Valencia 2010

Testing Indistinguishable Hypotheses?

★

Jan-Willem Romeijn

Faculty of Philosophy
University of Groningen

Contents

1 Conjunction fallacy

Linda is 31 years old, unmarried, assertive, and intelligent. She studied philosophy and wrote her thesis on social issues and justice. She was active in the campaign against the war in Iraq.
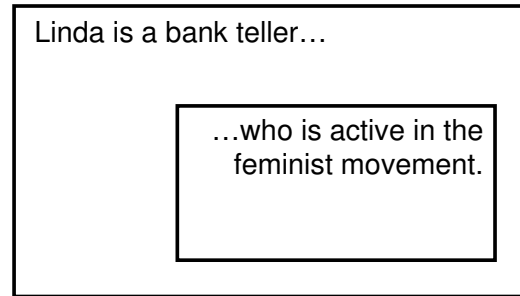

Which of the following two statements is more probable?

1. Linda is a bank teller.

2. Linda is a bank teller who is active in the feminist movement.

Probability and set inclusion

Statisticians will not give the answer that is often given by the probabilistically naive.

```
Linda is a bank teller…

        …who is active in the
          feminist movement.
```

Probability is a measure over sets, and the set of worlds in which Linda is a feminist bank teller is strictly included in the set of worlds in which she is a bank tellers.

Bayesian model selection

The idea behind BMS is to assign probabilities to models, and compare these models by means of the marginal, or average, likelihoods:

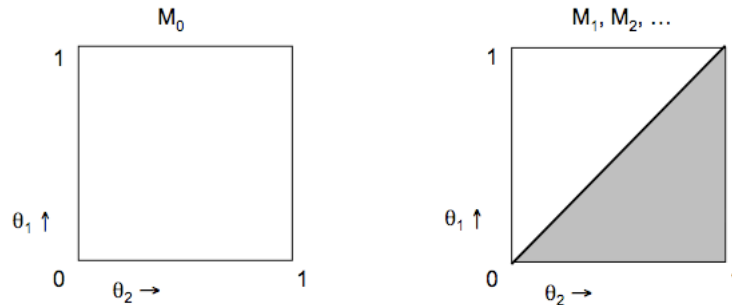$$\frac{P(M_1|x)}{P(M_0|x)} = \frac{P(x|M_1)}{P(x|M_0)} \frac{P(M_1)}{P(M_0)}$$

where

$$P(x|M_i) = \int_{M_i} P(x|\theta)P(\theta)d\theta.$$

For want of a prior over models, the likelihood ratio is often taken as the sole guide to choosing the model.

Model selection and Linda
It seems that Bayesian model selection is prone to committing the conjunction fallacy.

The probability of the constrained model may get larger than the probability of the unconstrained model. What do those probabilities refer to?

Bayesian information criterion

The BIC is an approximation to the marginal log-likelihood of a model in the long run:

$$P(x|M_i) \sim -\log P(x|\hat{\theta}_x) + d \log n$$

in which $\hat{\theta}_x$ is the maximum likelihood (ML) estimation of θ for the data x, and n is the size of the data set x. It seems that the BIC is prone to committing the same conjunction fallacy as BMS.

2 Model selection

Are other model selection tools also prone to the conjunction fallacy? And if not, can we understand BIC along the lines of the other ICs?

AIC We estimate the expected distance between the truth and the ML-estimation in the model.

DIC We estimate the expected predictive accuracy of the ML-estimation in the model.

These selection tools employ some penalty term, typically the number of model parameters d, as a measure of complexity. Note that prima facie this makes choosing between inequality constrained models different.

AIC

The general idea of the AIC is that we determine the estimated distance between the true chance $\theta^\star$ and the chance θ estimated in the model.

$$\Delta(\theta^\star, \theta) \sim E_{P(x|\theta^\star)}[-\log P(x|\theta)]$$

In this expression θ can still range over the whole model, while distances are of course between two points. This is resolved by taking an expectation towards a dummy data set y.
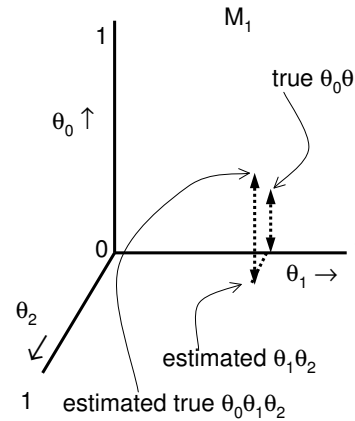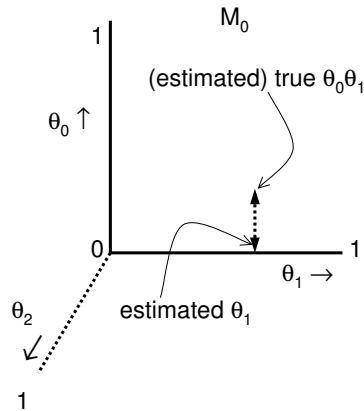
$$\Delta(\theta^\star, \theta) \sim E_{P(y|\theta^\star)}\left[E_{P(x|\theta^\star)}[-\log P(x|\hat{\theta}_y)]\right]$$

In integrating this data set out, the number of dimensions appears:

$$\Delta(\theta^\star, \theta) \sim E_{P(x|\theta^\star)}[-\log P(x|\hat{\theta}_x)] + d \sim \log P(x|\hat{\theta}_x) + d$$

AIC continued

Formulas are perhaps not the best way of explaining the procedure... Here is an artist's impression of how the distance to an unknown truth $\theta^\star$ is estimated.

DIC
In the DIC we determine a Bayesian expectation for the parameter $\overline{\theta}_y$, and we see how well this expected value predicts data from the true distribution, $P(x|\theta^\star)$.

$$E_{P(x|\theta^\star)}[-\log P(x|\overline{\theta}_y)] \sim -\log P(y|\overline{\theta}_y) + d$$

The predictive accuracy is measured by the logarithmic loss function:

$$-\log P(x|\theta)$$

This is also only term that matters to the distance function featuring in the AIC. Although different in interpretation, the loss function expressing predictive accuracy and the distance to the truth may be used interchangeably.

BMS

Clearly AIC and DIC do not commit the conjunction fallacy. But can we frame the BIC as the other ICs? There is some similarity in the use of average likelihoods:

$$-\log P(y|M) \sim E_{P(\theta)}[-\log P(y|\theta)].$$

But this is a far cry from presenting BMS on a par with the other ICs.

- The true parameter value $\theta^\star$ is not alluded to in the BMS.

- BMS does not rest on an expression involving all possible data x but only the data actually obtained.

In the following we are going to bite the bullet and see what else the probabilities featuring in BMS might mean.
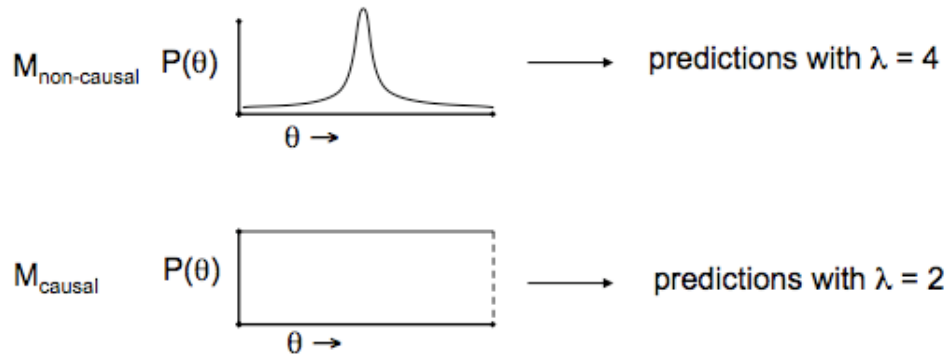
3 Comparing causal models

Say that we compare the following causal hypotheses about two binary variables:
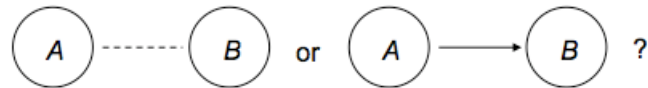
The non-causal model has a so-called simplex as its parameter space, with four independent probabilities adding up to one. The causal model is represented by a parameter for the probability of A, and two parameters for the conditional probabilities of B on A.

Different uniform priors

The different parameterisations, associated with a different causal picture, motivate different uniform priors over what is essentially the same parameter space.

$M_{\text{non-causal}}$ $P(\theta)$ $\longrightarrow$ predictions with $\lambda = 4$

$\theta \rightarrow$

M_{causal} $P(\theta)$ $\longrightarrow$ predictions with $\lambda = 2$

$\theta \rightarrow$

Different predictive properties

Depending on the prior, we find predictions that approach the true parameter value more or less quickly.

$$P(y_{t+1} = 1 | y_1 \cdots y_t) = \frac{\lambda}{t + \lambda} \frac{1}{2} + \frac{t}{t + \lambda} \frac{t_1}{t}.$$

The same expression also captures the expected values for the parameter, because

$$P(y_{t+1} = 1 | \theta, y_1 \cdots y_t) = \theta \qquad P(y_{t+1} = 1 | y_1 \cdots y_t) = \int \theta P(\theta) d\theta = \overline{\theta}_y.$$

The marginal likelihoods thus capture how the expected value for θ approaches the true value.

An example

For the causal hypotheses at hand, we can derive the following likelihood ratio for the two models as follows:

$$BF = \frac{P(y_1 \cdots y_t | M_{\text{causal}})}{P(y_1 \cdots y_t | M_{\text{non-causal}})} = \frac{6(t_0 + 1)(t_1 + 1)}{(t + 2)(t + 3)}$$

We find the following interesting points:

Number of observations t	Interval in which $BF > 1$
< 12	$-$
12	$\frac{1}{2}$
48	$[\frac{1}{4}, \frac{3}{4}]$
∞	$[\frac{1}{2} - \frac{1}{2\sqrt{3}}, \frac{1}{2} + \frac{1}{2\sqrt{3}}]$

This means that with less than 12 observations, the causal model always performs better! What is going on?
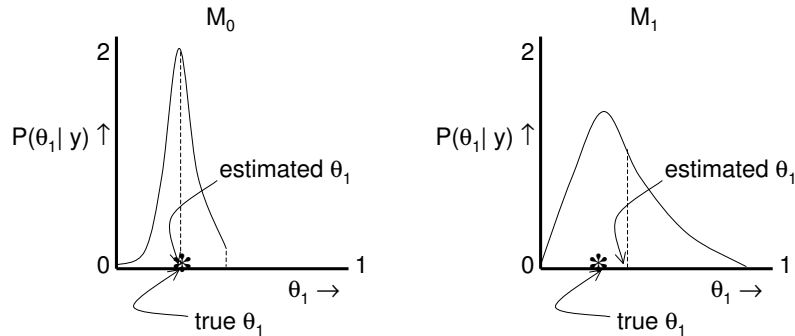
4 Convergence measure

In the foregoing example, the Bayes-factor measures how fast the expected value of the parameter $\overline{\theta}_y$ approaches the true value $\theta^\star$.
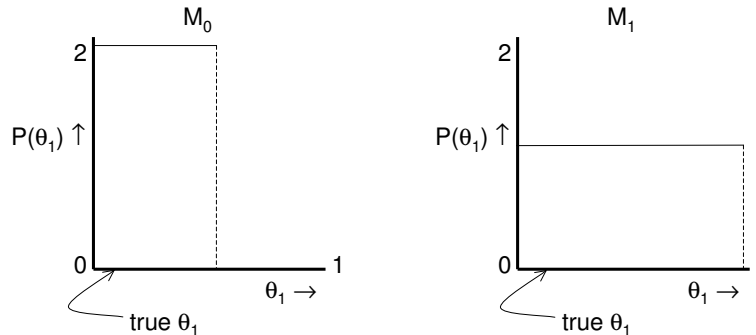
We can view BMS in this way more generally: we are comparing the priors on their convergence properties.

Choosing the best expectation

Both models generate an expected value for the parameter value $\overline{\theta}_y$.

If $\theta^\star$ lies within the restricted region, the expected value of the restricted model will be closer to the true value.

What marginal likelihood measures

Marginal likelihood thus combine the aforementioned aspects of model selection in a particular way.

- The Bayes' factors measure the online predictive performance of the models. This performance is determined by

  - the probability distributions in the model, and
  - the prior probability over the model.

- The online predictive performance is a measure for the distance to the true parameter value: it expresses the speed of approaching the maximum likelihood estimation.

However, the performance and distance are determined by the data that we have obtained, and not by all data that we could have got under the true distribution.

5   Future research

This leaves many questions unanswered. . .

- Can we derive a BIC for equi-dimensional models that only differ in the prior defined over them? For inequality-constrained priors the answer is affirmative, but the general case is unclear.

- Following the BIC, we can still view BMS as generating a kind of penalty term for complexity, although different from the penalty that the ICs give. But what exactly does it penalise for?

- We cannot interpret posterior model probabilities (PMPs) as probabilities that the true parameter value is included in the region. Is the concept of PMP terminally ill?