Contents lists available at ScienceDirect

Cognitive Psychology

journal homepage: www.elsevier.com/locate/cogpsych

Heuristics as Bayesian inference under extreme priors

Paula Parpart[a], Matt Jones[b], Bradley C. Love[a,c,*]

[a] University College London, United Kingdom
[b] University of Colorado Boulder, United States
[c] The Alan Turing Institute, United Kingdom

ARTICLE INFO

Keywords:
Heuristics
Bayesian inference
Decision making
Ridge regression

ABSTRACT

Simple heuristics are often regarded as tractable decision strategies because they ignore a great deal of information in the input data. One puzzle is why heuristics can outperform *full-information* models, such as linear regression, which make full use of the available information. These "less-is-more" effects, in which a relatively simpler model outperforms a more complex model, are prevalent throughout cognitive science, and are frequently argued to demonstrate an inherent advantage of simplifying computation or ignoring information. In contrast, we show at the computational level (where algorithmic restrictions are set aside) that it is never optimal to discard information. Through a formal Bayesian analysis, we prove that popular heuristics, such as tallying and take-the-best, are formally equivalent to Bayesian inference under the limit of infinitely strong priors. Varying the strength of the prior yields a continuum of Bayesian models with the heuristics at one end and ordinary regression at the other. Critically, intermediate models perform better across all our simulations, suggesting that down-weighting information with the appropriate prior is preferable to entirely ignoring it. Rather than because of their simplicity, our analyses suggest heuristics perform well because they implement strong priors that approximate the actual structure of the environment. We end by considering how new heuristics could be derived by infinitely strengthening the priors of other Bayesian models. These formal results have implications for work in psychology, machine learning and economics.

1. Introduction

Many real-world prediction problems involve binary classification based on available information, such as predicting whether Germany or England will win a soccer match based on the teams' statistics. A relatively simple decision procedure would use a rule to combine available information (i.e., *cues*), such as the teams' league position, the result of the last game between Germany and England, which team has scored more goals recently, and which team is home versus away. One such decision procedure, the *tallying heuristic*, simply checks which team is better on each cue and chooses the team that has more cues in its favor, ignoring any possible differences among cues in magnitude or predictive value (Czerlinski, Gigerenzer, & Goldstein, 1999; Dawes, 1979). In the scenario depicted in Fig. 1A this heuristic would choose England. Another algorithm, *take-the-best* (TTB), would base the decision on the best single cue that differentiates the two options. TTB works by ranking the cues according to their *cue validity* (i.e., predictive value), then sequentially proceeding from the most valid to least valid until a cue is found that favors one team over the other (Gigerenzer & Goldstein, 1996). Thus TTB terminates at the first discriminative cue, discarding all remaining cues.

In contrast to these heuristic algorithms, a *full-information model* such as linear regression would make use of all the cues, their

[*] Corresponding author.
E-mail address: b.love@ucl.ac.uk (B.C. Love).

https://doi.org/10.1016/j.cogpsych.2017.11.006
Accepted 30 November 2017

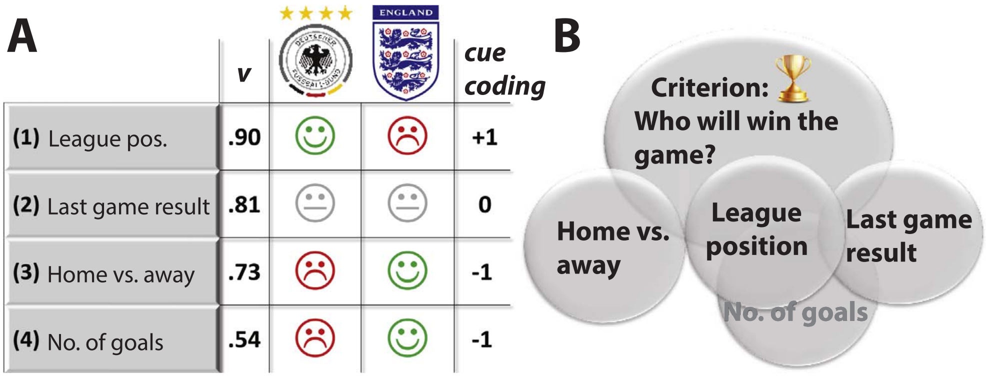

Fig. 1. Illustrative example of a binary prediction task. (A) Predicting whether Team Germany or England will win is based on four cues: league position, last game result, home vs. away match, and recent goal scoring. Cue validities (v) reflect the relative frequency with which each cue makes correct inferences across many team comparisons (formula in Appendix A). Smiley and frowning faces indicate which team is superior on each cue, whereas a grey face indicates the two teams are equal on that cue. For modeling, a cue is coded $+1$ when it favors the team on the left (Germany), -1 when it favors the team on the right (England), and 0 when the teams are equal along that cue. (B) Irrespective of cue validity, cues can co-vary (illustrated by overlap) with the criterion variable but also with each other. The heuristics considered here ignore this covariance among cues.

magnitudes, their predictive values, and observed covariation among them. For example, league position and number of goals scored are highly correlated, and this correlation influences the weights obtained from a regression model (Fig. 1B). Although such covariances naturally arise and can be meaningful, the cue validities used by the tallying and TTB heuristics completely ignore them (Martignon & Hoffrage, 1999). Instead, cue validities assess only the probability with which a single cue can identify the correct alternative, as the proportion of correct inferences made by that cue alone across a set of binary comparisons (formal definition in Appendix A). When two cues co-vary highly, they essentially provide the same information, but heuristics ignore this redundancy and treat the related cues as independent information sources. In the heuristic literature, the learner is usually assumed to learn cue validities from past experiences (i.e., the training data) (Gigerenzer & Goldstein, 1996; Gigerenzer & Todd, 1999).

Heuristics have a long history of study in cognitive science, where they are often viewed as more psychologically plausible than full-information models, because ignoring data makes the calculation easier and thus may be more compatible with inherent cognitive limitations (Bobadilla-Suarez & Love, 2018; Kahneman, 2003; Simon, 1990; Tversky & Kahneman, 1974). This view suggests that heuristics should underperform full-information models, with the loss in performance compensated by reduced computational cost. This prediction is challenged by observations of *less-is-more* effects, wherein heuristics sometimes outperform full-information models, such as linear regression, in real-world prediction tasks (Chater, Oaksford, Nakisa, & Redington, 2003; Czerlinski et al., 1999; Dawes, 1979; Gigerenzer & Goldstein, 1996; Goldstein & Gigerenzer, 2002; Hogarth & Karelaia, 2007; Katsikopoulos, Schooler, & Hertwig, 2010). These findings have been used to argue that ignoring information can actually improve performance, even in the absence of processing limitations. For example, Gigerenzer and Todd (1999) write, “There is a point where too much information and too much information processing can hurt” (p. 21). Likewise, Gigerenzer and Brighton (2009) conclude, “A less-is-more effect, however, means that minds would not gain anything from relying on complex strategies, even if direct costs and opportunity costs were zero” (p. 111).

Less-is-more arguments also arise in other domains of cognitive science, such as in claims that learning is more successful when processing capacity is (at least initially) restricted (Elman, 1993; Newport, 1990). Contrary to existing claims, we argue there is no inherent computational advantage to simplicity of information processing. Less-is-more effects can arise only when the space of models under consideration is limited to a particular family or architecture. At a computational level of analysis, where restrictions on algorithms are set aside (Marr, 1982), more information is always better.

We cast our argument in a Bayesian framework, wherein additional information (input data) is always helpful but must be correctly combined with appropriate prior knowledge. We first prove that the tallying and TTB heuristics are equivalent to Bayesian inference under the limit of an infinitely strong prior. This connection suggests that heuristics perform well because their relative inflexibility amounts to a strong inductive bias, one that is suitable for many real-world learning and decision problems.

We then use this connection to define a continuum of Bayesian models, determined by parametric variation in the strength of the prior. At one end of the continuum (infinitely diffuse prior), the Bayesian model is equivalent to a variant of linear regression, and at the other end (infinitely strong prior) it is equivalent to a heuristic. Although the Bayesian models mimic the heuristics perfectly in the limit, a crucial difference is that the Bayesian account regulates cue weights but never discards any information. The models are tested on classic datasets that have been used to demonstrate superiority of the heuristics over linear regression, and in all cases we find that best performance comes from intermediate models on the continuum, which do not entirely ignore cue weights or cue covariance but that nonetheless down-weight this information via the influence of their priors. These results suggest that the success of heuristics, and findings of less-is-more effects more broadly in cognitive science, are due not to a computational advantage of simplicity per se, but rather to the fact that simpler models can approximate strong priors that are well-suited to the true structure of the environment.