Inequality Constrained Modeling
Utrecht 2008

Testing Indistinguishable Hypotheses

⋆

Jan-Willem Romeijn
Faculty of Philosophy
University of Groningen

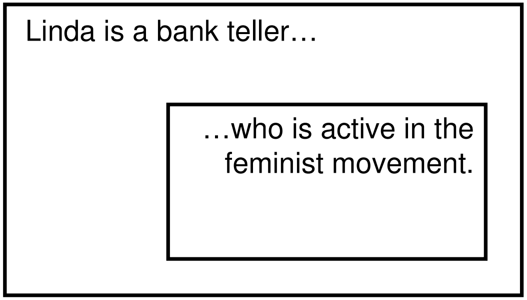
Linda the bank teller

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.

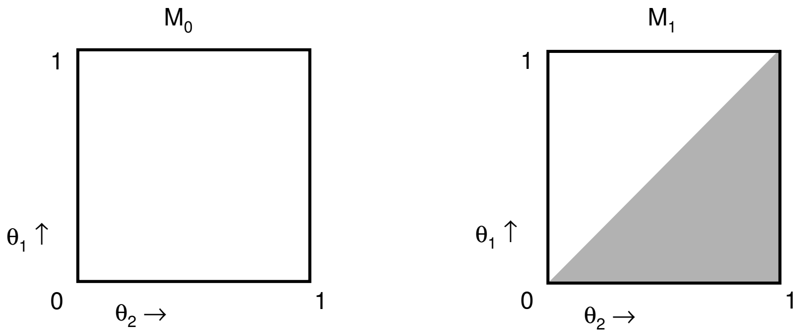
Probabilistic fallacy
Statisticians will not give the answer that is often given by the probabilistically naive.



Probability is a measure function over sets, and the set of feminist bank tellers is strictly included in the set of bank tellers.

Inequality constrained modeling

Nevertheless, it seems that some applications of Bayesian model selection make exactly the same kind of mistake.



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

Contents

1 Model selection

We can try to understand the use of Bayes' factors for inequality constrained models as a kind of model selection.

AIC We estimate the distance between the truth and the estimation following from the model.

DIC We determine the expected predictive accuracy of the estimation.

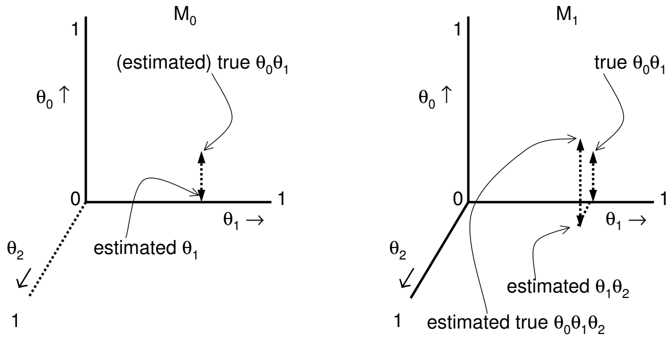
BIC We approximate the marginal likelihood of the models under comparison.

All these selection tools employ some penalty term, typically the number of model parameters d, as a measure of complexity. So choosing between inequality constrained models is prima facie different.

AIC

The general idea of the AIC is that we determine the estimated distance between the true chances and the chances estimated in the model.

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



DIC

In the DIC we determine an expected $\overline{\theta}_y$, 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 loss function:

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

This is also the log-likelihood of the hypotheses θ. As in the AIC, the loss can be viewed as a distance to the truth.

BMS

The use of likelihoods as contributing to such a distance suggests a relation to the use of Bayes-factors in BMS:

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

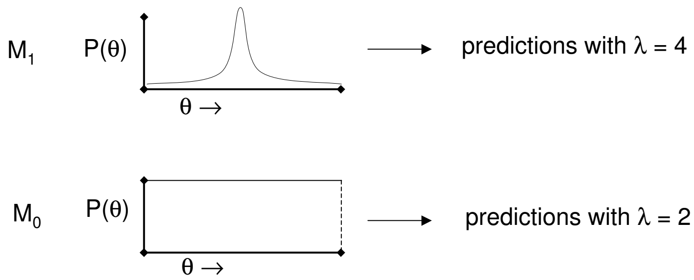
But this is a far cry from presenting BMS on a par with the other ICs.

- Up until now the true parameter value $\theta^\star$ is not alluded to in the BMS.
- In contrast to the ICs, BMS does not seem to rest on a distance involving all possible data x, as drawn from the true distribution of $\theta^\star$.

Perhaps we can present BMS as an approximation of such an IC. But the appeal of BMS seems that it has a stand-alone motivation.

2 Comparing causal models

Say that we compare two prior probability assignments over exactly the same space of Bernoulli hypotheses:



The different priors may be motivated by two different parameterisations of the same model, associated with a different causal picture.

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

Choosing $\lambda_0 = 2$ and $\lambda_1 = 4$, we can derive the likelihood ratio for the two models, as follows:

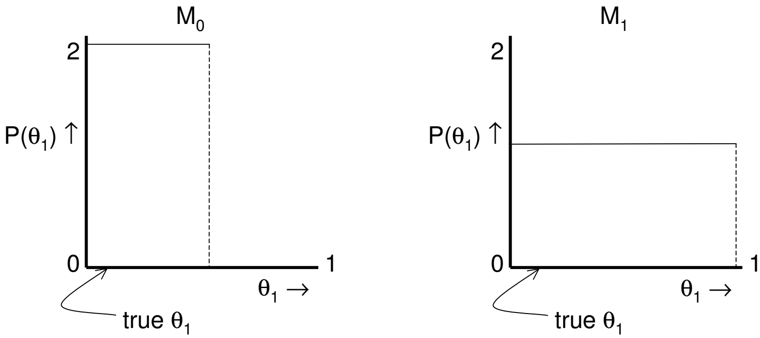
$$BF_{01} = \frac{P(y_1 \cdots y_t | M_1)}{P(y_1 \cdots y_t | M_0)} = \frac{6(t_0 + 1)(t_1 + 1)}{(t + 2)(t + 3)}$$

We find the following interesting points:

Number of observations t	Interval in which $BF_{01} > 1$
< 12	–
12	$\frac{1}{2}$
48	$[\frac{1}{4}, \frac{3}{4}]$
∞	$[\frac{1}{2} - \frac{1}{2\sqrt{3}}, \frac{1}{2} + \frac{1}{2\sqrt{3}}]$

3 Convergence measure

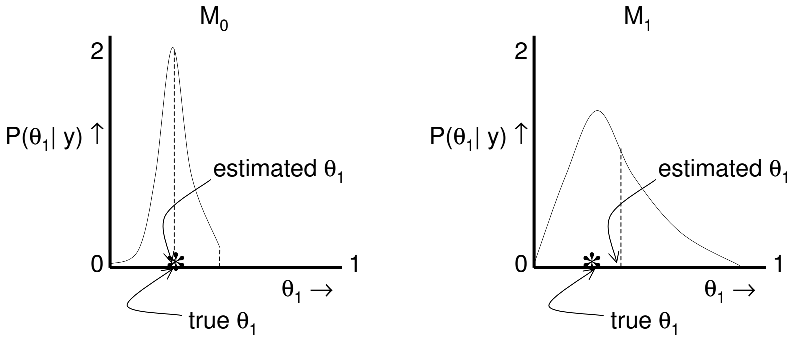
In the foregoing example, the Bayes-factor measures how fast the expected value of the parameter $\overline{\theta}_y$ approaches the true value $\theta^{\star}$.



We can view BMS with inequality constrained models in the same way. We compare the priors, and we look at the 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 combines the aforementioned aspects of model selection in a particular way.

- Bayes' factors measure the relative predictive performance of the models. This performance is determined by
  - the probability distributions in the model, and
  - the prior probability over the model.
- At the same time, this relative predictive preformance is a measure for the distance to the true parameter value.

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

4 Future research

This leaves many questions unanswered. Some of these relate to BMS as model selection:

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

- The BIC also starts with marginal likelihoods, but we can derive an analytic approximation of it in which fit and dimensionality show up. Perhaps a similar approximation can be found for the Bayes-factors in inequality-constrained models.

What does marginal likelihood measure?

Other research questions relate to the attempt to interpret PMPs as normal probabilities:

- If we take the Linda story seriously, we cannot interpret the PMPs as the probability that the true parameter value is included in the region. Does this mean that to emply PMPs we must work with non-nested models?
- The use of inequality-constrained models gives particular hypotheses a head start, so that we may find the truth more quickly. Are we thereby testing the inequalities?

Thank you

The slides for this talk will be available at http://www.philos.rug.nl/ romeyn.
For comments and questions, email j.w.romeijn@rug.nl.