Contents lists available at ScienceDirect

Journal of Mathematical Psychology

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

Extending Bayesian induction

Suyog H. Chandramouli, Richard M. Shiffrin [*]

Indiana University, United States

HIGHLIGHTS

- Extension of Bayesian Model Selection based on data distributions, termed BMS*.
- Induction is upon data which are a sample from the 'true' data generating distribution.
- Traditional BMS is a special case based on a restricted set of data distributions.

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Model Selection
Bayes Factor
BMS*
Data distributions

ABSTRACT

This article comments on "Harold Jeffreys's Default Bayes Factor Hypothesis Tests: Explanation, Extension, and Application in Psychology" by Ly, Verhagen and Wagenmakers (in this issue). Their article represents an excellent summary of the seminal contributions of Harold Jeffreys to Bayesian induction. We comment on a method to extend Bayesian induction that places the emphasis on data rather than models. Models are always wrong, acting as approximations to the data from which they derive and thereby explaining some of the main factors operating in the experiment. Our simple extension places priors and posteriors on the possible distributions of data outcomes, one of which represents the true state of the world—the observed data is a sample from that unknown true state. The proposed system infers the probability that a given data distribution is the true one, based on the observed sample of data; the posterior probabilities that given model instances provide best approximations to the truth can then be obtained directly.

1. Extending Bayesian induction

Harold Jeffreys had extraordinary insight into the use of Bayesian induction in scientific practice. We believe his ideas can be extended in ways that further rationalize and justify the approach. The article by Ly, Verghan, and Wagenmakers (in this issue) summarizes his use of Bayesian induction, leading to the result given in their Eq. (3) that the posterior odds for one model class over another is just the prior odds for those classes times the likelihood ratio: the probability of the data given one class divided by the probability of the data given the other class. Each model class consists of what (Shiffrin, Chandramouli, & Grünwald, in this issue) term model instances—model instances are members of a model class with all parameters having specified values. Thus $y = ax + b + \varepsilon$ is a class of linear functions with Gaussian noise $\varepsilon \sim N(0, \sigma)$; $y = 2x + 3 + N(0, 2)$ is a model instance within that class specifying a particular linear function with a particular value of Gaussian noise. For reasons given in Shiffrin et al. (in this issue) this commentary will instantiate all probabilities and data outcomes with discrete values (using suitably small intervals), so all integrals are replaced by finite sums. Bayesian induction implies that the posterior odds equals the Bayes Factor (BF) times the prior odds for the classes—the discrete version is given in Eq. (1). The classes being compared are made of model class instances that are here denoted θ_i and λ_i, respectively. Let $p_o(M_0) = \Sigma_i p_o(\theta_i, M_0)$ and $p_o(M_1) = \Sigma_i p_o(\lambda_i, M_1)$. Then, the Bayes Factor is the left hand term in Eq. (1): the sum across the model instances in one class of priors time likelihoods, divided by the sum of priors times likelihoods for the other class.

$$\frac{P(M_0|y)}{P(M_1|y)} = \frac{\sum_i \left[P(y|\theta_i, M_0)\, P_0(\theta_i|M_0)\right]}{\sum_i \left[P(y|\lambda_i, M_1)\, P_0(\lambda_i|M_1)\right]} \frac{P_0(M_0)}{P_0(M_1)}. \qquad (1)$$

2. Extending Bayesian induction

Bayesian induction focuses on model instances and classes. Each model instance predicts a distribution of data outcomes for the present study, and thereby assigns a probability to the observed data outcome. The model instances in all model classes

[*] Corresponding author.
E-mail address: shiffrin@indiana.edu (R.M. Shiffrin).

http://dx.doi.org/10.1016/j.jmp.2015.10.002
0022-2496/© 2015 Elsevier Inc. All rights reserved.

predict many different data distributions but certainly not all. For example a model class might propose that n, the number of successes observed in N trials, is binomial with probability p of success on each trial; for $N = 4$, the probabilities of n successes would then be ${}^4C_n p^n\ (1-p)^{4-n}$. Different values of the parameter p produce different binomial distributions for n. E.g. for $N = 4$ and $p = 0.5$ we would have the distribution [1/16; 1/4; 3/8; 1/4; 1/16] for 0 to 4 successes. However, binomial distributions hardly exhaust the total possible distributions of successes; non-binomial distributions include [0.20, 0.20, 0.20, 0.20, 0.20] and [0.6, 0, 0, 0.1, 0.3]. Shiffrin and Chandramouli (in press) proposed an extension of Bayesian induction in which priors are assigned to possible data distributions for the present experiment. The idea in short is that one (unknown) data distribution represents reality and a sample from that 'true' data distribution produces the observed data. Then Bayes Theorem is used to revise the prior distributions and produce posterior distributions for the possible data distributions. This general idea is not novel and is represented in a different form in a Journal of Mathematical Psychology article by Karabatsos (2006).

Where do model classes and model instances fit into the simple scheme? The model instances in all the classes under consideration each predict a data distribution. These predicted data distributions are only a small subset of the possible data distributions, but the priors and posteriors for each can be derived from the priors and posteriors for all data distributions by simple summation. One first defines a metric by which to compare distributions—we have termed this criterion G (it could for example be some form of Kullback–Leibler). Using G we can establish which data distributions are best matched by the predicted data distribution for each model instance. Ignoring the possibility of ties, which would not occur often enough to matter, 'best' match implies that all data distributions are partitioned into disjoint subsets; each subset contains distributions that are a best match to a given model instance. The priors (or posteriors) for a given instance are then a simple sum of the priors (or posteriors) for its best matching subset. Then the priors (or posteriors) for a model class are simply a sum of the priors (or posteriors) of the model instances in that class.

Thus what we are doing is using the observed data to carry out inference, but not inferring the probability that some model instance is 'true', but rather inferring the probability that some model instance is the *best approximation* to the truth. Of course 'best' depends on the choice of G, which depends in turn on one's goals of inference. When matching distributions one could for example choose to weight the center of distributions more than the extremes, or any other metric that satisfies one's inference goals.

A relatively full description of the proposed system and equations showing how it works are given in Shiffrin and Chandramouli (in press). A shorter description that nonetheless unpacks the brief comments above will be given in the following sections. It turns out that two figures provide the easiest way to depict how our extended version of Bayesian induction works—these are given in Figs. 1 and 2. Fig. 1 shows the situation holding before one collects data from the present experiment, based on one's prior knowledge. Fig. 2 is similar but depicts the new probabilities holding after the observed data has been taken into account.

3. Experimental outcomes

The possible measured outcomes of an entire experiment, or more precisely the subset or summary of what is measured that are reported and used for induction, are listed along the top of the table. Each potential outcome can be highly multidimensional and enormous in size (e.g. terabytes of data), but it is easiest to follow the exposition with a toy example: Suppose we wish to assess the probability $p(s)$ that a drug will produce a cure in a fixed amount of time, and apply the drug 10 times. The outcome is defined as the number of successes, n. There are 11 possible outcomes, ranging from 0 to 10, and thus there would be 11 columns in the table.

4. Data distributions

The real world produces a distribution of outcomes that will occur in the study, one of which is sampled when we carry out the study. This distribution is represented as a set of probabilities that sum to 1.0 and is represented is a row inside the table. For example if the drug had a true cure rate of exactly 0.4 for every person, the 'true' distribution would be binomial with $p(s) = 0.4$. There are numerous distributions possible, and we never know which is the 'true' one, but we use data to move our beliefs toward the truth. More precisely we use induction on the observed data to alter our current beliefs in a direction consistent with the data. Typically induction increases our confidence in certain distributions being true, and our confidence in certain model instances being best approximations to that truth.

Thus in the rows of the table we list *all* data distributions (a finite number because we have discretized probabilities). One such distribution would be the one just mentioned, binomial with $p(s) = 0.4$. Another distribution would assign 0.8 to $n = 0$, 0.1 to $n = 1$, and (0.1/9) to each of the remaining values of n. Another distribution would assign equal probabilities to all n.[1]

Why do we represent the 'true' state of the world that produces the data as a distribution? There are uncounted numbers of variables that affect the results that are unknown, un-measurable, or ignored, but would alter the results in what would be the most precise replication possible. Thus if we flipped a fair coin 100 times we might expect a binomial distribution for n with $p(s) = 0.5$. The uncountable number of unknown variables include the forces imported to the coin when it is released, the air currents and air pressure, the nature of the surface on which the coin lands, and the state of the gravitational field that depends on the current position of Jupiter with respect to our position on Earth. Such unknown variables are present in all studies including our toy example (success probability of a drug).

It should be emphasized that the possible distributions are not equally probable. Before we carry out the study or see the results we have a good deal of knowledge about the results that could be expected. We always have a good deal of such knowledge even for a study never done before. In our toy example we would not expect a distribution for n successes in 10 trials that would assign high probabilities to prime numbers (1, 2, 3, 5, 7) and low probabilities to other n's (0, 4, 6, 8, 9, 10). Such prior knowledge is a critical part of inference. The example just given is extreme, but most scientists are familiar with the importance of priors through interaction with their new graduate students: A student reports 'strange' results from a new study, results claimed to be accurate because they have been checked for errors in the program and analysis. The scientist nonetheless 'knows' the results are almost certainly in error. Subsequent investigation almost always reveals this to be the case. Even though the study may be new, the scientist's prior knowledge allows generally accurate assessment of the likelihood of error.

Thus the prior probabilities we assign to the possible data distributions are based on our prior knowledge before taking the

[1] Equal probabilities of data outcomes should not be confused with uniform probabilities of model instances (model instances are discussed shortly). If the various $p(s)$ values were equally likely then each would predict an equally likely binomial data distribution, and none would predict a uniform data distribution.

Please cite this article in press as: Chandramouli, S. H., & Shiffrin, R. M., Extending Bayesian induction. Journal of Mathematical Psychology (2015), http://dx.doi.org/10.1016/j.jmp.2015.10.002