Coherence, Explanation, and Hypothesis Selection

David H. Glass

Abstract

This paper provides a new approach to inference to the best explanation (IBE) based on a new coherence measure for comparing how well hypotheses explain the evidence. It addresses a number of criticisms of the use of probabilistic measures in this context by Clark Glymour ([2015]), including limitations of earlier work on IBE (Glass [2012]). Computer experiments are used to show that the new approach finds the truth with a high degree of accuracy in hypothesis selection tasks and that in some cases its accuracy is greater than hypothesis selection based on maximizing posterior probability. Hence, by overcoming some of the problems with the previous approach, this work provides a more adequate defence of IBE and suggests that IBE not only tracks truth but also has practical advantages over the previous approach. Applications of the new approach to parameter estimation and model selection are also explored.

1 Introduction

Inference to the best explanation (IBE) has often been proposed and defended as a mode of reasoning in both science and everyday life (see Lipton [2004]). IBE is also of particular relevance to debates about scientific realism since proponents of realism often appeal to it (see Psillos [1999]). The basic idea in IBE is that competing hypotheses are compared in terms of how well they explain the evidence in a given context and an inference made to the winning hypothesis. While IBE has intuitive appeal, it has come in for serious criticism (see particularly van Fraassen

[1989]), though significant defences have also been presented (see particularly Douven [1999], [2013]).

A significant challenge for advocates of IBE is to show how explanation is related to truth. Clearly, such a link is fundamental to IBE, but it is not immediately obvious why hypotheses that provide better explanations would be more likely to be true. One approach to this problem is to consider probabilistic measures of explanatory goodness and investigate, either analytically or via computer simulations, how successful they are at selecting the true hypothesis. Various measures of this kind have been proposed in the literature (see Schupbach [2011a] and references therein), but here the focus is on using two measures of coherence to rank explanations.

There has been considerable discussion of probabilistic approaches to coherence in the literature (see for example Bovens and Hartmann [2003]; Olsson [2005]). In previous work, it was argued that a particular coherence measure, the overlap measure, has some merit as a measure for ranking explanations (Glass [2007]) and, based on simulations, that it tracks truth when used for hypothesis selection (Glass [2012]). It was further shown that when there is uncertainty in the prior probabilities of the hypotheses, this measure can outperform other approaches to hypothesis selection including the approach that maximizes the posterior probability.

This work, along with other probabilistic measures of explanatory power, has received some significant criticisms recently (Glymour [2015]). In addition to raising general concerns about the ability of these approaches to handle cases involving ‘excellent but false explanations’ and ‘causal explanations’, Prof. Glymour presents a number of specific criticisms of the coherence-based approach to

hypothesis selection. Perhaps the most significant criticism in this category is that in cases where there is uncertainty in the priors the advantage of this approach decreases rapidly as the sample size increases. The focus of the original paper (Glass [2012]) was on IBE in the context where a hypothesis is selected on the basis of a piece of evidence. In modelling this probabilistically, the simulations involved making an inference on a single trial and hence a sample size of one. As the current work will show, not only is Glymour's criticism correct, but even when priors are known the approach does not track truth as closely when the sample size is larger. As such, more work is needed if IBE is to be defended.

In this paper, another coherence measure is proposed to provide an alternative way to rank explanations and it is shown to overcome the problem noted above and to address several other concerns identified by Glymour. Section 2 provides a brief overview of the various measures and considers Glymour's criticisms. The new measure is then presented in section 3 and its performance when applied to hypothesis selection is considered when the priors are known in section 4 and when there is uncertainty in the priors in section 5. Applications of the new approach to parameter estimation and model selection are explored in section 6 before conclusions are drawn in section 7.

2 A Response to Prof. Glymour

For a hypothesis h that provides an explanation of e, the measures considered by Glymour include Schupbach and Sprenger's ([2011]) measure of explanatory power:

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

an alternative measure of explanatory power proposed by Crupi and Tentori ([2012]):

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

and the overlap coherence measure (OCM) used to rank explanations by Glass ([2007], [2012]):

$$\mathcal{E}_{OCM}(e,h) = \frac{P(h \wedge e)}{P(h \vee e)}. \tag{3}$$

The focus here is on responding to criticisms insofar as they apply to $\mathcal{E}_{OCM}$, but the other measures are included for comparative purposes since all three can be used to compare hypotheses in terms of how well they explain the evidence. Other measures considered by Glymour include Myrvold's ([2003]) measure of unification, Wheeler's ([2009]) measure of coherence or 'focused correlation' and Fitelson's ([2003]) measure of coherence, but these measures were proposed to address different problems and are not so relevant to the current paper.[1]

Before responding to criticisms, it is worth noting some differences between $\mathcal{E}_{OCM}$ on the one hand and $\mathcal{E}_{SS}$ and $\mathcal{E}_{CT}$ on the other. $\mathcal{E}_{SS}$ and $\mathcal{E}_{CT}$ are both measures of confirmation or relevance since it follows from their definitions that if h explains e, then the degree of explanatory power is greater than (equal to / less than) 0 if and only if $P(e|h)$ is greater than (equal to / less than) $P(e)$. Schupbach and Sprenger seek to explicate explanatory power in terms of the ability of a hypothesis 'to decrease the degree to which we find the explanandum surprising' ([2011], p. 108) and Crupi and Tentori adopt this approach also. $\mathcal{E}_{OCM}$ by contrast is not a confirmation measure and so a high value does not necessarily imply

[1]Though Fitelson's measure has been used to rank explanations and compared with $\mathcal{E}_{OCM}$ previously (Glass [2012]).

positive relevance. This is because a high prior probability for a hypothesis could compensate to some extent for negative relevance.

A further difference is that in proposing $\mathcal{E}_{OCM}$ the focus was on using it in the context of IBE to compare different hypotheses in terms of how well they explain a common explanandum, e, so that h_1 is to be preferred over h_2 as an explanation for e if $\mathcal{E}_{OCM}(e, h_1) > \mathcal{E}_{OCM}(e, h_2)$ (see Glass [2012]). $\mathcal{E}_{SS}$ and $\mathcal{E}_{CT}$ are intended not merely for comparative purposes, but to provide a satisfactory absolute value of explanatory power. Of course, they can then be used to compare hypotheses, but in fact for a given explanandum both $\mathcal{E}_{SS}$ and $\mathcal{E}_{CT}$ will give the same ordering of hypotheses and this turns out to be the same ordering as given by the likelihood, i.e. h_1 is to preferred over h_2 as an explanation for e if $P(e|h_1) > P(e|h_2)$.

Much more could be said by way of comparison of the different approaches. For example, arguably $\mathcal{E}_{SS}$ and $\mathcal{E}_{CT}$ can be construed as alternative measures of the extent to which a hypothesis leads us to expect the explanandum given the truth of that hypothesis, whereas $\mathcal{E}_{OCM}$ also attempts to take into account how plausible the hypothesis is in the first place. This might mean that $\mathcal{E}_{SS}$ and $\mathcal{E}_{CT}$ should not be seen as rivals to $\mathcal{E}_{OCM}$, but rather as explicating a different concept. However, a detailed comparison is beyond the scope of this paper and so the above discussion is just intended to provide some relevant context before considering Glymour's criticisms.

2.1 Excellent but false explanations

The first problem arises from the fact that scientific theories can provide excellent explanations even though they are now known to be false. As Glymour points out, 'Newtonian explanations of various planetary orbits, or of Kepler's laws, still

count as explanations although the falsity of Newtonian theory is settled' ([2015], p. 594). The difficulty is that if the probability assigned to a theory is zero, the measures 'that are functions of the unconditional or conditional probability of a hypothesis are either undefined or give perverse results'. Note that $\mathcal{E}_{OCM}$ will have the value zero in such cases.[2]

Glymour discusses a suggestion from a reviewer to treat historical hypotheses as true and make the probabilities to be the subjective probabilities of those who justly accepted the theories. He rightly raises concerns about this proposal, mentioning problems about how to compare the explanatory power of historical theories that cannot both have probability one as well as difficulties concerning probability and acceptance. Apart from this particular proposal, though, Glymour also raises more general problems about which probabilities are to be used. Should we use our marginal probabilities for the hypotheses and the evidence or those of historical figures or perhaps some combination of the two? And this is further complicated by the fact that the same question can be raised for the various likelihoods. If we use our probability for the evidence when dealing with past cases then the probability of the evidence will be one, which presents problems for confirmation measures like $\mathcal{E}_{SS}$ and $\mathcal{E}_{CT}$.

Of course, many of these and related questions can be raised in the context of Bayesianism, but those debates will not be revisited here. To keep the discussion more focused, let us consider whether there are problems here that go beyond those faced by the Bayesian. And it seems there are. In dealing with historical

[2]If regularity is assumed, then the theory will not be assigned a probability of zero, but if it is assigned a very low value, then $\mathcal{E}_{OCM}$ will also be very low provided the probability of the evidence is not too low.

cases, Bayesians will attempt to capture key aspects of scientific reasoning by using probabilities (both marginal and conditional) that would have been reasonable in that context. Now, of course, questions can be raised about that, but proponents of the measures of explanatory power might appear to be making ahistorical claims about scientific theories and in that case Glymour's concerns seem particularly pertinent.

In response, it can be noted that this appearance is misleading, at least in the context in which $\mathcal{E}_{OCM}$ was proposed. As noted earlier, it was proposed to compare explanations in the context of IBE. This means that it only applies, just as Bayesianism does, in contexts where there are competing hypotheses, each with a probability greater than zero and less than one. Hence, $\mathcal{E}_{OCM}$ was not intended to provide a value of how well a theory that is now known to be false explains some body of evidence, a value which could then be compared with that of a current theory. If we are using $\mathcal{E}_{OCM}$ to implement IBE in order to compare current viable theories, it seems appropriate that $\mathcal{E}_{OCM}$ gives a value of zero to theories that are known to be false since they are no longer viable options and hence do not need to be considered in IBE. Alternatively, if we are using $\mathcal{E}_{OCM}$ to implement IBE in order to capture aspects of scientific reasoning in a historical case then a theory that is now known to be false could be given a non-zero probability (and non-zero value of $\mathcal{E}_{OCM}$) if it was a viable option at the time. Also, since the focus is on the comparative use of this measure, precision in determining the absolute values is not so important provided reasonable judgments can be made about the ranking of explanations.

In light of the comments made earlier, it is not so clear whether this response is open to the proponents of $\mathcal{E}_{SS}$ and $\mathcal{E}_{CT}$. It is doubtful whether they can focus

solely on comparative judgments since these measures will give the same ranking of hypotheses for a given explanandum. Furthermore, Glymour's concerns would need to be addressed if the explanatory merits of theories now regarded to be false are to be compared with those of current theories. However, by restricting their measures to particular historical contexts perhaps a response along similar lines is possible.

The claim here is that in terms of concerns about which probabilities should be used, if $\mathcal{E}_{OCM}$ is used as intended in the context of IBE, it faces no problems over and above those faced by Bayesianism. This of course will seem like a weak response to those who think that Bayesianism has no adequate response to these kinds of concerns, but to go further than that in this paper would take us too far off track.

A possible objection to the response outlined here is that since the approach is not intended to quantify the explanatory power of theories that are known to be false, it is inadequate as an approach to IBE; instead, it should be seen as an approach to model selection.[3] Clearly, as Glymour has pointed out, false theories can still count as explanations. Hence, if one assumes that any approach to IBE based on the sorts of quantitative measures considered here should provide reasonable non-zero values to false theories, which could then be compared with other theories, then this is a reasonable objection. However, it is not clear that IBE requires such a general comparison of theories. In his important book on IBE, Peter Lipton characterized it in terms of inferring the best explanation out of a pool of potential explanations. One option he considered would be to include all possible explanations within the pool, but Lipton's favoured option was to 'define the pool

[3]I would like to thank a reviewer for raising this objection.

more narrowly, so that the potential explanations are only the "live options": the serious candidates for an actual explanation' (Lipton [2004, p.59]). In either case, however, the pool of potential explanations would not include any theories known to be false. Understood in this way, using a measure such as $\mathcal{E}_{OCM}$ to compare 'live options' is compatible with IBE.

The suggestion that the approach being advocated should be understood in terms of model selection is reasonable, but understanding it in this way need not be incompatible with IBE. Whatever the merits of Bayesianism, its proponents seek to apply it not only as a general approach to scientific inference in contexts where IBE might be applied, but also to the problem of model selection in statistics. The current work could be seen as an attempt to do something similar for IBE. By articulating IBE in terms of probabilistic measures and evaluating its performance in various inference tasks, it provides an initial step towards an IBE-based approach to model selection. More will be said about the connection with model selection in section 6.

2.2 Causal explanation

Suppose e_1 and e_2 are probabilistically independent effects of a common cause, h. If h raises the probability of e_1 and e_2, Glymour points out that the various measures (apart from Myrvold's and Wheeler's) give a positive value. Glymour's concern is that the measures 'confound probability or predictive relations with explanatory relations' ([2015], p. 596). He acknowledges that a possible response to this problem is to claim that the measures are 'to be applied only in cases which, on other grounds, an explanatory relation obtains between two propositions' (Glymour [2015], p. 596). This point is emphasized by both Glass ([2007])

and Schupbach ([2011]), but Glymour thinks that this is inadequate.

To make his point he discusses another case where there is a common cause, h say, of e_1 and e_2, but this time e_1 is also a cause of e_2. He considers the particular case in which the probability relation between e_1 and e_2 arising from the causal pathway between them is exactly cancelled by the probability relation between them due to their common cause. He claims that 'even though the value of e_1 causes and helps to explain the value of e_2, according to all the probabilistic measures of explanatory power, e_1 has no power to explain e_2' (Glymour [2015], p. 596). While this is correct for the $\mathcal{E}_{SS}$ and $\mathcal{E}_{CT}$ measures since e_1 and e_2 are probabilistically independent and hence have an explanatory power of zero, it is not correct for the $\mathcal{E}_{OCM}$ measure. Consider the following example. Let

$P(e_2|e_1\&h) = 0.9$,

$P(e_2|{\sim}e_1\&h) = 0.8$,

$P(e_2|e_1\&\,{\sim}h) = 0.55$,

$P(e_2|{\sim}e_1\&\,{\sim}h) = 0.2$,

$P(e_1|h) = 0.3$,

$P(e_1|{\sim}h) = 0.8$,

$P(h) = 0.4$,

$P({\sim}h) = 0.6$.

It is easy to show that $P(e_2|e_1) = P(e_2|{\sim}e_1) = 0.62$ and hence e_1 and e_2 are independent even though they are positively dependent on each other given h and also given ${\sim}h$. If e_1 is a cause of e_2, then the positive influence between e_1 and

e_2 is cancelled by the probability relation between them due to their common cause. However, it turns out that $\mathcal{E}_{OCM}(e_2, e_1) = 0.4387$ whereas $\mathcal{E}_{SS}(e_2, e_1) = \mathcal{E}_{CT}(e_2, e_1) = 0$. Furthermore, $\mathcal{E}_{OCM}(e_2, \sim e_1) = 0.3212$ and so e_1 provides a better explanation of e_2 than does $\sim e_1$. Hence, $\mathcal{E}_{OCM}$ deals with this case in a satisfactory manner.

While this response is not open to advocates of $\mathcal{E}_{SS}$ and $\mathcal{E}_{CT}$, they could argue that in cases involving common causes the explanatory power is to be determined by conditioning on each value of the common causes. These measures will typically give different values, as will $\mathcal{E}_{OCM}$, in examples such as the one above when conditioning on h and $\sim h$ respectively. However, Glymour considers and rejects this response precisely on the grounds that 'conditioning on different values of the common causes will give different values to the measures of explanatory power' ([2015], p. 596). It is difficult to see why this is a problem. In the example above, the focus is on how well e_1 explains e_2, but since h is also causally and hence explanatorily relevant to e_1 and e_2, there is no reason to think that the extent to which e_1 explains e_2 should be the same irrespective of whether h is true or false. For example, when considering the likelihoods in the example given above, it is clear that the difference (and ratio) between the probability of e_2 given e_1 and e_2 given $\sim e_1$ is greater when conditioning on h than on $\sim h$.

In a final point about causal explanation, Glymour draws attention to an approach to explanatory power based on the difference between the probability of the effect, e, when a manipulation is carried out to force a causal variable H to take on the value true and the corresponding probability when a manipulation forces H to be false. Several points can be made in response. First, while there is certainly merit to this approach it is not immediately clear that it should be seen

as a rival to the approaches based on the various measures that are criticized by Glymour and, even if it is, more detailed argument would be need to show that it is superior to these approaches. Second, manipulation can be incorporated into these approaches (see Eva and Stern [2018]). Finally, while incorporating manipulation seems appropriate for quantifying causal influence, it is not clear that it is the right approach for quantifying explanatory power. As discussed earlier, $\mathcal{E}_{SS}$ and $\mathcal{E}_{CT}$ can be construed as measures of the extent to which a hypothesis leads us to expect the explanandum and it seems plausible to incorporate manipulation into measures of this kind. $\mathcal{E}_{OCM}$ also attempts to take into account how plausible the hypothesis is in the first place, but this is missing when manipulation is taken into account since the relevant hypothesis variable is simply forced to be true or false.

2.3 Finding the truth

In addition to the problems discussed so far, Glymour also draws attention to the limitations of the various measures in hypothesis selection and he focuses in particular on the use of the $\mathcal{E}_{OCM}$ measure by Glass ([2012]) since it seemed to show some merit in this regard.[4] A general problem for using these measures as statistical tests concerns how various probabilities, such as $P(e|\sim h)$ and $P(e)$, are to be acquired. Focusing on $P(e)$, he rightly points out that 'hypothesis selection by statistical testing requires comparing ratios so that the probability of the evidence does not appear' (Glymour [2015], p. 601). This criticism is quite legitimate and

[4]Glymour also presents criticisms of the psychological study carried out by Schupbach ([2011a]), but these will not be considered here.

so whatever the merits of the $\mathcal{E}_{OCM}$ in hypothesis selection, this presents a limitation on its practical usage. A new measure will be proposed in section 3 to address this problem.

Previous work employed computer experiments to compare how well various measures performed in a hypothesis selection task (Glass [2012]). Prior probabilities of hypotheses and likelihoods were randomly selected from uniform distributions and then one of the hypotheses was selected as the actual hypothesis based on the prior distribution. Either e, or $\sim e$ was then selected based on the likelihood for the actual hypothesis. Various measures were then used to select the best hypothesis (i.e. the one that had the highest score for a given measure) and if it matched the actual hypothesis it was counted as a success for that measure. This process was repeated ten million times to get an accurate picture of how well the various measures performed. The results showed that selecting the hypothesis with the greatest value of $\mathcal{E}_{OCM}$ gave results that closely tracked hypothesis selection by maximum posterior probability and outperformed all the other measures, including maximum likelihood.

The same paper also considered cases where there is uncertainty as modelled by a normal distribution in the prior probabilities or, to put it another way, where the prior probabilities are subjective and do not correspond to the objective probabilities. These subjective priors along with the actual likelihoods were then used in the various measures and the experiments re-run. The results showed that if there is sufficient uncertainty in the priors, then selecting the hypothesis with the greatest value of $\mathcal{E}_{OCM}$ gave more accurate results than hypothesis selection by maximum posterior probability.

Referring to work by Teng et al. ([unpublished]), Glymour notes that the re-

sults do not depend on the choice of the normal distribution to model uncertainty since similar results are obtained when a uniform distribution is used (see section 5 below for further discussion of this point). However, Glymour raises several concerns about these findings. First, he claims that the advantage of the $\mathcal{E}_{OCM}$ measure vanishes with small errors in the specification of likelihoods. It is true that the advantage over maximum posterior probability is lost when there are errors in the likelihoods, but as shown by experiments carried out by Glass and McCartney ([2014]), when there are errors in the likelihoods but not the priors, the approach based on the $\mathcal{E}_{OCM}$ measure still tracks maximum posterior probability quite closely and it performs much better than the other measures considered, including maximum likelihood.[5]

Second, Glymour says that the approach based on the $\mathcal{E}_{OCM}$ measure never dominates maximum likelihood. However, as just noted, it does in fact dominate it when there are errors in the likelihoods but not in the priors. It also dominates maximum likelihood in cases where there are no errors (for small sample sizes) in either the priors or the likelihoods. Where it fails to dominate maximum likelihood is in the case where there are errors in the priors. However, of the three approaches compared in this case (Glass [2012]) — maximum posterior probability, maximum likelihood and the approach based on the $\mathcal{E}_{OCM}$ measure — no approach dominates the other two for the range considered, which was for values of 0 to 1 for standard deviations in the prior probabilities. However, the $\mathcal{E}_{OCM}$ measure performed better than maximum likelihood for values from 0 to 0.7 and better

[5]More experiments would need to be carried out to investigate cases where there are errors in both priors and likelihoods, but based on experiments that have been carried out it seems likely that unless the errors for the prior were very large the $\mathcal{E}_{OCM}$ measure would perform better than maximum likelihood.

than maximum posterior probability for values from 0.4 to 1. Moreover, when the results were averaged over the entire range the $\mathcal{E}_{OCM}$ measure performed best.

Third, Glymour points out that in cases where there are no errors in the prior probabilities the advantage of the $\mathcal{E}_{OCM}$ measure over other measures vanishes when the sample size increases.[6] In fact, as we shall see in section 4 its advantage over maximum likelihood vanishes for a sample size of about 15, and so a significant advantage only occurs for very small sample sizes. This, alas, is correct and it also applies when there are errors in the priors (also shown in section 4).

This issue is closely related to another point mentioned by Glymour: the asymptotic behaviour of the various measures. The measure $\mathcal{E}_{OCM}(e,h)$ can be expressed as $\left[\frac{1}{P(h|e)} + \frac{1}{P(e|h)} - 1\right]^{-1}$. Since it incorporates both the likelihood and the posterior probability, this explains why it gives better results than maximum likelihood for small sample sizes (when there are no errors in the priors or likelihoods). What is the asymptotic behaviour of $\mathcal{E}_{OCM}$ in the limit of large sample sizes? Let e_n represent the evidence for a sample size of n. Taking the limit of the ratio of $\mathcal{E}_{OCM}$ to the likelihood gives:[7]

$$\lim_{n\to\infty} \frac{\mathcal{E}_{OCM}(e_n,h)}{P(e_n|h)} = 1, \tag{4}$$

which explains why the advantage of $\mathcal{E}_{OCM}(e,h)$ over maximum likelihood decreases as the sample size increases. This behaviour is evident from results pre-

[6] Glymour actually refers to the advantage over posterior probability, but since there is no advantage over posterior probabilities unless there are errors in the priors he presumably means the advantage over other approaches such as maximum likelihood.

[7] $\mathcal{E}_{OCM}(e_n,h) = P(e_n|h)P(h)/P(e_n \vee h)$ and hence the ratio can be written as $P(h)/P(e_n \vee h)$. Assuming the probability of each outcome is less than one, then the limit of $P(e_n)$ as n tends to infinity is zero and hence the limit of $P(e_n \vee h)$ as n tends to infinity is $P(h)$. Hence, the limit of $\mathcal{E}_{OCM}(e_n,h)/P(e_n|h)$ as n tends to infinity is 1.

sented in section 4.

In summary, Glymour has raised a number of very interesting challenges for measures of explanatory goodness. In offering a response, the focus has been on the $\mathcal{E}_{OCM}$ measure. It has been argued that the problem of 'excellent but false explanations' is no more serious for this measure than it is for Bayesianism and that adequate responses are available in the case of causal explanations. However, in the context of using $\mathcal{E}_{OCM}$ for hypothesis selection, there are two significant issues to which no adequate response has been provided. First, using this measure to compare hypotheses requires obtaining probabilities which are difficult to determine in practice and, second, the advantage of the $\mathcal{E}_{OCM}$ vanishes for larger (but still relatively small) sample sizes.

It could perhaps be argued that these problems serve to highlight limitations of using the $\mathcal{E}_{OCM}$ measure in practice, but do not undermine it as a measure that can be used to show that IBE tracks truth. Indeed, when IBE is formulated using $\mathcal{E}_{OCM}$ it tracks results obtained by maximizing posterior probability closely for very small sample sizes, performing better than maximum likelihood, and as the result in (4) shows, it gives the same results as maximum likelihood in the limit of large sample sizes. Nevertheless, a new measure will now be considered in order to address some of the practical limitations of the the $\mathcal{E}_{OCM}$ measure.

3 A New Measure for Comparing Explanations

For a hypothesis h that explains e, a simple measure of how good an explanation it is can be defined as the product of the likelihood and posterior probability:

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

which will be referred to as the product coherence measure (PCM). Note that given Bayes' theorem, it can also be expressed as $P(e|h)^2P(h)/P(e)$ or alternatively as $P(h|e)^2P(e)/P(h)$.

As the name suggests, $\mathcal{E}_{PCM}$ can be considered as a coherence measure. In fact, it has many features in common with the overlap coherence measure, $\mathcal{E}_{OCM}$. Clearly, its range is the interval $[0,1]$ with $\mathcal{E}_{PCM}(e,h) = 0$ when $P(e|h) = P(h|e) = 0$ and $\mathcal{E}_{PCM}(e,h) = 1$ when $P(e|h) = P(h|e) = 1$. If it is just used as a coherence measure (in which case it need not be assumed that h explains e according to some account of what constitutes an explanation) this means that consistent logically equivalent beliefs are maximally coherent (for example, $\mathcal{E}_{PCM}(e,e) = 1$), while logically inconsistent beliefs are incoherent (for example, $\mathcal{E}_{PCM}(e,{\sim}e) = 0$). Like $\mathcal{E}_{OCM}$, $\mathcal{E}_{PCM}$ depends only on the extent of agreement between two beliefs rather than how probable those beliefs are in the first place. More precisely, for fixed values of the relevant conditional probabilities, $P(e|h)$ and $P(h|e)$, it is independent of prior probabilities.[8]

Hence, like the overlap measure, $\mathcal{E}_{PCM}$ exhibits the characteristics of a particular type of coherence — coherence as agreement rather than as striking agreement — which has been argued in previous work to capture certain intuitions about coherence better than other approaches (Glass [2005]). However, arguably it also has some advantages over the overlap measure. Note that $\mathcal{E}_{PCM}$ can be

[8]See (Glass [2005]) for further discussion of this point. In fact, $\mathcal{E}_{PCM}$ was mentioned in a footnote in this paper.

expressed as:

$$\mathcal{E}_{PCM}(e,h) = \frac{P(e \wedge h)}{P(e)P(h)} \times P(e \wedge h), \tag{6}$$

where the first term on the right hand side is Shogenji's measure of coherence (Shogenji [1999]). Like many of the coherence measures proposed in the literature, Shogenji's measure is also a relevance or confirmation measure (in fact, it is the ratio measure, $P(h|e)/P(h)$ of confirmation), whereas this is not true of either $\mathcal{E}_{OCM}$ or $\mathcal{E}_{PCM}$. However, as equation (6) shows, $\mathcal{E}_{PCM}$ can be expressed as the product of a confirmation measure and the joint probability. By taking confirmation into account, it is sensitive to the dependence between h and e in a way that $\mathcal{E}_{OCM}$ is not.

Consider an example adapted from Bovens and Olsson ([2000]) to highlight the difference between coherence as agreement and coherence as striking agreement. Suppose there is a roulette wheel with one hundred numbers and in the first scenario Joe says the winning number is 49 or 50 and Amy says it is 50 or 51. In the second, scenario Joe says the winning number is 41, 42, ..., or 60 and Amy says it is 51, 52, ..., or 70. As measures of coherence as agreement, $\mathcal{E}_{OCM}$ or $\mathcal{E}_{PCM}$ will each yield the same degree of coherence in both cases, 1/3 for $\mathcal{E}_{OCM}$ and 1/4 for $\mathcal{E}_{PCM}$. However, now consider a third scenario where Joe says the winning number is 41, 42, ..., or 50 and Amy says it is 41, 42, ..., or 70. In this case, $\mathcal{E}_{OCM}$ gives the same result of 1/3 as in scenarios one and two since the relative overlap is unchanged, but $\mathcal{E}_{PCM}$ gives a higher coherence than in the other two scenarios (1/3 compared to 1/4). The difference lies in the fact that there is stronger dependence between the statements in this case since Joe's being correct entails that Amy is also correct.[9]

[9]This difference between $\mathcal{E}_{OCM}$ and $\mathcal{E}_{PCM}$ leads to a possible disadvantage to $\mathcal{E}_{PCM}$ because

A well known problem with the overlap coherence measure is that if it is extended to multiple belief in the obvious way so that $\mathcal{E}_{OCM}(h_1, \ldots, h_n) = P(h_1 \wedge \ldots \wedge h_n)/P(h_1 \vee \ldots \vee h_n)$, then coherence cannot increase as the number of beliefs increases. However, it seems clear that acquisition of further beliefs can enhance the coherence of previously held beliefs. For example, 'Tweety is a bird' does not cohere well with 'Tweety cannot fly', but combining these beliefs with 'Tweety is a penguin' results in much greater coherence. In light of equation (6), an obvious way to extend $\mathcal{E}_{PCM}$ to the general case is as follows:

$$\mathcal{E}_{PCM}(h_1, \ldots, h_n) = \frac{P(h_1 \wedge \ldots \wedge h_n)}{P(h_1) \ldots P(h_n)} \times P(h_1 \wedge \ldots \wedge h_n), \tag{7}$$

which avoids the problem.[10] Suppose for example that $h_3 = h_1 \wedge h_2$, then it is easy to show that $\mathcal{E}_{PCM}(h_1, h_2, h_3) > \mathcal{E}_{PCM}(h_1, h_2)$. Extending $\mathcal{E}_{PCM}$ in this way makes it a sort of hybrid between measures of agreement, which it is in the case of two beliefs, and measures of striking agreement. Arguably, this means it is able to capture the merits of different measures such as the overlap coherence measure and Shogenji's measure.

In previous work (Glass [2007]), it was argued that while there is merit to

$\mathcal{E}_{OCM}$ is known to be truth-conducive in the case of information pairs (Glass [2007]). It follows from these differences that $\mathcal{E}_{PCM}$ is not truth-conducive for the same set of *ceteris paribus* conditions, though it can be shown to be truth-conducive if more specific *ceteris paribus* conditions are defined. However, arguably truth-conduciveness should not be considered as an adequacy condition for coherence: coherence measures should be judged on other criteria and then their consequences for truth-conduciveness evaluated. As argued here, there are reasons to prefer $\mathcal{E}_{PCM}$ to $\mathcal{E}_{OCM}$ as a measure of coherence and it turns out to have better consequences for explanatory inference.

[10] Other generalizations could also avoid this problem. For example, it could be generalized to take j-wise (in)dependence, where $j < n$, into account, see Schupbach ([2011b]).

using either the posterior probability of a hypothesis, $P(h|e)$, or its likelihood, $P(e|h)$, to rank explanations, these approaches also face problems. It was also claimed that $\mathcal{E}_{OCM}$ takes account of their advantages while avoiding their problems. The same points could be made in favour of $\mathcal{E}_{PCM}$ since it is also a combination of posterior probability and likelihood. One particular reason why $\mathcal{E}_{PCM}$ is a plausible candidate for comparing explanations is that, like $\mathcal{E}_{OCM}$, it satisfies the explanation ranking condition (Glass [2007]), which can be stated as follows:

> For two hypotheses, h_1 and h_2 that explain e, if $P(e|h_1) > P(e|h_2)$ and $P(h_1|e) > P(h_2|e)$ then h_1 is a better explanation of e than h_2.

Furthermore, in terms of differences between $\mathcal{E}_{OCM}$ on the one hand and $\mathcal{E}_{SS}$ and $\mathcal{E}_{CT}$ on the other, the same points could be made about $\mathcal{E}_{PCM}$. In particular, like $\mathcal{E}_{OCM}$, $\mathcal{E}_{PCM}$ is not a confirmation measure and so a high prior probability for a hypothesis could compensate to some extent for negative relevance. In addition, the responses presented in defence of $\mathcal{E}_{OCM}$ in sections 2.1 and 2.2 to the 'excellent but false explanations' and 'causal explanation' objections, apply equally to $\mathcal{E}_{PCM}$.

Some advantages of $\mathcal{E}_{PCM}$ over $\mathcal{E}_{OCM}$ as a measure of coherence have been noted and one of these is also an advantage for $\mathcal{E}_{PCM}$ over $\mathcal{E}_{OCM}$ as a measure of explanation. As discussed earlier, equation (6) shows that $\mathcal{E}_{PCM}$ can be expressed as a product of a confirmation measure and the joint probability. Given that the approach to explanation in this paper takes into account how plausible the hypotheses are in the first place, there are reasons for preferring $\mathcal{E}_{OCM}$ or $\mathcal{E}_{PCM}$ to either $\mathcal{E}_{SS}$ or $\mathcal{E}_{CT}$ as a measure for comparing explanations. However, the dependence between the hypothesis and the evidence is clearly a factor in how well the hypothesis explains the evidence and this provides a reason to prefer $\mathcal{E}_{PCM}$

to $\mathcal{E}_{OCM}$ in the context of explanation. Suppose there are two hypotheses, h_1 and h_2, each of which provides a potential explanation of evidence e. Further suppose that $P(e) = 2/5$, $P(h_1) = 2/15$, $P(h_2) = 2/3$ and that h_1 entails e so that $P(e|h_1) = 1$ while h_2 is probabilistically independent of e so that $P(e|h_2) = 2/5$. It is easy to show that $\mathcal{E}_{OCM}(e, h_1) = \mathcal{E}_{OCM}(e, h_2) = 1/3$ and so $\mathcal{E}_{OCM}$ fails to discriminate between the hypotheses. By contrast, $\mathcal{E}_{PCM}(e, h_1) = 1/3 > 4/15 = \mathcal{E}_{PCM}(e, h_2)$. More generally, in cases where two hypotheses have equal relative overlap as measured by $\mathcal{E}_{OCM}$, but where one hypothesis, h_1, entails the evidence, while the other, h_2, neither entails nor is entailed by the evidence, the $\mathcal{E}_{PCM}$ measure will favour h_1.[11]

Recall that there were two problems for which no defence of $\mathcal{E}_{OCM}$ was provided in section 2. Does $\mathcal{E}_{PCM}$ fare any better? First, $\mathcal{E}_{OCM}$ requires the probability of the evidence to be determined and yet often this is unavailable. By contrast, the probability of the evidence is not needed when comparing either likelihoods or posterior probabilities. In the case of posterior probabilities, ratios can be taken, in which case the prior probability of the evidence cancels out so that just the likelihoods and priors are needed. Since $\mathcal{E}_{PCM}$ is just a product of the likelihood and posterior probability it also avoids the problem. Suppose two hypotheses h_1 and h_2 are to be compared. This can be done as follows:

[11] Since $\mathcal{E}_{OCM}(e, h)$ can be expressed as $\left[\frac{1}{P(h|e)} + \frac{1}{P(e|h)} - 1\right]^{-1}$, it follows that $\frac{1}{P(h_1|e)} + \frac{1}{P(e|h_1)} = \frac{1}{P(h_2|e)} + \frac{1}{P(e|h_2)}$. Furthermore, $P(e|h_1) = 1$ since h_1 entails e, and so $P(e|h_1)^{-1} < \min\{P(e|h_2)^{-1}, P(h_2|e)^{-1}\}$. Suppose $P(e|h_1)^{-1} = \min\{P(e|h_2)^{-1}, P(h_2|e)^{-1}\} - \delta$, where $\delta > 0$, and hence $P(h_1|e)^{-1} = \max\{P(e|h_2)^{-1}, P(h_2|e)^{-1}\} + \delta$. From this it can be shown that $P(e|h_1|e) \times P(h_1|e) > P(e|h_2) \times P(h_2|e)$ and so $\mathcal{E}_{PCM}(e, h_1) > \mathcal{E}_{PCM}(e, h_2)$.

$$\frac{\mathcal{E}_{PCM}(e,h_1)}{\mathcal{E}_{PCM}(e,h_2)} = \frac{P(e|h_1) \times P(h_1|e)}{P(e|h_2) \times P(h_2|e)} = \frac{P(e|h_1)^2 \times P(h_1)}{P(e|h_2)^2 \times P(h_2)}, \tag{8}$$

and hence there is no need to determine the probability of the evidence. Hence, from this practical point of view, $\mathcal{E}_{PCM}$ has a significant advantage over $\mathcal{E}_{OCM}$.[12]

Second, the advantage of comparing explanations using $\mathcal{E}_{OCM}$ over likelihood decreases with increase in sample size and this was explained by the fact that in the limit of large sample size the ratio of $\mathcal{E}_{OCM}$ to the likelihood tends to one, but this is not the case with $\mathcal{E}_{PCM}$. As with $\mathcal{E}_{OCM}$, $\mathcal{E}_{PCM}$ combines both the posterior probability and likelihood, but taking the limit of the ratio of $\mathcal{E}_{PCM}$ to the likelihood trivially gives:

$$\lim_{n\to\infty} \frac{\mathcal{E}_{PCM}(e_n,h)}{P(e_n|h)} = \lim_{n\to\infty} P(h|e_n), \tag{9}$$

and so the influence of the posterior probability is retained as sample size increases, which is not the case for $\mathcal{E}_{OCM}$ as is clear from equation (4). The significance of this contrast between $\mathcal{E}_{PCM}$ and $\mathcal{E}_{OCM}$ will be investigated experimentally in section 4.

Previous work made a case for $\mathcal{E}_{OCM}$ as a measure for comparing explanations Glass [2007, 2012] based on its properties as a coherence measure. However, as we saw in section 2, two shortcomings with this approach were identified. To address these issues, a new coherence measure has been proposed, $\mathcal{E}_{PCM}$. The previous arguments in support of $\mathcal{E}_{OCM}$ all apply to $\mathcal{E}_{PCM}$ as well, but further motivation for $\mathcal{E}_{PCM}$ has been provided both in terms of its merits as a measure of

[12]Of course, $\mathcal{E}_{PCM}$ requires the prior probabilities of the hypotheses, but in this respect it is no worse than hypothesis selection based on posterior probability. Indeed, as a reviewer has pointed out, it only requires the ratio of priors which may be easier to estimate. It also does not require probabilities such as $P(e|\sim h)$.

coherence and as a measure for comparing explanations. In particular, the way in which it incorporates dependence between the hypothesis and evidence has been highlighted as an important benefit in the context of explanation. Since $\mathcal{E}_{PCM}$ is also able to address the two issues identified, it has clear advantages for explanatory inference. In terms of how it deals with the second of these issues, it appears that inference based on $\mathcal{E}_{PCM}$ will still have an advantage over other approaches such as maximum likelihood as sample size increases. Computer simulations will now be used to investigate this in more detail.

4 IBE and Truth Tracking Revisited

In order to determine how the new measure fares when used in hypothesis selection, computer simulations were carried out as in the earlier paper (Glass [2012]). Consider the case of two mutually exclusive and jointly exhaustive hypotheses, h_1 and h_2, each of which can bring about either e or $\sim e$. A prior probability is obtained from the uniform distribution and assigned to $P(h_1)$ and hence $P(h_2) = 1 - P(h_1)$. By sampling this distribution, one hypothesis is selected and designated the actual hypothesis. Values are also obtained from a uniform distribution for the likelihoods of the hypotheses, $P(e|h_1)$ and $P(e|h_2)$. By sampling the conditional distribution for the actual hypothesis, the outcome is determined to be either e or $\sim e$.

Four hypothesis selection strategies can now be considered. Given knowledge of the prior probabilities, the likelihoods, and the outcome, each strategy uses a different measure to try to identify the actual hypothesis. The strategies are given

below:

MPE: most probable explanation; selects the hypothesis with the maximum posterior probability,

ML: selects the hypothesis with the maximum likelihood,

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

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

For a given strategy, if it correctly identifies the actual hypothesis, this is counted as a success, otherwise it is a failure. The process is then repeated 10,000,000 times with different priors and likelihoods and the accuracy of each strategy at selecting the actual hypothesis is determined. Further approaches could be defined based on the $\mathcal{E}_{SS}$ and $\mathcal{E}_{CT}$ measures presented in section 2, but these approaches give the same ordering as ML and so would yield identical results.

As described above, there are just two hypotheses and one outcome each time, but this can be generalized for multiple hypotheses, as in the earlier paper, and for multiple outcomes (i.e. increased sample size). For multiple outcomes, this is achieved by sampling the conditional distribution for the probability of the evidence given the actual hypothesis the required number of times to get a sequence $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ where x_i is either e or $\sim e$. The probability of this sequence of outcomes given each hypothesis can then be determined under the assumption that the outcomes are independent and identically distributed and these probabilities used in the four different strategies to select the best hypothesis. In Figure 1, percentage accuracy is plotted for each of the hypothesis selection approaches against sample size for the case of two hypotheses. As pointed out previously, the OCM approach achieves an accuracy very close to that of MPE when the sample

size is one, but as sample size increases it does not track the MPE result so closely. When the sample size is 20, for example, the accuracy of OCM is about the same as ML and some way short of MPE. This is consistent with the result in equation (4) which showed that the ratio of $\mathcal{E}_{OCM}$ to likelihood tends to one in the limit of large sample size. It should be noted that all of these approaches eventually converge to an accuracy of one as sample size increases, but it is clear that the OCM approach has limited merit when compared with ML since it only has a significant advantage for very low sample sizes.

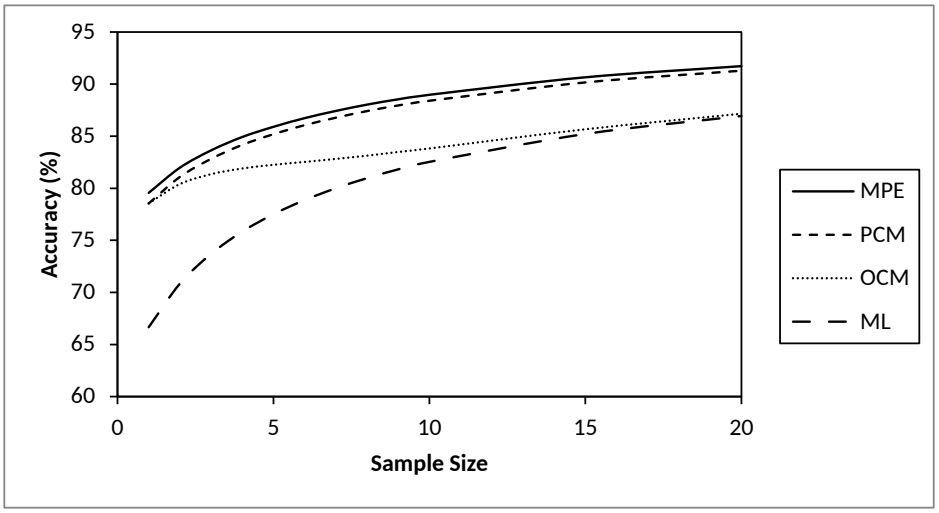


Figure 1: Accuracy plotted as a function of sample size for each of the different hypothesis selection approaches when there are two hypotheses.

By contrast, the new measure PCM obtains the same level of accuracy as OCM for a sample size of one, but continues to track the MPE result closely for larger sample sizes. Again, this is consistent with the result given in equation (9). This advantage of PCM over OCM is also obtained as the number of hypotheses varies. Previous work (Glass [2012]) found that the OCM result tracks MPE very closely

as the number of hypotheses increases when the sample size is one. Unfortunately, OCM does not perform so well when the sample size is larger as indicated in Figure 2. For a sample size of 20, OCM gets progressively worse relative to MPE as the number of hypotheses increases from 2 to 10. Again, however, PCM performs much better. It is much closer to MPE to start with and continues to track it closely for larger numbers of hypotheses.

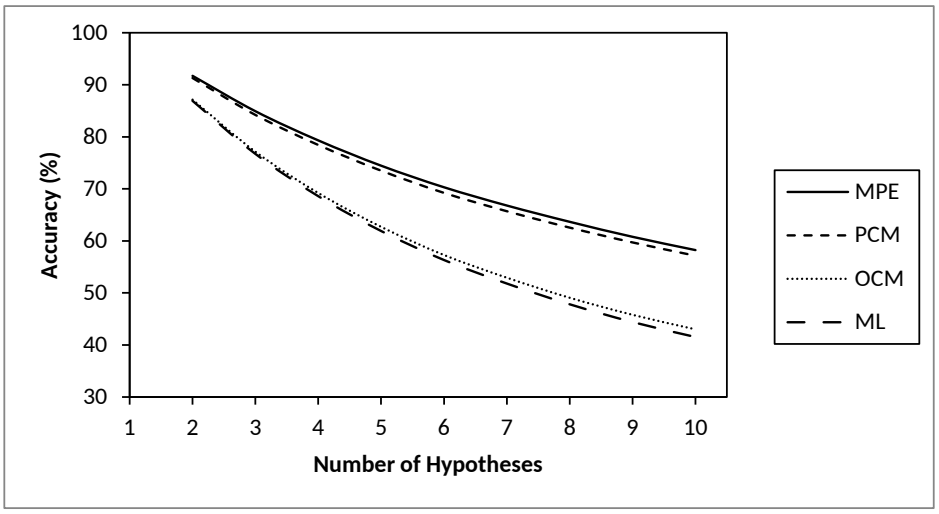


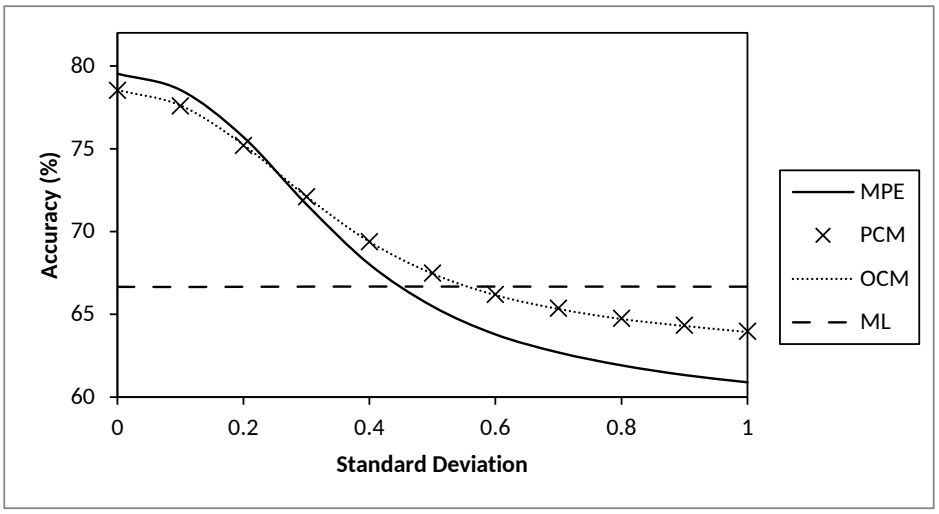
Figure 2: Accuracy plotted as a function of the number of competing hypotheses for each of the different hypothesis selection approaches when the sample size is 20.

These results for the PCM approach to hypothesis selection are very encouraging. In terms of IBE, if PCM is an appropriate measure for comparing explanations, then these results go some way to showing that IBE does track truth. Furthermore, it does significantly better than ML, and hence than approaches based on the measures $\mathcal{E}_{SS}$ and $\mathcal{E}_{CT}$. Let us now consider how PCM performs whenever there is uncertainty in the priors.

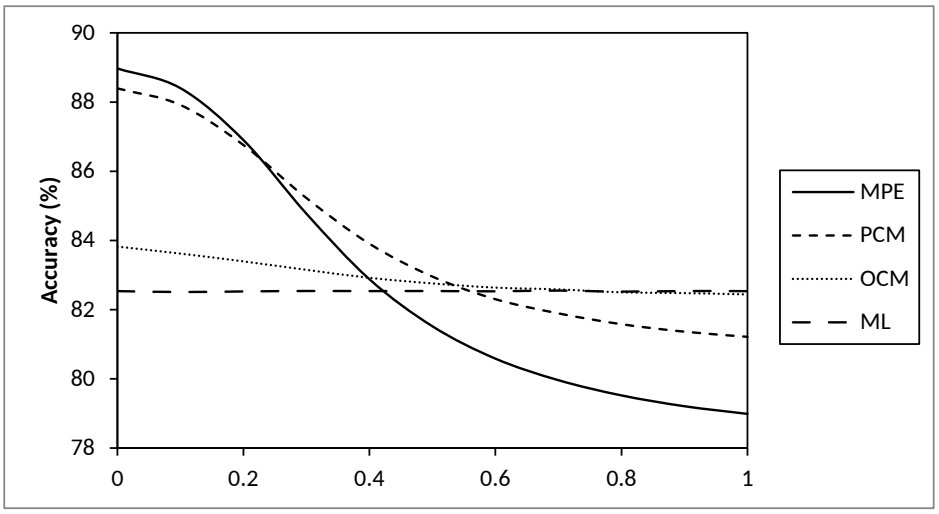
5 Hypothesis Selection under Uncertainty

This section implements the computer simulations carried out in the earlier paper (Glass [2012]), but now the PCM approach is included and the simulations are extended to the case where the sample size is greater than one. The probability model for priors and likelihoods is the same as in section 4, but in this case the true prior probability distribution is no longer assumed to be known. Instead of adopting the true prior for h_1, an incorrect value is obtained in the following way. A number is drawn from a normal distribution with mean zero and a specified standard deviation and this is added to the true prior $P(h_1)$. We can think of this as an agent's subjective prior, $P'(h_1)$, provided it lies between 0 and 1. If it does not, the process is repeated until a value is obtained in the desired interval. The corresponding value for h_2 is then $P'(h_2) = 1 - P'(h_1)$. This provides a way of representing uncertainty in the agent's knowledge of priors.

Results are presented in Figure 3 for the case of two hypotheses. In Figure 3a, the results are for a sample size of one and so correspond to Figure 3 in (Glass [2012]). Note that the results for PCM and OCM are indistinguishable. As expected, when the standard deviation is small, corresponding to low uncertainty, MPE outperforms all the other approaches. It is also not surprising that for very large values, corresponding to a high degree of uncertainty, ML outperforms all other approaches since it does not depend on the priors at all. However, for intermediate values between about 0.25 and 0.55 PCM and OCM outperform both MPE and ML. Indeed, as reported in the earlier paper, if the results are averaged



(a)



(b)

Figure 3: Accuracy plotted as a function of the standard deviation in the case of two hypotheses for a sample size of (a) 1, and (b) 10.

over the entire range then PCM and OCM come out on top. This suggests that these approaches are best for hypothesis selection if the degree of uncertainty in

the priors is unknown.

However, as Glymour ([2015]) has pointed out there is a problem with the OCM approach for greater sample sizes. This is illustrated in Figure 3b where the sample size is 10. Note that for low values of the standard deviation, OCM performs much worse than MPE and it has little advantage over ML. Note, however, that the results are very different for PCM. It still exhibits a similar advantage over MPE and ML as in Figure 3a.

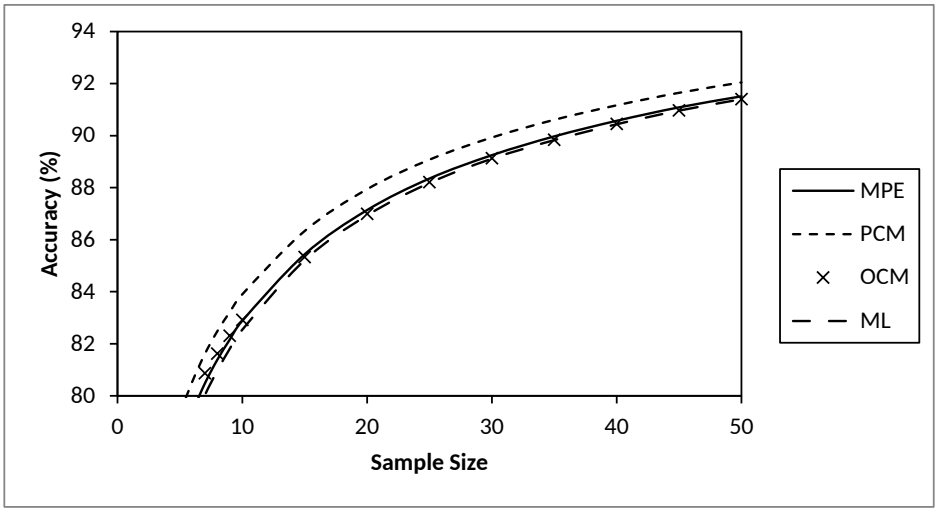
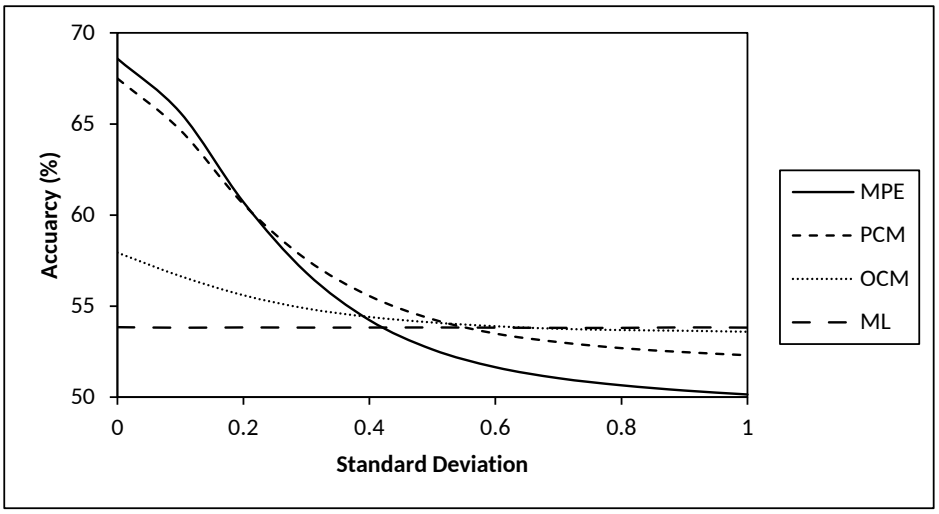
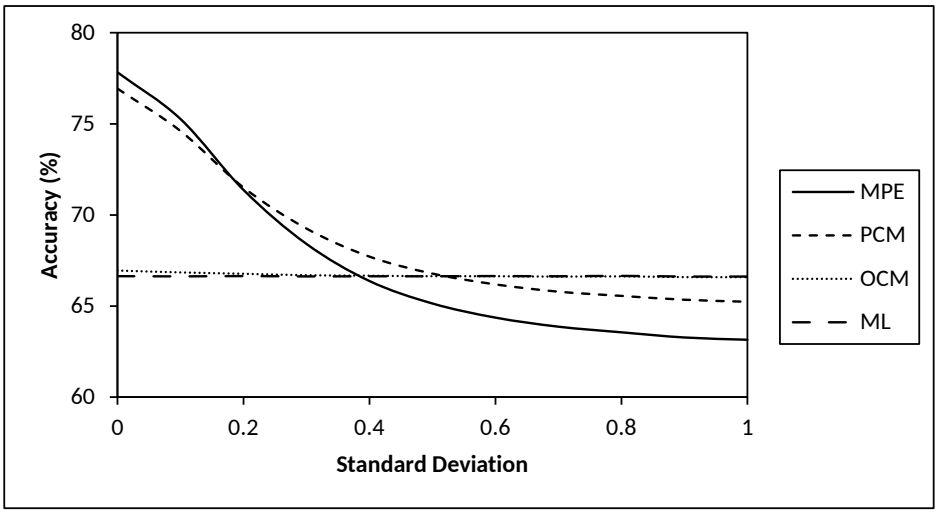


Figure 4: Accuracy plotted as a function of sample size in the case of two hypotheses for a standard deviation of 0.4.

Figure 4 presents results for a fixed value of the standard deviation (0.4) as a function of sample size for the case of two hypotheses. This is a value where PCM and OCM outperform MPE and ML for a sample size of one as can be seen from Figure 3a, but when the sample size has increased to 10, OCM has lost any advantage it had. By contrast PCM retains its advantage for much larger sample sizes.



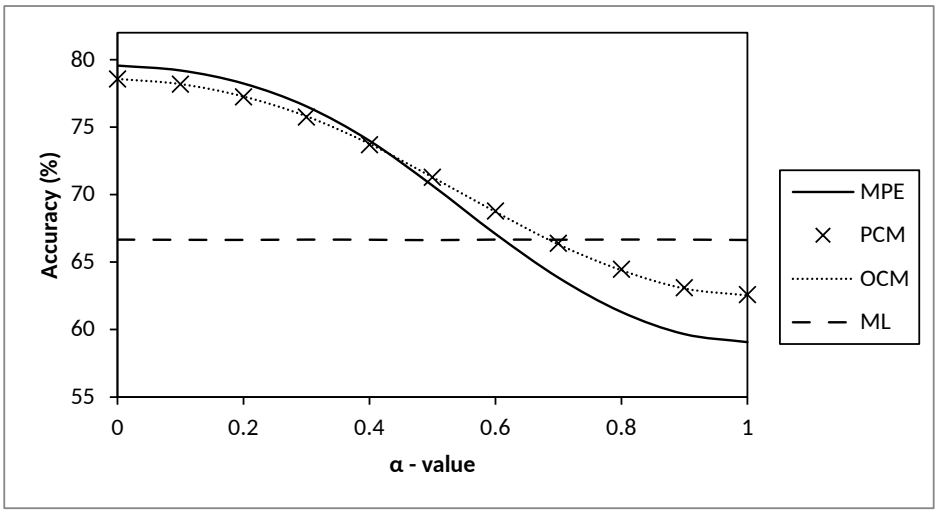
(a)



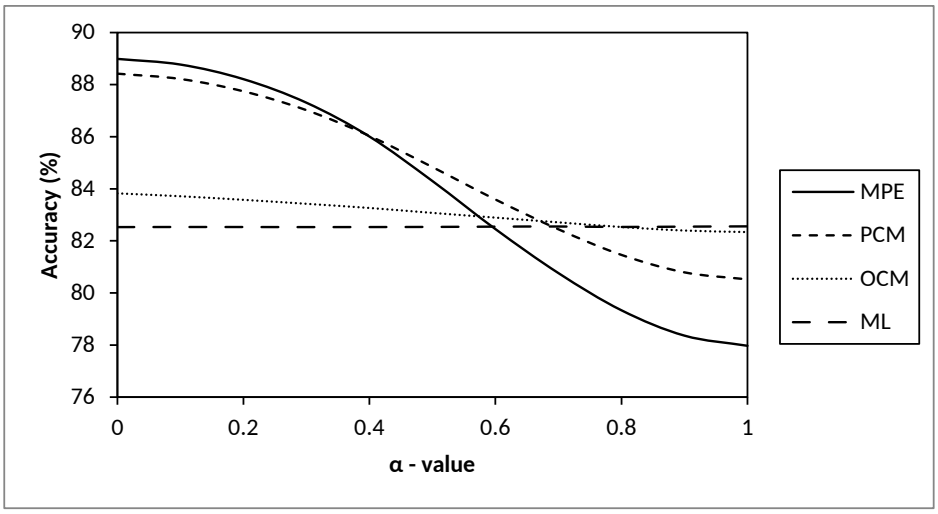
(b)

Figure 5: Accuracy plotted as a function of the standard deviation in the case of five hypotheses for a sample size of (a) 10, and (b) 30.

Figure 5 presents results for the case of five hypotheses. Clearly, similar behaviour is found. The relative merits of PCM over MPE and ML are very similar



(a)



(b)

Figure 6: Accuracy plotted as α varies in the uniform distribution used to represent uncertainty in the case of two hypotheses for a sample size of (a) 1, and (b) 10.

for sample sizes of 10 and 30, whereas OCM has lost most of its advantage over MPE and ML for a sample size of 10 and is almost indistinguishable from ML for

a sample size of 30.

As pointed out in section 2.3, Glymour ([2015]) notes that the results for OCM do not depend on the normal distribution to model uncertainty in the priors since similar results are obtained with a uniform distribution. In order to investigate this point further and in particular to see whether the same applies for PCM, the following experiment has been carried out. Priors are drawn from a uniform distribution in the interval $[\max\{P(h_1)-\alpha, 0\}, \min\{P(h_1)+\alpha, 1\}]$ (or equivalently from the interval $[P(h_1) - \alpha, P(h_1) + \alpha]$ with resampling if the value does not lie between 0 and 1). Results obtained for the case of two hypotheses and sample sizes of one and 10 are presented in Figure 6. Overall, these results illustrate very similar behaviour to those presented for the normal distribution in Figure 3. Note, however, that PCM only outperforms MPE for values greater than about 0.4, whereas in the case of the normal distribution it outperforms MPE for a standard deviation of about 0.25. This can be explained as follows. For $\alpha = 0.25$ the subjective priors are within 0.25 of the true value whereas this is not the case for a standard deviation of 0.25 where the subjective priors can be much greater. Hence, it seems plausible that $\alpha = 0.25$ corresponds to a lower degree of uncertainty than a standard deviation of 0.25 and so the advantages of PCM are greater in the latter case.

Hence, in addition to the success of the PCM approach when the priors are known (section 4), the results in this section show that it also performs much better than OCM when the priors are not known accurately. Its advantages persist when there are more hypotheses to be compared, when the sample size increases, and when uncertainty in the priors is modelled in a different manner.

6 Parameter Estimation and Model Selection

The focus so far has been on applying the PCM and OCM approaches to IBE and, in particular, to selecting the best hypothesis from a finite number of mutually exclusive hypotheses. However, we can think of the different hypotheses as different values of a discrete random variable and this raises the question whether a similar approach can be applied to continuous variables. To answer this question the PCM approach will be applied to simple examples of parameter estimation, where the goal is to obtain point estimates of an unobserved quantity, and the results compared with two other approaches. After that, we consider whether the PCM approach can be applied to the model selection problem.

6.1 Parameter estimation

Let X be random variable with observed data $\mathbf{x}$ and suppose we want to use this data to estimate a parameter θ. The maximum likelihood approach to parameter estimation obtains the value θ that maximizes the likelihood function. This can be expressed in terms of maximizing the probability density of the observed data as follows:

$$\hat{\theta}_{ML} = \arg\max_{\theta \in \Theta} p(\mathbf{x}|\, \theta). \tag{10}$$

An alternative approach to parameter estimation called *maximum a posteriori* or MAP defines a prior distribution $\pi(\theta)$ over the parameter θ and then maximizes the posterior probability of θ given the data:

$$\hat{\theta}_{MAP} = \arg\max_{\theta \in \Theta} p(\mathbf{x}|\, \theta)\pi(\theta). \tag{11}$$

Note that this approach corresponds to the MPE approach that was used in the

discrete case.

Following a similar strategy, a new approach to parameter estimation based on PCM can be defined as follows:

$$\hat{\theta}_{PCM} = \arg\max_{\theta \in \Theta} p(\mathbf{x}|\,\theta)^2 \pi(\theta). \tag{12}$$

Let us see how these approaches apply in the case of n Bernoulli trials such as the tossing of a coin with an unknown bias, $\theta \in [0, 1]$, so that the probability of m heads in n tosses is given by the binomial distribution:

$$p(m|\,\theta) = \binom{n}{m} \theta^m (1-\theta)^{n-m}. \tag{13}$$

The maximum likelihood estimation is simply the proportion of heads:

$$\hat{\theta}_{ML} = \frac{m}{n}. \tag{14}$$

MAP estimation requires a suitable prior distribution to be chosen. Adopting a beta distribution with parameters α and β so that:

$$\pi(\theta|\alpha, \beta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1} \tag{15}$$

results in the following MAP estimate:

$$\hat{\theta}_{MAP} = \frac{m + \alpha - 1}{n + \alpha + \beta - 2}. \tag{16}$$

The PCM estimate can be obtained by substituting (13) and (15) into (12), and then maximizing with respect to θ by taking the log, then the derivative with respect to θ, and then setting to zero. This results in the following estimate:

$$\hat{\theta}_{PCM} = \frac{2m + \alpha - 1}{2n + \alpha + \beta - 2}. \tag{17}$$

Before commenting on this result, let us briefly consider parameter estimation for a one dimensional Gaussian distribution with mean μ and variance σ^2 from observations $\mathbf{x} = (x_1, \ldots, x_n)$. The maximum likelihood estimation for the mean is:

$$\hat{\mu}_{ML} = \frac{1}{n} \sum_{j=1}^{n} x_j, \tag{18}$$

while the MAP estimate is:

$$\hat{\mu}_{MAP} = \frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2} \hat{\mu}_{ML} + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2} \mu_0, \tag{19}$$

where μ_0 and σ_0^2 are the mean and variance respectively of the prior distribution. It is easy to show that PCM gives the following estimate:

$$\hat{\mu}_{PCM} = \frac{2n\sigma_0^2}{2n\sigma_0^2 + \sigma^2} \hat{\mu}_{ML} + \frac{\sigma^2}{2n\sigma_0^2 + \sigma^2} \mu_0. \tag{20}$$

For both the binomial and Gaussian cases, the PCM result is obtained from the MAP result by a factor of two being applied to the data (terms involving m and n). This derives from the fact that maximizing $p(\mathbf{x}|\,\theta)^2 \pi(\theta)$ in equation (12) is equivalent to maximizing $p(\mathbf{x}, \mathbf{x}|\,\theta)\pi(\theta)$ in equation (11) which corresponds to the data points in $\mathbf{x}$ having occurred twice.

What do these results tell us? It can be shown that the PCM estimates lie between the ML and MAP estimates. That is, for the binomial case:

$$\min\{\hat{\theta}_{ML}, \hat{\theta}_{MAP}\} \leq \hat{\theta}_{PCM} \leq \max\{\hat{\theta}_{ML}, \hat{\theta}_{MAP}\} \tag{21}$$

and similarly for the Gaussian case:

$$\min\{\hat{\mu}_{ML}, \hat{\mu}_{MAP}\} \leq \hat{\mu}_{PCM} \leq \max\{\hat{\mu}_{ML}, \hat{\mu}_{MAP}\}. \tag{22}$$

This shows that an approach to IBE based on PCM can be applied not only to scientific inference in a general sense, but also to parameter estimation and that

in doing so it produces sensible results. Furthermore, these results suggest that this approach might have benefits over the ML and MAP estimates in some cases. PCM provides a way of taking prior probabilities into account without giving them as much weight as MAP. And just as this gave rise to better results in some cases where there was uncertainty in the priors in section 5, so it could result in better parameter estimates in some cases.

Having shown that PCM can give rise to reasonable results in the case of parameter estimation, let us now consider whether it might also be applied to the model selection problem.

6.2 Model selection

In the model selection problem, we can think of a model as a family of statistical hypotheses such as polynomials of a given order used to fit data in a regression problem. Suppose we have data $\mathbf{x}$ and a set of models $\mathcal{M}$. The goal in model selection is to select the model that scores best according to a specified criterion, where the criterion is intended to represent a trade-off between the complexity of a model and how well it fits the data. The Akaike Information Criterion (AIC) is a well-known approach that is based on the classical statistical procedure of estimation and is given by (Akaike [1973]):

$$AIC(M, \mathbf{x}) = -2 \log p(\mathbf{x} \,|\, \hat{\theta}_{ML}) + 2k, \tag{23}$$

where $M \in \mathcal{M}$, p is a probability density, $\hat{\theta}_{ML}$ is the maximum likelihood estimate, and k is the number of parameters to be estimated.

From a Bayesian perspective, the posterior probabilities of two models, M_i

and M_j can be compared as follows:

$$\frac{p(M_i|\,\mathbf{x})}{p(M_j|\,\mathbf{x})} = \frac{p(\mathbf{x}|M_i)}{p(\mathbf{x}|M_j)} \times \frac{p(M_i)}{p(M_j)}, \tag{24}$$

where the term $p(\mathbf{x}|M_i)/p(\mathbf{x}|M_j)$ is the Bayes factor (Kass and Raftery [1995]). Adopting a MAP approach, we wish to find the model M_{MAP} which maximizes $p(M|\,\mathbf{x})$:

$$M_{MAP} = \arg\max_{M \in \mathcal{M}} p(\mathbf{x}\,|M)p(M) \tag{25}$$

or equivalently we can minimize the negative log to get:

$$M_{MAP} = \arg\min_{M \in \mathcal{M}} \left[-\log p(\mathbf{x}|M) - \log p(M)\right], \tag{26}$$

where $-\log p(\mathbf{x}|M)$ relates to how well the model fits the data and $-\log p(M)$ is a penalty term, where smaller values of $p(M)$ result in greater penalties.

How might the PCM approach be used for model selection? Based on equation (8), we can compare two models using the following expression corresponding to (24):

$$\frac{\mathcal{E}_{PCM}(\mathbf{x}, M_i)}{\mathcal{E}_{PCM}(\mathbf{x}, M_j)} = \frac{p(\mathbf{x}|M_i)^2}{p(\mathbf{x}|M_j)^2} \times \frac{p(M_i)}{p(M_j)}. \tag{27}$$

Hence, we can select the model that maximizes $p(\mathbf{x}|M)^2 p(M)$:

$$M_{PCM} = \arg\max_{M \in \mathcal{M}} p(\mathbf{x}|M)^2 p(M) \tag{28}$$

or equivalently we can minimize the negative log:

$$M_{PCM} = \arg\min_{M \in \mathcal{M}} \left[-2\log p(\mathbf{x}|M) - \log p(M)\right], \tag{29}$$

and so by comparing (29) with (26) we see that the PCM approach gives more weight to the data and less to the penalty than does MAP. The challenge for Bayesian approaches such as MAP and hence for PCM as well is to determine

the factor $p(\mathbf{x}|M)$ and to identify appropriate priors. One approach is to use that adopted in the Bayesian Information Criterion (BIC) which employs the Laplace approximation to integrate over the parameter space and assumes that the number of data points, n, is large so that only terms that depend on n are taken into account (Schwarz [1978]). This means that the priors drop out. Using the PCM approach with this approximation yields the the same result as the standard BIC approach to model selection, which can be expressed as follows:

$$BIC(M, \mathbf{x}) = -2 \log p(\mathbf{x}|\, \hat{\theta}_{ML}) + k \log n. \tag{30}$$

These results are somewhat encouraging. Based on (29), we see that an approach to model selection based on PCM is similar to MAP, but gives more weight to the data. Clearly, in cases where a uniform prior distribution is adopted over models there will be no difference between the approaches. Similarly, given the assumptions underlying BIC, PCM gives rise to the same results. Hence, just as Bayesian approaches can be applied to model selection, these results suggest that IBE based on PCM can be similarly applied. Could PCM have advantages over other approaches such as BIC? One direction for future work in this area would be to consider other approximations where differences in the priors of the models would differentiate between the approaches. Related to this, another direction would be to investigate how the PCM approach might be related to other approaches such as the minimum message length (see Wallace and Dowe [1999]). Model selection has given rise to debate between Bayesians and non-Bayesians (see Forster and Sober [1994]; Dowe et al. [2007]), so it would be interesting to see whether the preliminary work here on an approach motivated by IBE might be extended to contribute to that debate.

7 Conclusion

An earlier paper set out to show that IBE tracks truth when the overlap coherence measure (OCM) was used to compare explanations (Glass [2012]). Interestingly, when IBE was formulated in this way it was extremely successful at finding the true hypothesis, almost as good as an approach based on maximizing posterior probability and better than maximum likelihood. Even more surprising was the discovery that in some cases where there was uncertainty in the prior probabilities IBE was *more successful* at finding the truth than maximizing posterior probability or maximizing likelihood. These results appeared to achieve more than is needed to defend IBE as a mode of reasoning.

However, Glymour ([2015]) identified a number of general problems for various measures that seek to use probability to quantify how well a hypothesis explains the evidence. Responses have been presented here to his objections concerning 'excellent but false explanations' and 'causal explanation', but some of his criticisms of the work on hypothesis selection described above have been accepted. First, it would be difficult to use OCM in practice since it requires determining the probability of the evidence which is often unavailable and, second, the advantages of this approach over maximum likelihood vanish for larger (but still relatively small) sample sizes.

To address these issues, a new measure (the product coherence measure, PCM) has been proposed which has several advantages compared to OCM and solves both problems. Hence, this new version of IBE, which uses PCM to compare explanations, is more successful at tracking the truth than the previous version based on OCM. Some preliminary work has also been presented to show how this approach might be applied to parameter estimation and model selection. Frequent

criticisms of IBE are that it is not clearly defined and that the connection between explanation and truth has not been established. However, if PCM provides an adequate measure for comparing explanations, then not only does the account of IBE provided here address both these criticisms, but IBE is shown to track the truth much more closely than might have been expected. Furthermore, since it is more accurate at finding the truth than standard approaches such as maximizing posterior probability or maximizing likelihood in cases involving uncertainty and since these advantages persist for larger sample sizes, IBE may well have scientific as well as philosophical merit.

Acknowledgements

The author would like to thank Prof. Clark Glymour, Dr. Jonah Schupbach, and reviewers for very helpful comments and suggestions.

School of Computing
Ulster University
Newtownabbey, UK
dh.glass@ulster.ac.uk

References

Akaike, H. [1973]: 'Information theory and an extension of the maximum likelihood principle', in B. N. Petrov and F. Cśaki (eds.), *2nd International Symposium on Information Theory*, Budapest: Akadémiai Kiadó, pp. 267–281.

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

Bovens, L. and Olsson, E. [2000]: 'Coherence, reliability and Bayesian networks', *Mind*, **109**, pp. 685–719.

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

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

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

Dowe, D. L., Gardner, S. and Oppy, G. [2007]: 'Bayes not bust! why simplicity is no problem for Bayesians', *The British Journal for the Philosophy of Science*, **58**, pp. 709–754.

Eva, B. and Stern, R. [2018]: 'Causal explanatory power', *The British Journal for the Philosophy of Science*, axy012, ¡https://doi.org/10.1093/bjps/axy012¿.

Fitelson, B. [2003]: 'A probabilistic theory of coherence', *Analysis*, **63**, pp. 194–199.

Forster, M. and Sober, E. [1994]: 'How to tell when simpler, more unified, or less ad hoc theories will provide more accurate predictions', *The British Journal for the Philosophy of Science*, **45**, pp. 1–35.

Glass, D. H. [2005]: ‘Problems with priors in probabilistic measures of coherence’, *Erkenntnis*, **63**, pp. 375–385.

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

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

Glass, D. H. and McCartney, M. [2014]: ‘Explanatory inference under uncertainty’, in E. Corchado, J. A. Lozano, H. Quintián and H. Yin (eds.), *Intelligent Data Engineering and Automated Learning – IDEAL 2014. Lecture Notes in Computer Science*, volume 8669, pp. 215–222.

Glymour, C. [2015]: ‘Probability and the explanatory virtues’, *The British Journal for the Philosophy of Science*, **66**, pp. 591–604.

Kass, R. E. and Raftery, A. E. [1995]: ‘Bayes factors’, *Journal of the American Statistical Association*, **90**, pp. 773–795.

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

Myrvold, W. C. [2003]: ‘A Bayesian account of the virtue of unification’, *Philosophy of Science*, **70(2)**, pp. 399–423.

Olsson, E. J. [2005]: *Against Coherence*, Oxford: Oxford University Press.

Psillos, S. [1999]: *Scientific Realism: how science tracks truth*, London: Routledge.

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

Schupbach, J. N. [2011b]: ‘New hope for Shogenji’s coherence measure’, *The British Journal for the Philosophy of Science*, **62**, pp. 125–142.

Schupbach, J. N. and Sprenger, J. [2011]: ‘The logic of explanatory power’, *Philosophy of Science*, **78**, pp. 105–127.

Schwarz, G. [1978]: ‘Estimating the dimension of a model’, *The Annals of Statistics*, **6**, pp. 461–464.

Shogenji, T. [1999]: ‘Is coherence truth-conducive?’, *Analysis*, **59**, pp. 338–345.

Teng, C. M., Ramsey, J. D. and Glymour, C. [unpublished]: ‘Accuracy of hypothesis selection and inference to the best explanation’.

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

Wallace, C. S. and Dowe, D. L. [1999]: ‘Minimum message length and kolmogorov complexity’, *The Computer Journal*, **42**, pp. 270–283.

Wheeler, G. [2009]: ‘Focused correlation and confirmation’, *The British Journal for the Philosophy of Science*, **60**, pp. 79–100.