Date 15.04.2015

Subjectivity in evidence

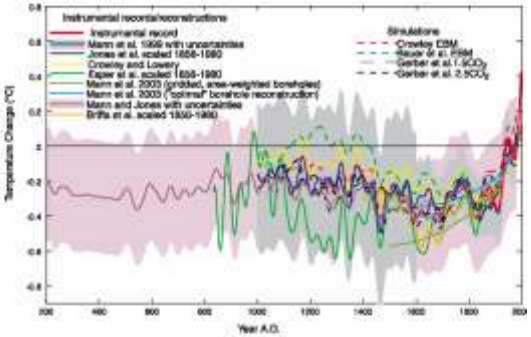
Three studies in Bayesian model evaluation

Presentation for PSF conference 2016

Jan-Willem Romeijn
Faculty of Philosophy
University of Groningen

Data vs evidence

Data do not all by themselves present evidence one way or another. They become evidence when confronted with theory.


Instrumental records/reconstructions
Instrumental record
Mann et al. 1999 with uncertainties
Jones et al. scaled 1856-1980
Crowley and Lowery
Esper et al. scaled 1856-1980
Mann et al. 2003 (gridded, area-weighted boreholes)
Mann et al. 2003 ("optimal" borehole reconstruction)
Mann and Jones with uncertainties
Briffa et al. scaled 1930-1960
Simulations
Crowley EBM
Bauer et al. EMM
Gerber et al. 1.5CO₂
Gerber et al. 2.5CO₂
Temperature Change (°C)
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
200
400
600
800
1000
1200
1400
1600
1800
2000
Year A.D.

Accordingly, whether or not data confirm or disconfirm a statistical model is dependent on the models that frame them.

This paper

I will argue this point in three confirmation-theoretic analyses that use Bayesian model evaluation.

The main insight is that evidence in model evaluation is a *subjective* notion, but that this is something to enjoy rather than lament.

Programme

① The Bayesian evaluation of statistical models

② A Bayesian notion of evidence

③ Testing priors: a model for abduction?

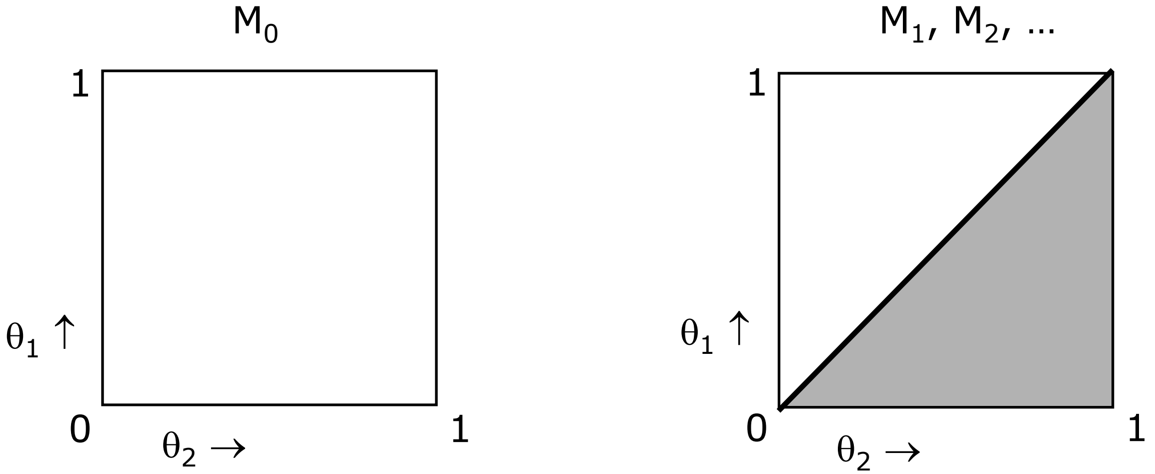
④ Calibrating and testing the model: no double counting

⑤ Perfect fit at little cost: Bayes to the rescue

⑥ Discussion

① Bayesian Model evaluation

Model selection helps us to determine what statistical parameters to include in our model.



Making the data more probable is not the only criterion for this: we can include too many parameters.

Model selection tools

Several information criteria (ICs) are on offer. Remarkably, they lead to roughly the same evaluation of models:

Model score = Fit of best hypothesis in model
- Penalty for model complexity
= Log-Likelihood of best hypothesis
- Adjusted number of parameters

Notably, the information criteria only *approximate* good-making model features, e.g., past predictive performance.

BIC, BFs, and BMS

The Bayesian information criterion (BIC) approximates the marginal likelihood of the model.

$$P(E|\mathcal{M}) = \sum_{H \in \mathcal{M}} P(E|H)\, P(H|\mathcal{M})$$

The marginal likelihoods can be used to compute the Bayes factor (BF) and the posterior probability in Bayesian model selection (BMS).

BME, not BMS

In Bayesian model evaluation, the probability of a model is determined in the usual Bayesian way.

$$P(\mathcal{M}|E) = P(\mathcal{M})\,\frac{P(E|\mathcal{M})}{P(E)}$$

The probability of evidence is again a weighted average:

$$P(E) = \sum_{\mathcal{M}} P(\mathcal{M})\,P(E|\mathcal{M})$$

A better account of confirmation?

Henderson *et al* (2010) argue that we can illuminate scientific confirmation by casting it in the format of Bayesian model evaluation, and I agree.

- › Theories are best understood as offering a range of hypotheses that share certain features.
- › This view on theory elegantly captures the informative-ness and the simplicity of theory.
- › We might even hope to capture the explanatory virtues of theory.

② Subjectivity of evidence

We call data *evidence* when it impacts on our opinion (cf. Morey *et al* 2016).

› Data are spelled out in terms of constraints to a set of possible worlds or samples.

› Our opinion can be spelled out as a probability function over possible samples.

› In a Bayesian theory of evidence, the impact of data on opinion is modelled as conditioning.

Dependence on the prior

In BME, the impact of data on the probability of a model, as measured by the marginal likelihood, depends crucially on the prior over the hypotheses within the model.

$$P(E|\mathcal{M}) = \sum_{H\in\mathcal{M}} P(E|H)\,P(H|\mathcal{M})\,dH$$

This makes the notion of evidence within BME highly subjective: the prior might well be chosen on subjective grounds.

Using the subjectivity

In what follows we consider three cases in which this aspect of BME is entirely benign or even beneficial.

› A model of Bayesian abduction: exploiting the fact that priors are motivated by theoretical considerations.

› A model of use-novelty: explaining the fact that we may sometimes “double-count” evidence.

› A model of implicit complexity: using priors within models to avoid faulty curve-fitting.

③ Abducted by Bayesians?

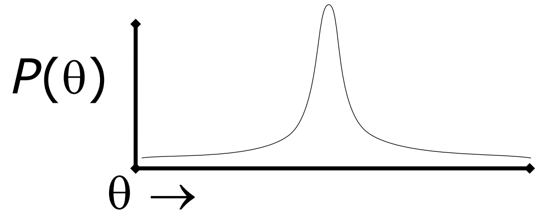
Romeijn (2013) and Henderson (2013) offer similar Bayesian models for abductive inference, based on BMS.

The idea from Romeijn (2013) can be illustrated with a coin that may be from my wallet, or from a conjurer’s box.

Comparing the priors

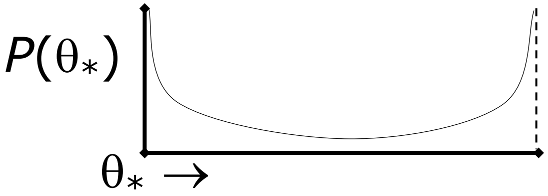
The difference between the normal and the magical coin is expressed in a prior probability over the bias hypotheses.

Θ:



$\longrightarrow$ predictions c_λ with $\lambda = 12$

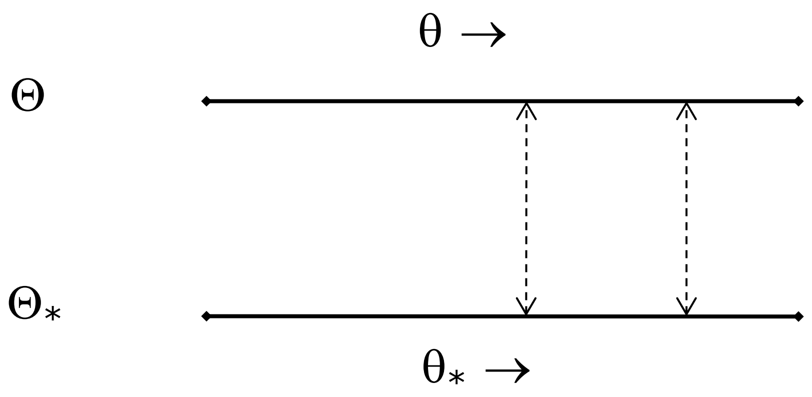
Θ_*:



$\longrightarrow$ predictions c_λ with $\lambda = 1/3$

Infusing the data with theory

The distinction between the models for the normal and magical coins is strictly theoretical.



In virtue of the priors, the data nevertheless impacts differently on the two models.

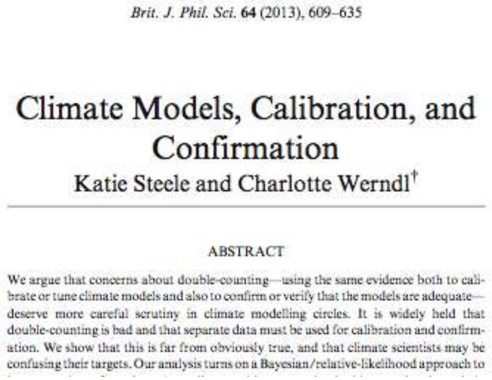
Employing theory-ladenness

We need not regret that priors make the impact of data, and hence the evidence, subjective.

Instead we can use this to make the data speak about theoretical distinctions, as expressed in the priors.

④ The double use of evidence

Several authors have argued that evidence must be *use-novel*: it may not be used twice.



Brit. J. Phil. Sci. **64** (2013), 609–635

Climate Models, Calibration, and Confirmation

Katie Steele and Charlotte Werndl[†]

ABSTRACT

We argue that concerns about double-counting—using the same evidence both to calibrate or tune climate models and also to confirm or verify that the models are adequate—deserve more careful scrutiny in climate modelling circles. It is widely held that double-counting is bad and that separate data must be used for calibration and confirmation. We show that this is far from obviously true, and that climate scientists may be confusing their targets. Our analysis turns on a Bayesian/relative-likelihood approach to

Recently Steele and Werndl (2013) have nuanced this: double-counting may be okay under certain circumstances.

Calibrating and testing at once

In a Bayesian evaluation of the models, we adapt the probability within the model and of the model in one single operation.

$$\begin{aligned} P(H|\mathcal{M}) &\quad \rightarrow \quad P(H|\mathcal{M} \cap E) \\ P(\mathcal{M}) &\quad \rightarrow \quad P(\mathcal{M}|E) \end{aligned}$$

It is inherent to this way of evaluating models that it is dependent on the priors within the models that are compared.

Use-novel evidence

Against the backdrop of Bayesian model evaluation we can propose a framing-dependent notion of use-novelty.

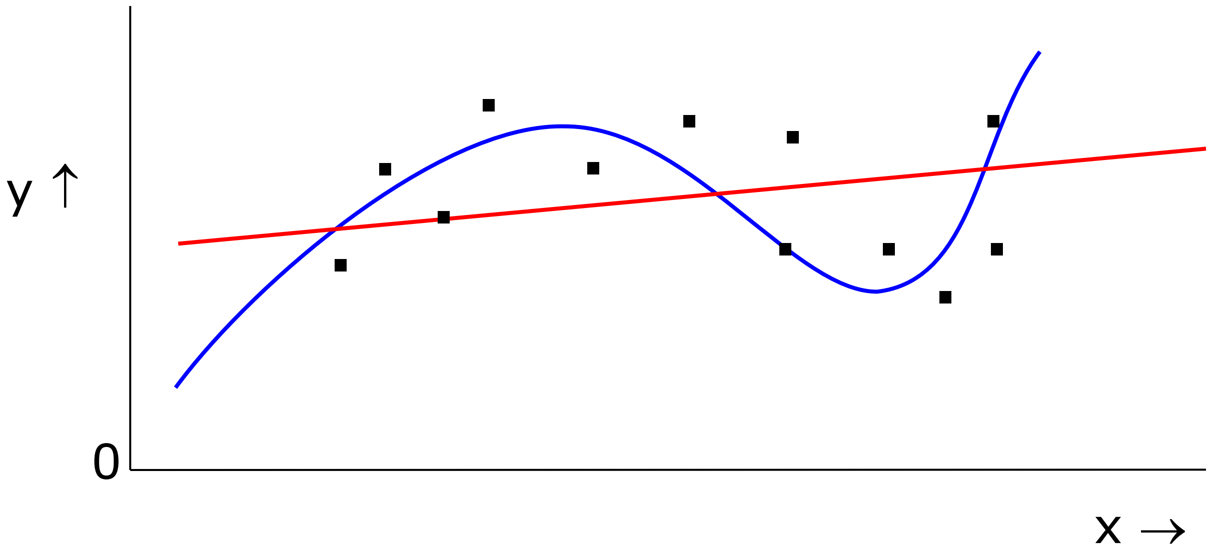
› One learning experience must be represented uniquely by one datum.

› We use this datum by conditioning on it. Its evidential value is determined by its impact on opinion.

› After one conditioning operation, the datum ceases to be evidence: renewed conditioning will not impact on our opinions.

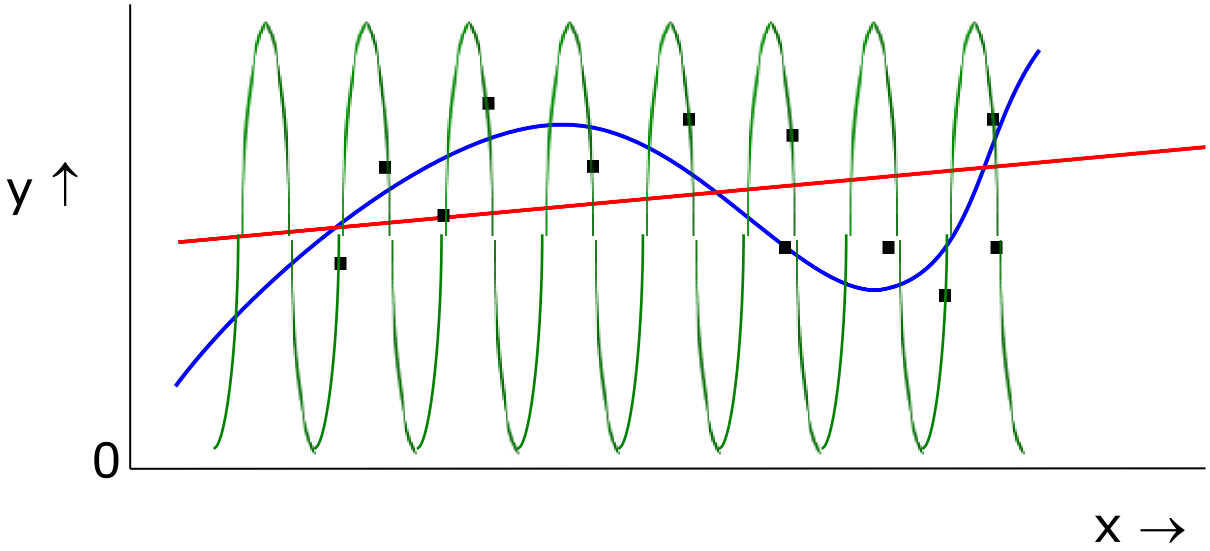
⑤ Perfect fit at little cost

In curve fitting, we choose between candidate curves from families that differ in the number of adjustable parameters.



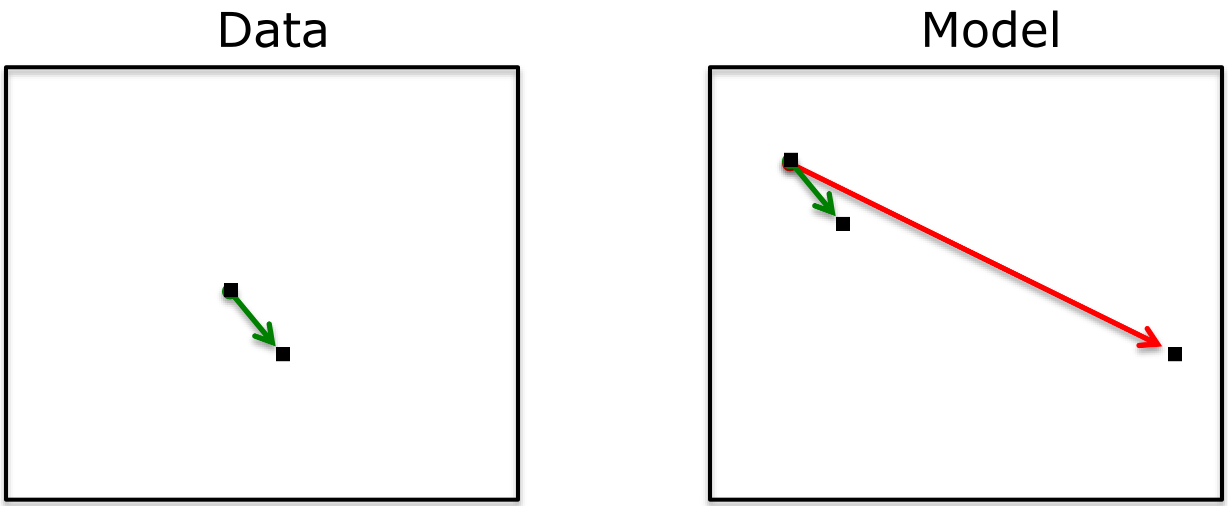
Implicit complexity

How to evaluate the sine model below? It has only three parameters, it is fully generic, and its fit is perfect! The known ICs give the wrong answer.



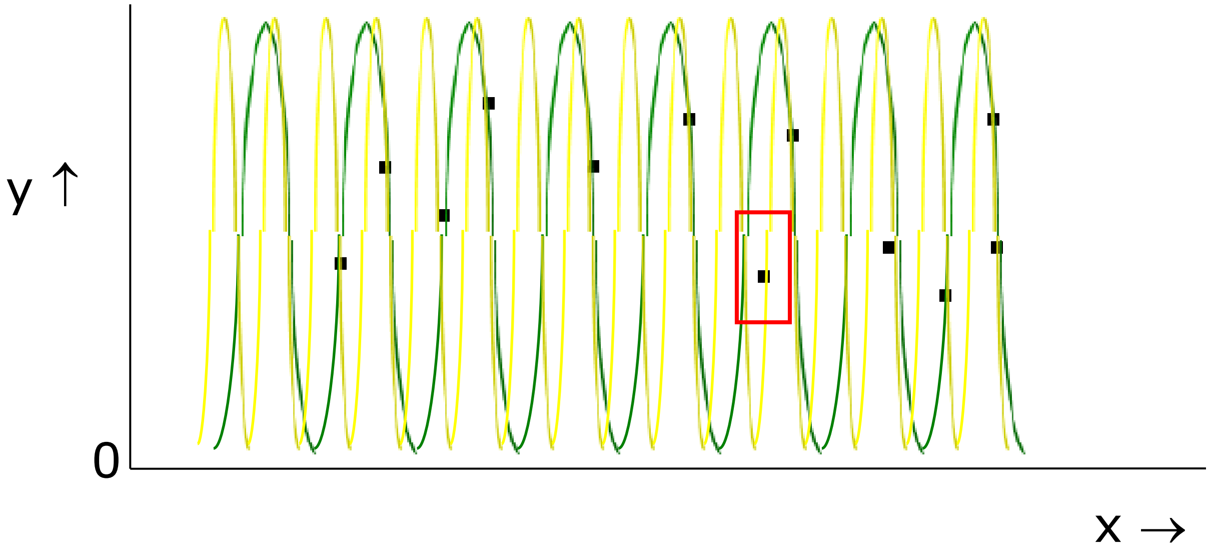
Model sensitivity

The best estimate in a model must be robust under tiny changes in the data: a good model is not skittish.



Sensitivity of estimations

The sine model suffers from exactly that defect: nudging the data space cause the best estimate to change radically.



Sensitivity in the BIC

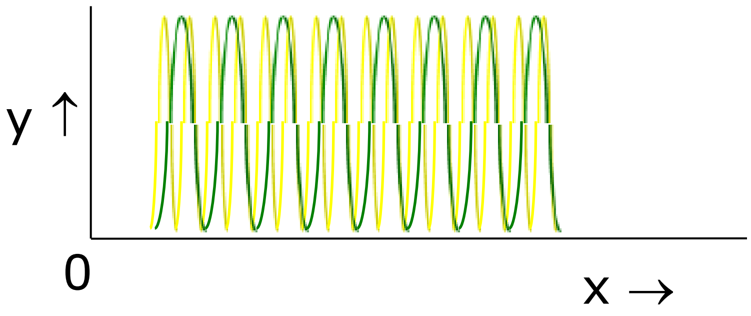
The Bayesian information criterion is an approximation of the marginal likelihood of the model.

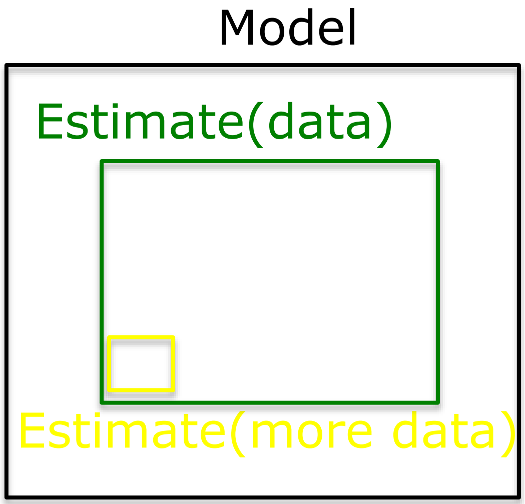
Score(M) = Fit × Order(n) – Dimension × Order(log n)
– Fisher Information × Order(1) + …

The sensitivity shows up in this approximation as the Fisher information, but in the wrong order in data size.

Priors to the rescue

The sine model is degenerate: many hypotheses fit the data. With additional data points, the set of best estimates gets smaller.





Diminishing priors

The measure of the set with best estimates also diminishes rapidly with the addition of new data points.

$$\text{Score}(M) \quad = \quad \text{Fit} \times \text{Order}(n) - \text{Dimension} \times \text{Order}(\log n)$$
$$- \text{Diminishing prior} \times \text{Order}(n) + \ldots$$

This will show up in a properly approximated marginal likelihood: the prior term will act as substantial penalty.

Subjective evaluation

This diminishing of the prior is dependent on what prior was chosen at the outset.

This is a version of the so-called “sub-family problem”: we can always construct the prior with hindsight to give the best fitting hypotheses a high prior in advance.

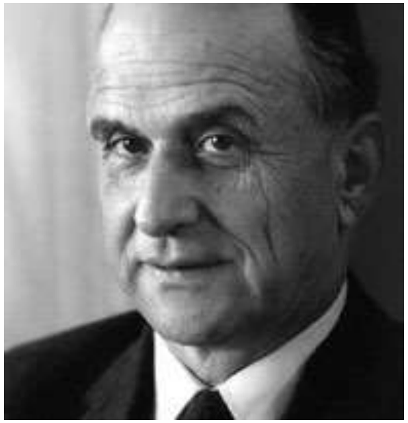
⑥ The upshot

Bayesian model evaluation nicely accommodates a number of challenging cases of evidence handling.

The fact that correct evidence handling requires subjective input rather works to its advantage.

A philosophical classic

This insight echoes well-known facts about induction: the choice of language and perspective is central.



It drives home the subjective and contextual nature of evidence.

Thanks

With questions and remarks, please email:

`j.w.romeijn@rug.nl`

Slides will be made available on my website:

`http://www.philos.rug.nl/~romeyn`