is consistent with the results for the OCM approach in [18]. Also, both MPE and PCM outperform LR which in turn outperforms ML. When the catch-all hypothesis is selected at random the results in figure 2a show that MPE and PCM perform much worse than MPE without a catch-all, but as the number of hypotheses under consideration increases this difference becomes much less significant; being ignorant of one of ten hypotheses has very little effect on accuracy.

Figure 2b displays results for MPE and PCM which are almost identical to those of MPE without a catch-all. Hence being ignorant of one low likelihood hypothesis has very little impact on these approaches to inference even for small numbers of hypotheses. By contrast the results for LR and ML change very little between the two scenarios.

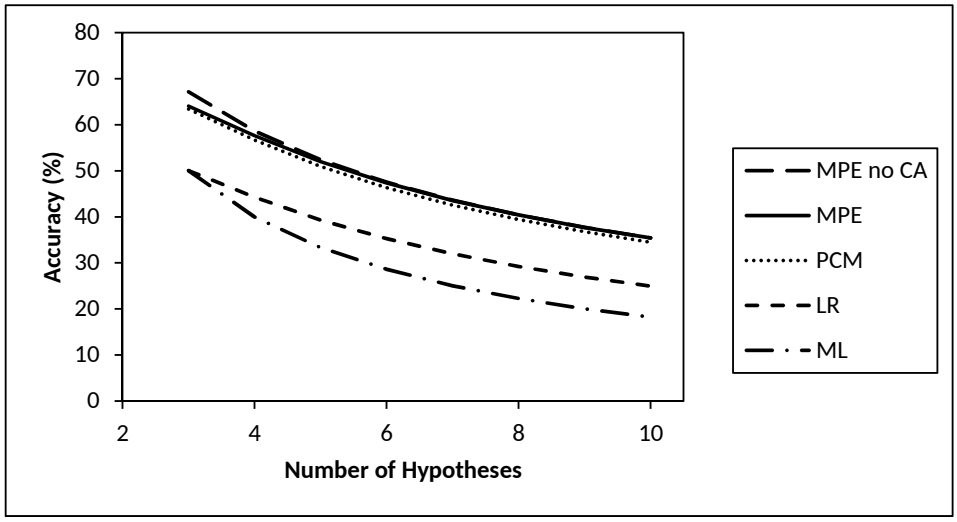
4.2 Selection of plausible hypotheses

Figure 3 shows results for the scenario where the agent restricts the pool of plausible hypotheses to just the two which have the highest prior probability according to the agent's subjective probability distribution. As before, there is no uncertainty in the agent's distribution and only a single piece of evidence is considered. Figure 3a presents results for the case where there is no catch-all, while 3b includes a catch-all selected at random. It should be noted that the restriction to two hypotheses has no effect if there are just two hypotheses in figure 3a or just three hypotheses in 3b (since in this case the agent is only considering two anyway, having excluded the catch-all).

Despite the fact that excluding all but two hypotheses seems very drastic, at least for higher numbers of hypotheses, it has almost no effect on the results for MPE or PCM. This can be seen by comparing the results for these approaches in figure 3a for MPE and PCM with those in figure 2a for MPE with no catch-all and for the results in figure 3b for these approaches with those for the same approaches in 2a. This seems like a very surprising result. Further investigation shows that



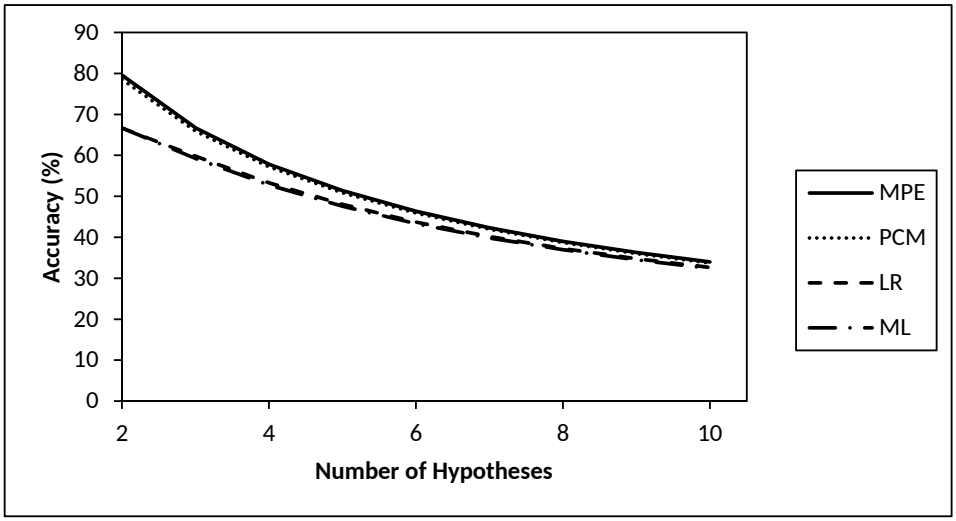
(a)



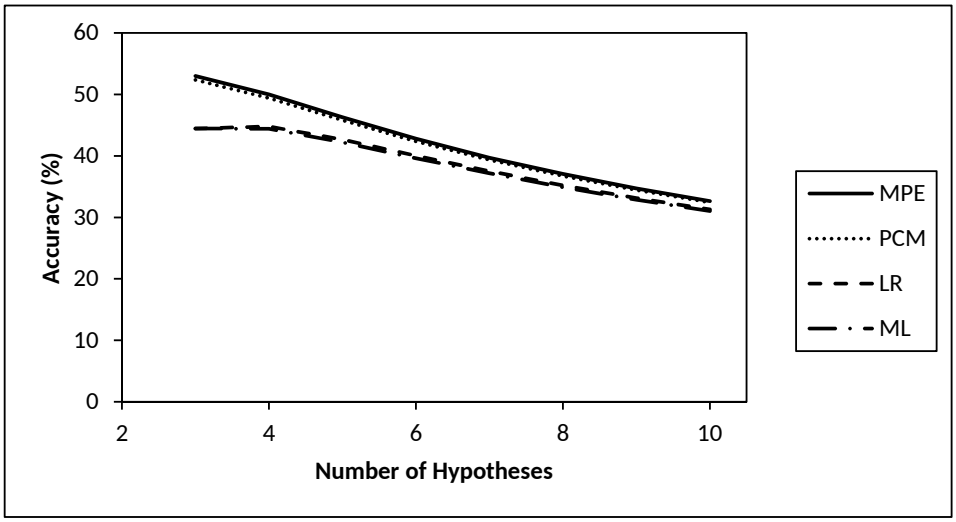
(b)

Figure 2: Results when a catch-all hypothesis is included so that the agent is making an inference based on $n-1$ hypotheses instead of all n. a) The catch-all hypothesis is selected randomly from the set of all hypotheses. b) The hypothesis with the lowest likelihood is selected as the catch-all. In both cases, results for MPE with no catch-all are also presented for comparison.

MPE and PCM only succeed in identifying the actual hypothesis in very few cases if it is not one of the top two in terms of prior probability.



(a)



(b)

Figure 3: Results for the case where the agent excludes all hypotheses except the two that have the highest prior probability according to the agent's distribution. a) No catch-all is included. b) A catch-all hypothesis is selected at random.

Both sets of results in figure 3 also show that LR and ML perform much better when only the top two hypotheses are considered. It can be seen that the gap between these two approaches and MPE and PCM has narrowed considerably compared to figure 2 and in fact all the approaches appear to converge for higher

numbers of hypotheses. Again, this initially seems surprising since most of the hypotheses are excluded yet the results improve. Further investigation has shown that although these approaches do better than MPE and PCM at identifying the actual hypothesis when it has a low prior probability, this is more than outweighed by the fact that they do much worse than MPE and PCM when the actual hypothesis has a high prior probability. Hence, by only considering hypotheses with high prior probabilities this reduces the number of failures and so enhances their accuracy.

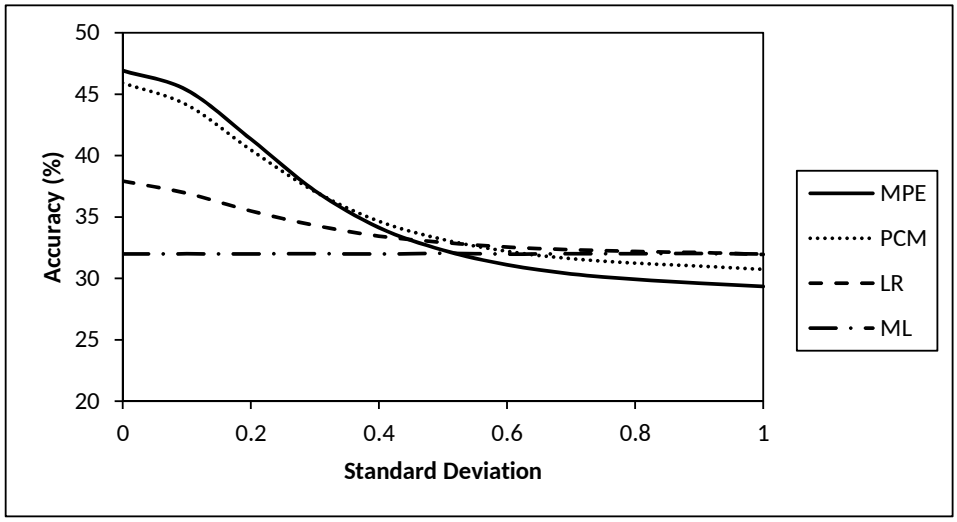
4.3 Uncertainty in the priors

Figure 4 shows results for scenarios where there is uncertainty in the agent's prior probability distribution with increasing standard deviation representing a larger random error and hence greater uncertainty. Figure 4a presents results for the case where there is a catch-all, but otherwise the agent does not restrict the number of hypotheses, while 4b does not include a catch-all, but now the agent does restrict the pool of hypotheses to just two. As was shown for the OCM approach in [18], the PCM approach tracks the MPE approach very closely and actually achieves a higher accuracy in figure 4a when the standard deviation is above about 0.4. While LR and ML also do better than MPE for greater uncertainty, they do much worse for lower values of uncertainty.

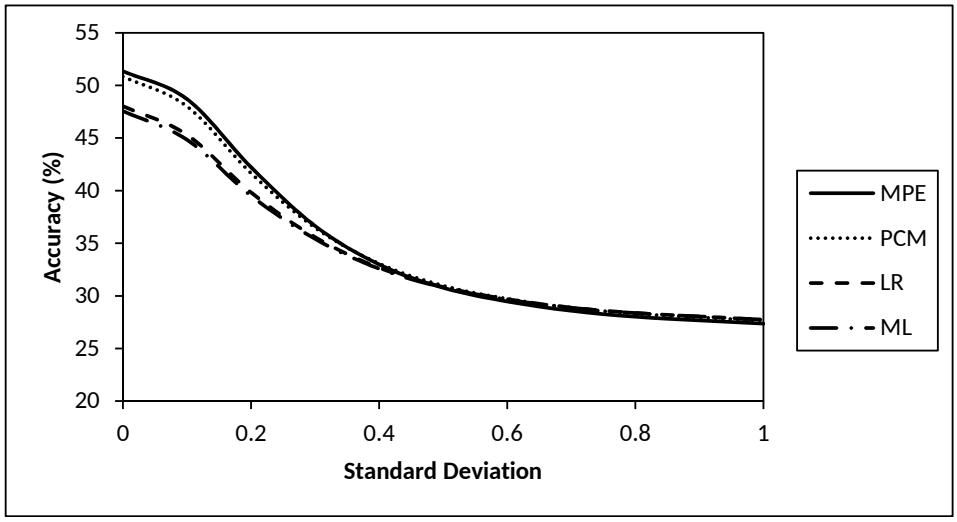
Figure 4b shows, as before, that LR and ML do much better when the pool of plausible hypotheses is restricted to just two, but this is only true for low values of the standard deviation and hence uncertainty. For higher values, the advantage over MPE is lost and indeed there is very little difference between any of the approaches.

4.4 Increasing the sample size

Figure 5 shows results for scenarios where the number of evidence samples is varied. Both sets of results include a catch-all, but not uncertainty. Figure 5a confirms a finding presented in [21], but now in the case where there is ignorance of the catch-all hypothesis. It is that PCM continues to track the MPE result as the sample size increases whereas OCM does not and instead converges to the ML result. LR is shown to lie midway between ML and MPE/PCM and continues to do so as the



(a)

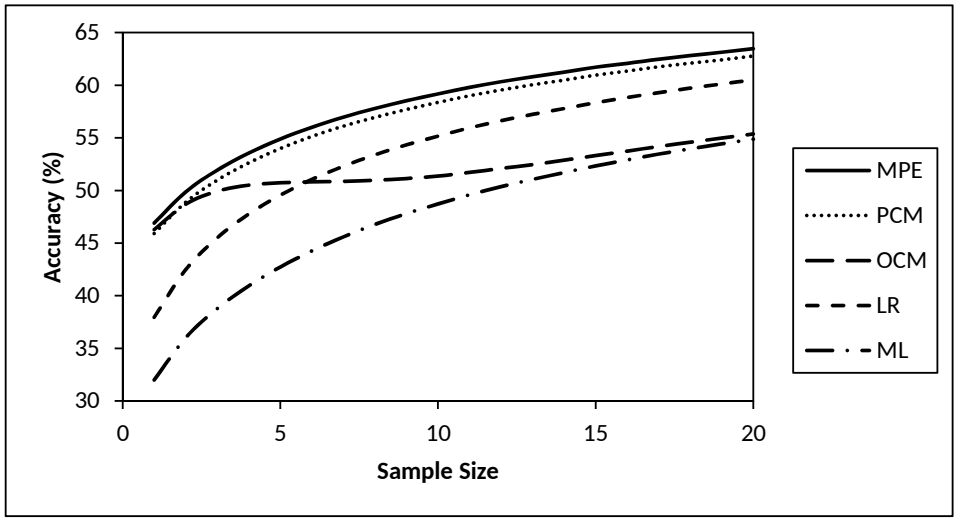


(b)

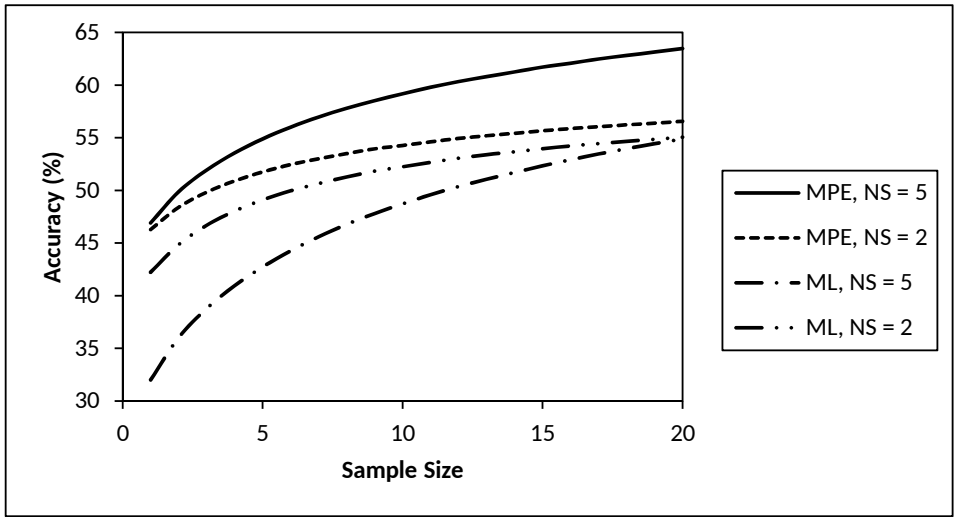
Figure 4: Results for the case where the number of hypotheses is five and there is uncertainty in the agent's prior probability distribution, with the standard deviation being that of the random error. a) A catch-all hypothesis is selected at random, but otherwise the agent does not restrict the number of hypotheses. b) No catch-all is included, but now the agent does restrict the pool of hypotheses to just two.

sample size increases.

Figure 5b compares results where there is no restriction in the pool of plausible



(a)



(b)

Figure 5: Results for the case where the number of hypotheses is five and where the number of pieces evidence sampled is varied. Both figures include a catch-all hypothesis, but no uncertainty. a) The agent does not restrict the pool of plausible hypotheses. b) Results for MPE and ML to compare the case where there is no restriction of plausible hypotheses, $NS = 5$, to that where it is restricted to just two, $NS = 2$.

hypotheses to those where it is restricted to just two hypotheses. Results are only shown for MPE and ML, but similar results could be presented for PCM and LR. As noted earlier, restricting the plausible hypotheses to just two has little effect on the accuracy of MPE when there is just one sample, but improves the accuracy of ML substantially. This can be seen in this figure also for low sample sizes. However, as the sample size increases the accuracy of MPE does indeed drop off compared to the no-restriction case, while the advantage of ML over the no-restriction case diminishes.

These results help to provide an explanation of the surprising results in figure 3. With only one piece of evidence MPE is typically unable to correctly identify the actual hypothesis when it has a low prior probability and so the restricted case works well. However, when much more evidence is available MPE is able to discriminate between the hypotheses more accurately and so the advantages of the no-restriction case can now be seen as might have been expected. Similarly, since the significance of the priors is reduced when there is more evidence available, there is then less difference between the restricted and non-restricted versions of ML.

5 Conclusion

Several probabilistic versions of IBE have been compared through the use of computer simulations. In particular, explanatory approaches based on the product coherence measure (PCM), the overlap coherence measure (OCM), maximum likelihood (ML) and the likelihood ratio (LR) have been evaluated to see how well they compare with the most probable explanation (MPE) approach that simply selects the hypothesis that has the greatest posterior probability. The results show that PCM performs much better than ML and LR in almost all cases and better than OCM for larger sample sizes. PCM also tracks the accuracy of the MPE very well in all cases that have been considered and actually performs better for high levels of uncertainty in the priors. Hence, if PCM is considered to be a viable approach to IBE, then IBE is a successful mode of inference.

The results also show that IBE, at least when implemented using PCM, still

performs well when there is ignorance about one of the hypotheses as represented by the catch-all hypothesis. It might have been expected that the performance of IBE would diminish considerably when there is a preliminary selection of plausible hypotheses. When the agent's prior probabilities are used to achieve this, the performance is remarkably robust (and improves for ML and LR) even when just two hypotheses are selected and the others discarded. The results do tail off somewhat for larger sample sizes, but the surprising nature of this result for low sample sizes could have implications for abductive inference and inference more generally since it suggests preliminary selection of hypotheses, which could be significant in computational terms, could be carried out with little effect on accuracy when the sample size is small.

Future directions for research include investigating IBE in contexts where there is uncertainty in the likelihoods as well as the priors and where the hypotheses are not assumed to be mutually exclusive. It would also be interesting to explore how these approaches compare to other probabilistic approaches to IBE that involve alternatives to Bayesian updating by giving a boost to hypotheses that provide better explanations [8, 27], including how such approaches can be applied in a social network of interacting agents [28] and extended to handle uncertain evidence [29]. As in the current work, these alternative approaches have presented different computational results concerning IBE and have highlighted various complementary merits to IBE as a mode of inference.

Acknowledgements

The author would like to thank Mark McCartney for helpful discussions.

Appendix A. Equivalence of approaches for selecting priors

As discussed in section 3.2, a Dirichlet distribution was used to obtain prior probabilities. It was pointed out that this approach gave the same results as an alternative

approach to obtaining priors used in earlier work [18,21]. Here it is shown that these approaches are in fact equivalent to each other.

In the current work, n priors were obtained by sampling a uniform Dirichlet distribution. The Dirichlet distribution of order $n \geq 2$ with parameters $\alpha_1, \ldots, \alpha_n > 0$ has a probability density function given by:

$$f(x_1, \ldots, x_n; \alpha_i, \ldots, \alpha_n) = \frac{\Gamma(\sum_{i=1}^{n} \alpha_i)}{\prod_{i=1}^{n} \Gamma(\alpha_i)} \prod_{i=1}^{n} x_i^{\alpha_i - 1}. \tag{A.1}$$

For a uniform distribution, f_1, $\alpha_i = 1$ for all i, which gives:

$$f_1(x_1, \ldots, x_n) = \frac{\Gamma(n)}{\Gamma(1)^n} = (n-1)! \tag{A.2}$$

In the earlier work, n priors were obtained by first of all selecting $n-1$ values randomly from the uniform distribution over the interval $[0, 1]$, which we will denote $\mathbf{z} = (z_1, \ldots, z_{n-1})$. Since the values are selected independently, the probability density, g, is just a product of $n-1$ uniform distributions:

$$g(z_1, \ldots, z_{n-1}) = 1. \tag{A.3}$$

These values together with the value one were then put in ascending order to give a sequence which we shall call $\mathbf{y} = (y_1, \ldots, y_{n-1}, y_n)$, where $y_n = 1$, for each $i \in \{1, \ldots, n-1\}$, $\exists j \in \{1, \ldots, n-1\}$ such that $y_i = z_j$, and $y_{i+1} > y_i$. In going from $\mathbf{z}$ to $\mathbf{y}$, note that there are $(n-1)!$ unordered sequences of random numbers that result in the same ordered sequence $\mathbf{y}$. These unordered sequences are just the $(n-1)!$ permutations of the values $y_1, \ldots, y_{n-1}$. For example, when $n = 3$, the random selected sequences $\mathbf{z_1} = (0.43, 0.29)$ and $\mathbf{z_2} = (0.29, 0.43)$, where there is no restriction on the order of the numbers, both result in the same ordered sequence $\mathbf{y} = (0.29, 0.43, 1)$. According to equation (A.3), the probability density for each of the $(n-1)!$ unordered sequences is the same and so the probability density, h, for the ordered sequences is given by:

$$h(y_1, \ldots, y_{n-1}, 1) = (n-1)! \times g(z_1, \ldots, z_{n-1}) = (n-1)! \tag{A.4}$$

where $(z_1, \ldots, z_{n-1})$ is one of the sequences that results in $(y_1, \ldots, y_{n-1})$ when the values are put in ascending order.

The final step in the second approach is to convert the sequence y into a sequence that sums to one and so can be assigned as probabilities. This is achieved by taking the difference between subsequent values in y to obtain $\mathbf{x} = (x_1, \ldots, x_n)$, where $x_1 = y_1$ and $x_i = y_i - y_{i-1}$ for all i from 2 to $n-1$ and $x_n = 1 - y_{n-1}$. With this transformation of variables, the probability density for $\mathbf{x}$, denoted f_2 is given by:

$$f_2(x_1, \ldots, x_n) = h(y_1, \ldots, 1) \times |J| \tag{A.5}$$

where J is the $(n-1)$-dimensional Jacobian matrix (since the probability densities are actually defined on a $(n-1)$-dimensional space) with elements $J_{ij} = \partial y_i / \partial x_j$. Since $y_i = \sum_{j=1}^{i} x_j$, it follows that $\partial y_i / \partial x_j = 1$ if $i \geq j$ and 0 otherwise, from which it further follows that $|J| = 1$. Hence,

$$f_2(x_1, \ldots, x_n) = h(y_1, \ldots, 1) = (n-1)! \tag{A.6}$$

and so by comparison with equation (A.2) we see that the probability densities from the two approaches are the same.

Appendix B. Supplementary material

Supplementary material related to this article can be found online at `https://doi.org/10.1016/j.ijar.2018.09.004`.

References

[1] J. R. Josephson and S. G. Josephson. *Abductive Inference: Computation, Philosophy and Technology*. Cambridge University Press, Cambridge, 1994.

[2] P. Lipton. *Inference to the Best Explanation*. Routledge, London, 2nd edition, 2004.

[3] I. Douven. Abduction. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2016 edition, 2016.

[4] L. Magnani and T. Bertolotti, editors. *Springer Handbook of Model-Based Science*. Springer, 2017.

[5] B. C. van Fraassen. *Laws and Symmetry*. Clarendon Press, Oxford, 1989.

[6] H. Leitgeb and R. Pettigrew. An objective justification of Bayesianism II: The consequences of minimizing inaccuracy. *Philosophy of Science*, 77(2):236–272, 2010.

[7] I. Douven. Inference to the best explanation made coherent. *Philosophy of Science*, 66:S424–S435, 1999.

[8] I. Douven. Inference to the best explanation, dutch books, and inaccuracy minimisation. *The Philosophical Quarterly*, 63(252):428–444, 2013.

[9] D. Poole. Probabilistic Horn abduction and Bayesian networks. *Artificial Intelligence*, 64(1):81 – 129, 1993.

[10] C. Lacave and F. J. Diez. A review of explanation methods for Bayesian networks. *The Knowledge Engineering Review*, 17(2):107–127, 2002.

[11] J. Park and A. Darwiche. Complexity results and approximation strategies for map explanations. *Journal of Artificial Intelligence Research*, 21:101–133, 2004.

[12] C. Yuan, H. Lim, and T. Lu. Most relevant explanation in Bayesian networks. *Journal of Artificial Intelligence Research*, 42:309–352, 2011.

[13] J. Kwisthout. Most probable explanations in bayesian networks: Complexity and tractability. *International Journal of Approximate Reasoning*, 52(9):1452–1469, 2011.

[14] J. Kwisthout. Most frugal explanations in Bayesian networks. *Artificial Intelligence*, 218:56 – 73, 2015.

[15] U. Chajewska and J. Y. Halpern. Defining explanation in probabilistic systems. In *Proceedings of the 13th Conference on Uncertainty in AI*, pages 62–71, 1997.

[16] J. N. Schupbach and J. Sprenger. The logic of explanatory power. *Philosophy of Science*, 78(1):105–127, 2011.

[17] D. H. Glass. Coherence measures and inference to the best explanation. *Synthese*, 157:275–296, 2007.

[18] D. H. Glass. Inference to the best explanation: does it track truth? *Synthese*, 185:411–427, 2012.

[19] V. Crupi and K. Tentori. A second look at the logic of explanatory power (with two novel representation theorems). *Philosophy of Science*, 79(3):365–385, 2012.

[20] J. N. Schupbach. Comparing probabilistic measures of explanatory power. *Philosophy of Science*, 78(5):813–829, 2011.

[21] D. H. Glass. Coherence, explanation, and hypothesis selection. *The British Journal for the Philosophy of Science*, page axy063, 2018.

[22] I. J. Good. Weight of evidence, corroboration, explanatory power, information, and the utility of experiments. *Journal of the Royal Statistical Society: Series B*, 22:319–331, 1960.

[23] T. McGrew. Confirmation, heuristics and explanatory reasoning. *The British Journal for the Philosophy of Science*, 54:553–567, 2003.

[24] C. Glymour. Probability and the explanatory virtues. *The British Journal for the Philosophy of Science*, 66(3):591–604, 2015.

[25] J. N. Schupbach and D. H. Glass. Hypothesis competition beyond mutual exclusivity. *Philosophy of Science*, 84(5):810–824, 2017.

[26] D. H. Glass and M. McCartney. Explanatory inference under uncertainty. In Emilio Corchado, José A. Lozano, Héctor Quintián, and Hujun Yin, editors, *Intelligent Data Engineering and Automated Learning – IDEAL 2014. Lecture Notes in Computer Science*, volume 8669, pages 215–222. 2014.

[27] I. Douven. Inference to the best explanation: What is it? and why should we care? In K. McCain and T. Poston, editors, *Best Explanations: New Essays on*

Inference to the Best Explanation, pages 7–24. Oxford University Press, Oxford, 01 2017.

[28] I. Douven and S. Wenmackers. Inference to the best explanation versus Bayes's rule in a social setting. *The British Journal for the Philosophy of Science*, 68(2):535–570, 2017.

[29] B. Trpin and M. Pellert. Inference to the best explanation in uncertain evidential situations. *The British Journal for the Philosophy of Science*, page axy027, 2018.