A Comparison of Explanatory Measures in Abductive Inference

Jian-Dong Huang(✉), David H. Glass(✉), and Mark McCartney(✉)

School of Computing, Ulster University,
Newtownabbey, Co. Antrim, Northern Ireland BT37 0QB, UK
{jd.huang, dh.glass, m.mccartney}@ulster.ac.uk

Abstract. Computer simulations have been carried out to investigate the performance of two measures for abductive inference, Maximum Likelihood (ML), and Product Coherence Measure (PCM), by comparing them with a third approach, Most Probable Explanation (MPE). These have been realized through experiments that compare outcomes from a specified model (the correct model) with those from incorrect models which assume that the hypotheses are mutually exclusive or independent. The results show that PCM tracks the results of MPE more closely than ML when the degree of competition is greater than 0 and hence is able to infer explanations that are more likely to be true under such a condition. Experiments on the robustness of the measures with respect to incorrect model assumptions show that ML is more robust in general, but that MPE and PCM are more robust when the degree of competition is positive. The results also show that in general it is more reasonable to assume the hypotheses in question are independent than to assume they are mutually exclusive.

Keywords: Inference to the Best Explanation (IBE) · Explanatory reasoning · Hypotheses competition · Abduction

1 Introduction

In modern literature, abduction refers to the study of explanatory reasoning in justifying hypotheses, or Inference to the Best Explanation (IBE) that considers a number of plausible candidate hypotheses in a given evidential context and then compares these hypotheses in order to make an inference to the one that best explains the relevant evidence [1–11].

In conventional studies involving IBE, significant attention has been paid to dealing with the hypotheses being mutually exclusive [12–15], and the measure for identifying the most plausible hypothesis was typically chosen as the maximized posterior probability [16, 17] termed Most Probable Explanation (MPE). However, modern studies have highlighted situations where hypotheses can be in competition even though they are not mutually exclusive and can compete to varying degrees [18, 19]. Meanwhile, alternative measures to MPE have been considered and applied [12, 19–29], such as Maximum Likelihood (ML) and Product Coherence Measure (PCM) [11]. The reality that hypotheses often have various degrees of competition gives rise to the necessity of examining the characteristics of abductive inference in such a context, along with the

M.-J. Lesot et al. (Eds.): IPMU 2020, CCIS 1239, pp. 304–317, 2020.
https://doi.org/10.1007/978-3-030-50153-2_23

characteristics of the functioning and performance of the explanatory measures used to make inferences in these contexts. This then motivates a study to compare the performance of several measures when conducting abductive inference under the assumption that they may be competing to some extent, with an objective of identifying the most suitable measure(s) as the criteria/criterion for the inference to the best explanation, thus benefiting abductive inference research. Therefore, in this study, comparison of the performance and functionality among the measures MPE, ML, and PCM in identifying the best explanation has been carried out. We consider a number of different probability model settings, as an extension of the study of abductive inference under various degrees of competition between candidate hypotheses [19].

2 Competing Hypotheses and Degree of Competition

We start by giving an example to illustrate the competition concept. Suppose a detective has two main suspects in a murder inquiry, Smith and Jones. The detective tries to determine which hypothesis best explains all the relevant evidence by treating the suspects as two competing hypotheses and reasoning abductively. The hypotheses can be represented as:

H_S: Smith committed the murder
H_J: Jones committed the murder

In general, the hypotheses need not be assumed to be mutually exclusive, since both Smith and Jones could have colluded in committing the murder and hence it would be improper to assume that $P(H_S\&H_J) = 0$. Clearly, if the two hypotheses are known to be mutually exclusive (if Smith and Jones could not have colluded), then they are competing hypotheses. In reality, it might be difficult to establish mutual exclusion, but in many cases, it would still be reasonable to treat them as competing hypotheses. Perhaps, for example, in light of the evidence it is very unlikely but not impossible that Smith and Jones colluded.

Glass [19] proposed a definition for competing hypotheses: Let each of H_1 and H_2 be hypotheses and E evidence under consideration and suppose that $P(H_1\&E)$ and $P(H_2\&E)$ are greater than zero. Hypotheses H_1 and H_2 are said to be competing hypotheses with respect to evidence E if and only if $P(H_1|H_2\&E) < P(H_1|E)$. Because the competition is a symmetric concept, the formula can also be expressed as $P(H_2|H_1\&E) < P(H_2|E)$.

Schupbach and Glass [18] recently defined a measure of the degree of competition between two hypotheses, H_1 and H_2, with respect to evidence E, as the average degree to which H_1 and H_2 disconfirm each other given E:

$$\mathrm{Comp}(H_1, H_2|E) = \frac{1}{2} \times [C_l(H_1, \neg H_2|E) + C_l(H_2, \neg H_1|E)] \quad (1)$$

where C_l is the likelihood ratio measure of confirmation conditioned on E, that is,

$$C_l(H_1,\ H_2 \mid E) = \log \frac{P\ (H_1 \mid H_2 \& E)}{P\ (H_1 \mid \neg H_2 \& E)} \tag{2}$$

By the definition, Comp has increasingly positive values to the extent that H_1 and H_2 disconfirm one another given E, and increasingly negative values to the extent that H_1 and H_2 confirm one another given E, and zero when H_1 and H_2 are probabilistically independent given E. Note that if H_1 and H_2 are assumed to be mutually exclusive, then $P(H_1 \mid H_2 \& E) = 0$ and hence this would be a case with the highest possible degree of competition. H_1 and H_2 can be said to compete with respect to E if Comp > 0. It can then be shown that H_1 and H_2 compete with respect to E if the condition in the definition is met. Since the measure of competition lies in the range $[-\infty, \infty]$, their alternative measure, which lies in the range [− 1, 1], has been used for convenience in this study. It is given by [18]

$$\text{dcomp} = \frac{1}{2} \times [C_k(H_1, \neg H_2 \mid E) + C_k(H_2, \neg H_1 \mid E)] \tag{3}$$

where C_k is the confirmation measure proposed by Kemeny and Oppenheim [30] when conditioned on E,

$$C_k(H_1,\ H_2 \mid E) = \frac{P\ (H_1 \mid H_2 \& E) - P(H_1 \mid \neg H_2 \& E)}{P\ (H_1 \mid H_2 \& E) + P(H_1 \mid \neg H_2 \& E)} \tag{4}$$

3 Probability Model and Experiment Design

Computer simulations were carried out to investigate the performance and functionality of the different measures when incorrectly assuming hypotheses to be mutually exclusive or independent, on making inferences. Each of the experiments concerned generating a probability model involving evidence E and hypotheses H_1, H_2 and a catchall hypothesis, $H_c = \neg H_1 \& \neg H_2$, with a specified degree of competition between H_1 and H_2 given E [19]. This model was stipulated to be the correct model, $Prob_0$, and three different incorrect probability models, modified from that of $Prob_0$, in which H_1 and H_2 were treated as mutually exclusive for two experiments (named MEx1 and MEx2 respectively), and treated as independent for a third experiment (named IND), were used for the inference.

These are intended to represent simplifying assumptions that might be made in practice when the true probability model is unknown. The goal is then to evaluate several versions of abductive inference under these assumptions. Each of the experiments were repeated a large number of times ($N = 10^6$) to sample the distribution over the variables and obtain a meaningful average, and a Degree of Agreement (*DA*) is defined as follows:

Let N_identified be the number of times the hypothesis identified by the incorrect model agrees with the correct model, N_total be the total number of observations, under a given degree of competition dcomp, then

$$DA(\%) = 100 \times \frac{\text{N\_identified}}{\text{N\_total}} \tag{5}$$

Note that the *DA* is a parameter closely linked to the degree of competition, or the *DA* is a function of dcomp, because N_identified is a function of dcomp.

We also use *DA* Index (*DAI*) to represent the average Degree of Agreement over the interval [−1, 1] containing successively (with a certain step) n points of dcomp:

$$DAI = \frac{\sum_{k=1}^{n} (DA)_k}{n} \tag{6}$$

The simulations are an extension of a study in which the measure MPE was used as the standard for hypothesis identification [19] since here ML and PCM are also used. Details of how the correct probability model and the incorrect models were constructed can be found in [19]. Design of the extended experiments can be sketched as follows [11, 19]:

Firstly, for a specified value of the degree of competition, dcomp (Eq. (3)), a probability model was defined and stipulated as the correct model Prob_0, involving hypotheses H_1, H_2, a catchall H_c, and evidence E, where H_1 and H_2 are not assumed to be mutually exclusive; the initial parameters in the model are randomly generated from a uniform distribution.

For MEx1, a mutually exclusive probability model Prob_1 is obtained from the original model, Prob_0, by replacing H_1 with $H_1\&\neg H_2$ and H_2 with $H_2\&\neg H_1$, setting:

$$\text{Prob}_1(H_1) = \text{Prob}_0(H_1\&\neg H_2) \tag{7}$$

$$\text{Prob}_1(H_2) = \text{Prob}_0(H_2\&\neg H_1) \tag{8}$$

$$\text{Prob}_1(H_1\&H_2) = 0 \tag{9}$$

$$\text{Prob}_1(H_c) = 1 - \text{Prob}_1(H_1) - \text{Prob}_1(H_2) \tag{10}$$

$$\text{Prob}_1(E \mid H_1) = \text{Prob}_0(E \mid H_1\&\neg H_2) \tag{11}$$

$$\text{Prob}_1(E \mid H_2) = \text{Prob}_0(E \mid H_2\&\neg H_1) \tag{12}$$

$$\text{Prob}_1(E \mid H_c) = \text{Prob}_0(E \mid H_c) \tag{13}$$

In the second experiment MEx2, another mutually exclusive probability model, Prob_2, is obtained from the original model Prob_0,

$$\text{Prob}_2(\text{H}_1) = \text{Prob}_0(\text{H}_1) \times \frac{\text{Prob}_0(\text{H}_1 \vee \text{H}_2)}{\text{Prob}_0(\text{H}_1) + \text{Prob}_0(\text{H}_2)} \tag{14}$$

$$\text{Prob}_2(\text{H}_2) = \text{Prob}_0(\text{H}_2) \times \frac{\text{Prob}_0(\text{H}_1 \vee \text{H}_2)}{\text{Prob}_0(\text{H}_1) + \text{Prob}_0(\text{H}_2)} \tag{15}$$

The probabilities for the likelihood terms are set in the same way as for Prob_1, and the probability for the catchall hypothesis, H_c, is similarly set to

$$\text{Prob}_2(\text{H}_c) = 1 - \text{Prob}_2(\text{H}_1) - \text{Prob}_2(\text{H}_2) \tag{16}$$

since H_1 and H_2 are assumed to be mutually exclusive [19]; and we have:

$$\text{Prob}_2(\text{H}_1 \& \text{H}_2) = 0 \tag{17}$$

$$\text{Prob}_2(\text{E} \mid \text{H}_1) = \text{Prob}_0(\text{E} \mid \text{H}_1 \& \neg \text{H}_2) \tag{18}$$

$$\text{Prob}_2(\text{E} \mid \text{H}_2) = \text{Prob}_0(\text{E} \mid \text{H}_2 \& \neg \text{H}_1) \tag{19}$$

In contrast to the models Prob_1 and Prob_2, the third experiment IND treats H_1 and H_2 as independent:

$$\text{Prob}_3(\text{H}_1) = \text{Prob}_0(\text{H}_1) \times \frac{\text{Prob}_0(\text{H}_1 \vee \text{H}_2)}{\text{Prob}_0(\text{H}_1) + \text{Prob}_0(\text{H}_2)} \tag{20}$$

$$\text{Prob}_3(\text{H}_2) = \text{Prob}_0(\text{H}_2) \times \frac{\text{Prob}_0(\text{H}_1 \vee \text{H}_2)}{\text{Prob}_0(\text{H}_1) + \text{Prob}_0(\text{H}_2)} \tag{21}$$

$$\text{Prob}_3(\text{H}_1 \& \text{H}_2) = \text{Prob}_3(\text{H}_1) \times \text{Prob}_3(\text{H}_2) \tag{22}$$

$$\text{Prob}_3(\text{H}_c) = 1 - \text{Prob}_3(\text{H}_1 \vee \text{H}_2) \tag{23}$$

$$\begin{aligned} &\text{Prob}_3(\text{E}|\text{H}_1) = \\ &\text{Prob}_0(\text{E}|\text{H}_1 \& \text{H}_2) \times \text{Prob}_3(\text{H}_2) + \text{Prob}_0(\text{E}|\text{H}_1 \& \neg \text{H}_2) \times \text{Prob}_3(\neg \text{H}_2) \end{aligned} \tag{24}$$

$$\begin{aligned} &\text{Prob}_3(\text{E}|\text{H}_2) = \\ &\text{Prob}_0(\text{E}|\text{H}_2 \& H_1) \times \text{Prob}_3(\text{H}_1) + \text{Prob}_0(\text{E}|\text{H}_2 \& \neg \text{H}_1) \times \text{Prob}_3(\neg \text{H}_1) \end{aligned} \tag{25}$$

This is because $Prob_3$ provides a compromise between incorrectly treating hypotheses as mutually exclusive and fully taking into account the dependence between them; we need to consider that in some cases hypotheses are modelled as being independent as well.

Secondly, abductive inference was carried out under a given degree of competition dcomp, for the correct models to find the hypothesis which maximizes the selected measure (MPE, ML, PCM) for evidence E. If an inference made by the incorrect model (mutually exclusive or independent) in identifying the hypothesis agreed with the inference made using the correct model, this inference is then counted as a success.

Thirdly, the above process was repeated $N = 10^6$ times and the number of total successful inferences S was obtained. The accuracy, or the degree of agreement between the inference of the incorrect model (mutually exclusive or independent) and the correct model, was defined as S/N (percentage success). The accuracy values reflect how often an incorrect model identifies the same hypothesis as the correct model.

Finally, the process was repeated for a range of values of the degree of competition between −0.9 and 0.9, with a step of 0.1.

There are a number of different measures proposed in the literature, to quantify how well a hypothesis H explains evidence E. For example, the Measure of Explanatory Power proposed by Schupbach and Sprenger [22]; the measure proposed by Crupi and Tentori [24]; the measure by Good [25] and McGrew [26]; the Likelihood Ratio measure [19]; the Overlap Coherence Measure used to rank explanations by Glass [27–29]; and the Product Coherence Measure by Glass [11, 12]. In this study, the following measures have been used for the inference in the computer simulation:

(1) MPE: Most Probable Explanation; selects the hypothesis with the maximum posterior probability, of the hypotheses in light of the evidence,

$$\mathrm{MPE} = \underset{H_i, i\in\{1,2,C\}}{\operatorname{argmax}} \; P(H_i|E) \tag{26}$$

(2) ML: selects the hypothesis with the Maximum Likelihood,

$$\mathrm{ML} = \underset{H_i, i\in\{1,2,C\}}{\operatorname{argmax}} \; P(E|H_i) \tag{27}$$

(3) PCM: selects the hypothesis with the maximum value of the Product Coherence Measure [11]:

$$\mathrm{PCM} = \underset{H_i, i\in\{1,2,C\}}{\operatorname{argmax}} \; [P(H_i|E) \times P\,(E\,|H_i)] \tag{28}$$

Arguably, MPE is not a good measure of explanation [20–27]. In the example of the murder suspects, the probability that both Smith and Jones are guilty, $P(H_S\&H_J|E)$, will obviously always be less than or equal to that of the individual hypotheses, $P(H_S|E)$ or $P(H_J|E)$. However, if MPE is used as a measure of explanation, this means that the joint

explanation that Smith and Jones committed the murder can never provide a better explanation than the individual explanations that Smith (or Jones) committed the murder. More generally, this means that it only makes sense to use MPE for a fixed number of explanatory variables, but arguably in various contexts, such as explanation in Bayesian networks, it is desirable to compare different numbers of explanatory variables in order to obtain explanations that are neither too simple nor too complex [21].

But MPE is still a useful measure to include for comparison since we are interested in whether inferences made using explanatory measures such as ML and PCM have a high probability of being correct.

As a further extension of [19], the performance of these three measures was examined and compared in the computer simulation of abductive inference for identifying the most probably correct explanation. The computer simulations were carried out with the procedures described earlier. In reality we typically do not know the true model, so we are evaluating how well abductive inference works with different incorrect assumptions, mutually exclusive or independent. Bearing this in mind, within each of the experiments, two groups of comparisons were made:

Group 1: Here the assumption is that explanatory approaches (ML and PCM) should be compared against MPE as the standard to see how good ML and PCM are at inferring hypotheses that are probably true. Thus, with MPE as a standard in the hypothesis identification in the correct model, this group is to find out the degree of agreement of hypothesis identification made from the correct model against that from an incorrect model, with MPE, ML, and PCM respectively as the criterion in the hypothesis identification in the incorrect model. These experiments are repeated for each of the incorrect models (MEx1, MEx2, IND). The inferences with the three measures have been abbreviated as:

- MPE_F versus MPE_T: using MPE criterion in the incorrect (False or _F) model against using MPE criterion in the correct (True or _T) model to infer a hypothesis;
- ML_F versus MPE_T: using Maximum Likelihood (ML) criterion in the incorrect (_F) model against using MPE criterion in the correct (_T) model to infer a hypothesis;
- PCM_F versus MPE_T: using Product Coherence Measure (PCM) criterion in the incorrect (_F) model against using MPE criterion in the correct (_T) model to infer a hypothesis;

Group 2: In abductive inference we do not necessarily need to consider MPE as the only standard. Therefore, in this group, each of the three measures was applied in the inference as the criterion for both incorrect model and the correct model, i.e., taking each of the three measures as the standard, to find out the degree of agreement of hypothesis identification made from the correct model against that from an incorrect model. These experiments can be seen as evaluating the robustness of each of the measures with respect to the incorrect model assumptions (mutually exclusive or independent). Again, these experiments are repeated for each of the incorrect models (MEx1, MEx2, IND). The inferences have been abbreviated as:

- MPE_F versus MPE_T: this is the same as in the Group1;

- ML_F versus ML_T: using Maximum Likelihood (ML) criterion in the incorrect (_F) model against using ML criterion in the correct (_T) to infer a hypothesis;
- PCM_F versus PCM_T: using Product Coherence Measure (PCM) criterion in the incorrect (_F) model against using PCM criterion in the correct (_T) model to infer a hypothesis;

The output of the two groups are expressed by Degree of Agreement, *DA*, and *DAI*, formulated in (5) and (6). It should be noted that this metric should be interpreted in one of two ways, depending on what comparison is being made. For Group 1, the *DA* reflects the Accuracy of the corresponding measure, i.e., given that we presume that MPE is a standard for identifying the true hypothesis, the degree of agreement with the identification by MPE is then viewed as the accuracy of the relevant measure, and the higher the *DA* is, the better the Accuracy the measure possesses. In Group 2, the same measure is used in the incorrect model and the correct model, and in this case the output of the three experiments will show how close the three series of output will be, i.e., the degree of consistency of the same measure under different experiments. In this case we say that the higher the degree of agreement the more *robust* the measure is. Therefore, under the circumstance of Group 2, we say that the Degree of Agreement reflects the Robustness of the measure. Accuracy and Robustness are then used in this work to represent the relevant properties of the explanatory measures.

4 Results

Graphs were plotted to illustrate the results of MEx1, MEx2 and IND, for comparison of the performance (Accuracy and Robustness) of the three explanatory measures, MPE (Most Probable Explanation), ML (Maximum Likelihood), and PCM (Product Coherence Measure).

Figure 1-1 shows that when the MPE is used as the standard, ML and the PCM have lower degree of agreement with the identification using MPE, but PCM is much closer to MPE than ML as found in [12]. When dcomp < 0, the curves drop to below 50%, suggesting that for negative degree of competition all three measures result in poor agreement with the output of using the standard MPE and PCM performs slightly better than MPE.

Figure 1-2 and 1-3 exhibit a similar trend as in Fig. 1-1 when dcomp > 0; but in the range of dcomp < 0, the curves are better ordered from high to low without crossing. As expected, the MPE curve is higher than those of ML and PCM, and noticeably the MPE and PCM curves are all above 50% in the whole range [−0.9, 0.9]. The degree of agreement for the measure ML appears lower, with the value less than 50% in the majority of the interval for all three experiments.

In Fig. 1-3, all the curves for MPE and PCM are above 60% when dcomp > 0, showing that the PCM performs very well compared to ML in identifying the most probably correct explanation, with MPE as the standard. For dcomp < 0, the output of MPE and PCM still have their accuracy greater than 50%. However, the ML curve

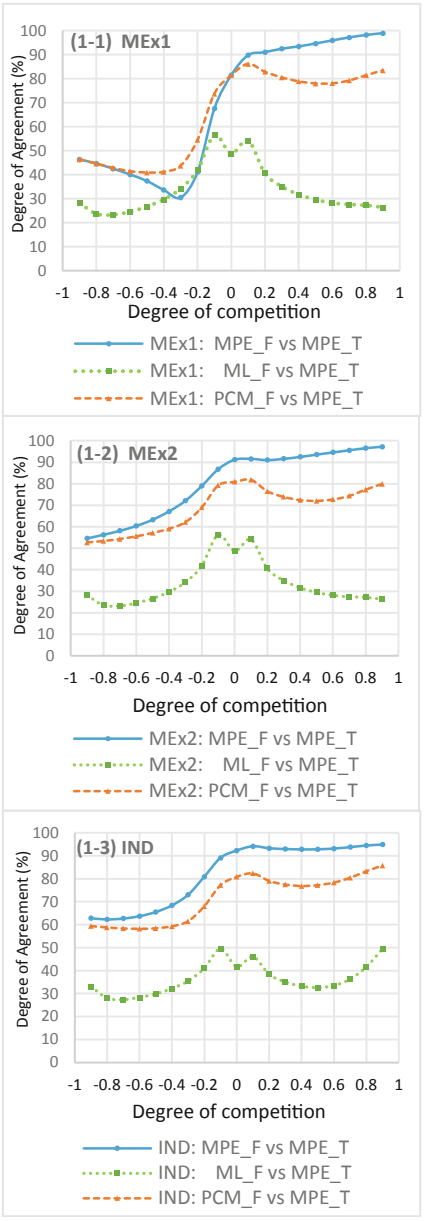

Fig. 1. Degree of agreement (Accuracy) between the output of the incorrect model (in MEx1, MEx2, and IND) using each of the measures and the correct model ($Prob_0$) using MPE to infer a hypothesis.

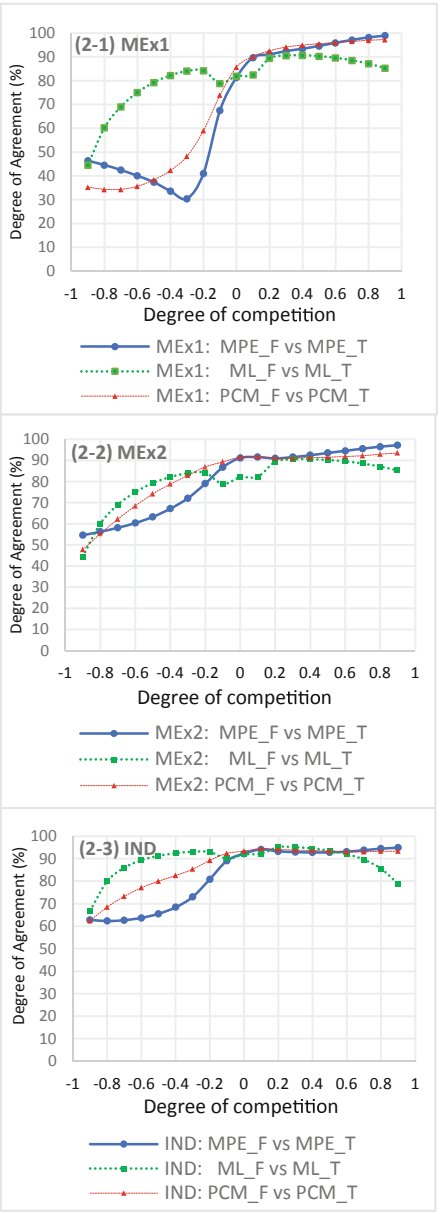

Fig. 2. Degree of agreement (Robustness) between the output of the incorrect model (in MEx1, MEx2, and IND) using each of the measures and the correct model ($Prob_0$) using the corresponding measure to infer a hypothesis.

output drops to below 30% when dcomp is in [−0.8,−0.6]. In all the three curves, PCM performs much better than ML.

In Fig. 1-1 through 1-3, it appears that the curves of the MPE measure are higher than PCM's for dcomp > 0 but the PCM curve is higher than the MPE curve for MEx1 from −0.6 to 0. In the majority of the range [−0.9, 0.9], MPE curves are higher. For MPE itself, it shows good degrees of agreement with the correct model in identifying the hypotheses when dcomp > 0; whilst in the case of dcomp < 0, it results in poor agreement with the output of the correct model.

In general, the experiment results indicate that, with the MPE measure and correct model used as the standard in hypothesis identification (a) for dcomp > 0 (or dcomp > 0.1 for MEx1), the MPE curves are all above 90%; for dcomp < 0, the curves are above 50% for MEx2 and IND. This is the same as one of the results in [19], implying that to presume the hypotheses to be mutually exclusive or independent appears reasonable especially for the situation in which the hypotheses compete to some degree (dcomp > 0), in the context of the experiments; (b) the measure ML has low degree of agreement with MPE in hypotheses identification; and (c) PCM results in closer degrees of agreement with the output of the MPE measure.

However, the above features do not necessarily mean that the ML and PCM are 'worse' measures than the MPE measure. Although MPE is often referred to in the artificial intelligence literature as the most probable explanation, arguably, this is an inadequate definition of 'best explanation' [11, 12, 19] but it nevertheless provides a standard against which to compare the various explanatory measures to determine how good they are at identifying hypotheses that are probably true.

The curves merely reflect the degree of agreement of the identifications made by ML or PCM with MPE. Therefore, a further comparison as illustrated in Fig. 2, examines the performance and Robustness when using the measures of ML and PCM with the correct model as the standard for the same measures with the incorrect models (experiment Group 2). This shall reveal more significant information on the performance of the measures.

It can be seen in Figs. 2-1 to 2-3, that in the range of dcomp > 0, MPE and PCM show a better degree of agreement with the identification of the correct model in the hypothesis identification (greater than 90%), whilst when dcomp < 0, ML performs better than the other two, with the degree of agreement largely above 50%, and for IND it goes up to 90% in (−0.6,−0.2). These features reflect that MPE does not always perform better than the other measures. PCM has a high degree of agreement similar to the MPE when dcomp > 0 (the difference is less than 5%) but ML performs better than MPE and PCM in the majority of the range of dcomp < 0.

Among the three figures of Fig. 2-1 through 2-3, ML has its highest curve in Fig. 2-3 (for the experiment IND), and PCM has its highest one in Fig. 2-3 as well. The curves of PCM are only slightly lower than MPE in all three figures when dcomp > 0, i.e. the PCM curve and the MPE curve are very close when dcomp > 0.

Moreover, curves for PCM are above their MPE counterparts in the majority of the range of dcomp < 0. For dcomp < 0, PCM perform better than MPE whilst when dcomp > 0 the two measures perform similarly.

The PCM curve for IND in Fig. 2-3 is above 60% in the whole range −0.9 to 0.9. It is obvious that, with PCM as the standard, presuming the hypotheses to be mutually exclusive and independent are both reasonable when dcomp > 0; and presuming them as being independent appears more reasonable when dcomp < 0, under the condition of the experiments.

Further, for a quantitative understanding of the properties of the measures, Fig. 3 and 4 give the average values of the Degree of Agreement over the interval [−0.9, 0],

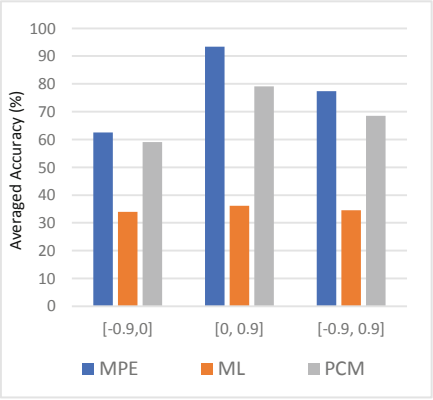

Fig. 3. Averaged *DAI* for the three experiments, showing a comparison of the performance of the measures ML and PCM against the presumed standard MPE for inferring a hypothesis in the experiment group 1, for the situations of dcomp $\leq$ 0, dcomp $\geq$ 0 and the whole interval [−0.9, 0.9].

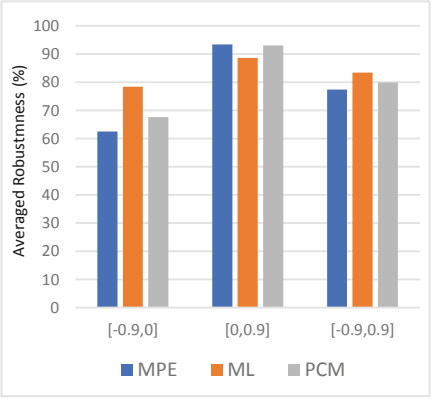

Fig. 4. Averaged Robustness for the output of the three experiments, illustrating a comparison of consistency of the measures for inferring a hypothesis under the incorrect model and correct model in the experiment group 2, for the situation of dcomp $\geq$ 0, dcomp $\leq$ 0, and for the whole range of [−0.9, 0.9].

[0, 0.9] and the whole range of [−0.9, 0.9], with Fig. 3 reflecting the Accuracy of the measures and Fig. 4 the Robustness of the measures. It can be seen that PCM has much higher Accuracy than ML but ML has slightly higher Robustness.

5 Conclusions

Computer simulations have been carried out to investigate the performance, in terms of Accuracy and Robustness, of two measures for abductive inference, Maximum Likelihood (ML), and Product Coherence Measure (PCM). This has been done by

comparing them with a third approach, Most Probable Explanation (MPE), for the identification of the best explanation. The results show that:

1. It appears appropriate to represent the functioning characteristics of the measures separately according to the sign of the degree of competition (dcomp) of the hypotheses, which can be calculated in practice using the Eqs. (1) through (4).
2. In terms of Accuracy, the results show that PCM tracks the results of MPE much more closely than ML especially when the degree of competition is positive, hence it is able to infer explanations that are much more likely to be true under such a condition.
3. Experiments on the Robustness of the measures with respect to incorrect model assumptions show that ML is in general more robust, although it is only slightly more robust than PCM. It performs better than MPE and PCM when the hypotheses are not competing (dcomp < 0) and in general. MPE and PCM are more robust and similar to each other when the degree of competition is positive; in general, PCM is more robust than MPE.
4. Presuming the hypotheses in question to be mutually exclusive appears reasonable when the hypotheses are competing (dcomp > 0) but could result in a low degree of agreement (accuracy) when they are not (dcomp < 0).
5. The experimental results also show that it is more reasonable to assume that the hypotheses are independent than to assume that they are mutually exclusive, both in the case of competing hypotheses and non-competing hypotheses.

Overall, the results show that PCM performs much better in terms of accuracy and only slightly worse in terms of robustness than ML. Hence, PCM seems preferable to ML as a measure for abductive inference. One limitation of the current work is that MPE has been used as a standard for determining accuracy. Future work will include simulations that designate hypotheses as true or false and then evaluate all three measures (MPE, PCM and ML) on an equal footing. Also, in the current work the different measures are used to infer the single best hypothesis, but since the hypotheses are not assumed to be mutually exclusive more than one hypothesis could be true. There is scope for comparing single hypotheses such as H_1 or H_2 with conjunctive hypotheses involving two or more hypotheses such as $H_1\&H_2$. Clearly, such a conjunction cannot be more probable than one of its conjuncts, so PCM and ML might be expected to have benefits over MPE in such contexts.

Acknowledgments. This publication was supported by a grant from the John Templeton Foundation (Grant ID 61115). The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of the John Templeton Foundation. The authors would like to thank anonymous reviewers for helpful comments and suggestions.

References

1. Douven, I.: Abduction. In: Zalta, E.N. (ed.) The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University, winter 2016 edition (2016)
2. Lipton, P.: Inference to the best explanation, 2nd edn. Taylor and Francis, London (2004)

3. Josephson, J.R., Susan, G.J.: Abductive Inference: Computation. Philosophy, Technology. Cambridge University Press, Cambridge (1996)
4. Harman, G.H.: The inference to the best explanation. Philos. Rev. **74**(1), 88–95 (1965)
5. Thagard, P.R.: The best explanation: Criteria for theory choice. J. Philos. **75**(2), 76–92 (1978)
6. Ben-Menahem, Y.: The inference to the best explanation. Erkenntnis **33**(3), 319–344 (1990)
7. Douven, I.: Inference to the best explanation made coherent. Philos. Sci. **66**, S424–S435 (1999)
8. Niiniluoto, I.: Defending abduction. Philos. Sci. **66**, S436–S451 (1999)
9. Gabbay, D., Woods, J.: The reach of abduction: Insight and Trial. 289–294 (2005)
10. Douven, I.: Inference to the best explanation, Dutch books, and inaccuracy minimization. Philos. Q. **63**(252), 428–444 (2013)
11. Glass, D.H.: Coherence, explanation, and hypothesis selection. Br. J. Philos. Sci. axy063 (2018). https://doi.org/10.1093/bjps/axy063
12. Glass, D.H.: An evaluation of probabilistic approaches to inference to the best explanation. Int. J. Approximate Reasoning **103**, 184–194 (2018)
13. Fenton, N., Neil, M., Lagnado, D., Marsh, W., Yet, B., Constantinou, A.: How to model mutually exclusive events based on independent causal pathways in Bayesian network models. Knowl.-Based Syst. **113**, 39–50 (2016)
14. Lam, F.C., Yeap, W.K.: Bayesian updating: on the interpretation of exhaustive and mutually exclusive assumptions. Artif. Intell. **53**(2–3), 245–254 (1992)
15. Norman Fenton, M.N., Lagnado, D.: Modelling mutually exclusive causes in Bayesian networks. Submitted to IEEE Transactions on Knowledge and Data Engineering (2011)
16. Pearl, J.: Probabilistic reasoning in intelligent systems: networks of plausible inference. Elsevier (2014)
17. Shimony, S.E.: Explanation, irrelevance and statistical independence. In: Proceedings of the Ninth National Conference on Artificial Intelligence-Volume **1**, 482–487 (1991)
18. Schupbach, J.N., Glass, D.H.: Hypothesis competition beyond mutual exclusivity. Philos. Sci. **84**(5), 810–824 (2017)
19. Glass, D.H.: Competing hypotheses and abductive inference. Ann. Math. Artif. Intel. 1–18 (2019). https://doi.org/10.1007/s10472-019-09630-0
20. Chajewska, U., Halpern, J.Y.: Defining explanation in probabilistic systems. In: Proceedings of the Thirteenth conference on Uncertainty in Artificial Intelligence, pp. 62–71. Morgan Kaufmann Publishers Inc. (1997)
21. Yuan, C., Lim, H., Lu, T.C.: Most relevant explanation in Bayesian networks. J. Artif. Intell. Res. **42**, 309–352 (2011)
22. Schupbach, J.N., Sprenger, J.: The logic of explanatory power. Philos. Sci. **78**(1), 105–127 (2011)
23. Schupbach, J.N.: Comparing probabilistic measures of explanatory power. Philos. Sci. **78**(5), 813–829 (2011)
24. Crupi, V., Tentori, K.: A second look at the logic of explanatory power (with two novel representation theorems). Philos. Sci. **79**(3), 365–385 (2012)
25. Good, I.J.: Weight of evidence, corroboration, explanatory power, information and the utility of experiments. J. Roy. Stat. Soc.: Ser. B (Methodol.) **22**(2), 319–331 (1960)
26. Mc Grew, T.: Confirmation, heuristics, and explanatory reasoning. Br. J. Philos. Sci. **54**(4), 553–567 (2003)
27. Glass, D.H.: Inference to the best explanation: does it track truth? Synthese **185**(3), 411–427 (2012). https://doi.org/10.1007/s11229-010-9829-9

28. Glass, D.H., McCartney, M.: Explanatory Inference under Uncertainty. In: International Conference on Intelligent Data Engineering and Automated Learning, pp. 215–222. Springer, Cham (2014)
29. Glass, D.H.: Coherence measures and inference to the best explanation. Synthese **157**, 275–296 (2007). https://doi.org/10.1007/s11229-006-9055-7
30. Kemeny, J.G., Oppenheim, P.: Degree of factual support. Philos. Sci. **19**(4), 307–324 (1952)