university of groningen

NWO
Netherlands Organisation for Scientific Research

Novel Predictions Workshop
Düsseldorf 2011

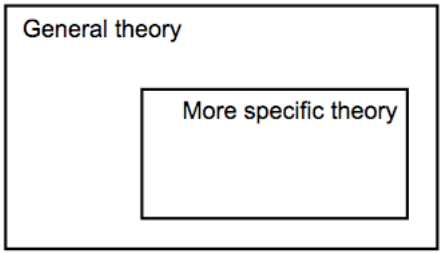
# Specificity, Accommodation and the Sub-family Problem

⋆

Jan-Willem Romeijn
Faculty of Philosophy
University of Groningen

## Specificity vs probability

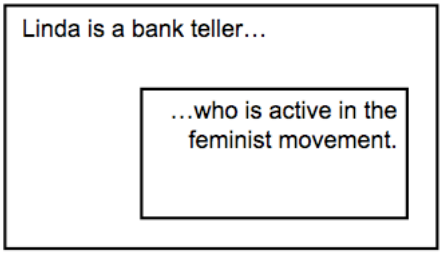
We prefer true scientific theories that allow for few possibilities over true ones that allow for many. Theories must be specific.




By the very nature of probability as a measure of sets, it seems that this preference cannot be captured by probabilistic confirmation theory.

**Specificity and Linda**

We know our intuitive preference for specific hypotheses from Linda the bank teller. Recent confirmation-theoretic solutions to this problem employ the comparative aspect of confirmation.




In this talk I will provide a solution to the problem of specificifty based on model selection techniques.

# Contents

# 1 Theories as sets of hypotheses

Theory appraisal concerns the comparison of hypotheses from a given set $M_i$. In statistics, for example, we have

$$\hat{\theta}(D_n, M_i) = \{H_\theta : P(D_n|H_\theta) \text{ is maximal}\}.$$

A closer look at theories shows that they are better understood as sets of possible ways the world can be. Each theory comprises of a set of particular hypotheses:

$$M_i = \{H_\theta : \theta \in R\}.$$

The specificity of a theory is determined by the size of the region $R$.

**Example theories**

We will consider a comparison between the following two theories:

$$M_0 = \{H_\theta : \theta \in [0, 1]\},$$

$$M_1 = \left\{H_\theta : \theta \in \left[0, \frac{1}{2}\right]\right\},$$

where $M_0$ is a so-called encompassing theory, and $M_1$ constrained. In this setup the theories $M_0$ and $M_1$ in some sense overlap so that a comparison of posteriors is nonsensical, but this can be amended easily.

## 2 Theory appraisal by model selection

Model selection tools facilitate the comparison of sets of hypotheses on their ability to accommodate the data. The Bayesian information criterion ($BIC$) compares by approximated marginal likelihood:

$$P(D_n|M_i) \approx -2 \log P(D_n|H_{\hat{\theta}} \cap M_i) + d_i \log(n) = BIC(M_i).$$

When choosing between theories, we trade specificity against fit, expressed in

- the maximum likelihood term $P(D_n|H_{\hat{\theta}} \cap M_i)$ and
- the dimensionality term $d_i$.

Model selection by $IC$s provides a probabilistic account of theory appraisal that factors in specificity, as captured by dimensionality.

**A shortcoming of standard *IC*s**

Now consider the theories $M_0$ and $M_1$ of the foregoing and say that the best fitting hypothesis $H_{\hat{\theta}}$ is included in both:

$$\hat{\theta}(D_n, M_0) = \hat{\theta}(D_n, M_1) = \frac{1}{3}.$$

Then the maximum likelihood terms are equal, as are the dimensions of the theories and the number of observations:

$$d_0 = d_1,$$
$$P(D_n | H_{\hat{\theta}} \cap M_0) = P(D_n | H_{\hat{\theta}} \cap M_1).$$

In fact, none of the standard *IC*s (Akaike, Bayesian, Deviance) captures the difference in specificity between the constrained and encompassing theories $M_0$ and $M_1$.

## 3 Adapting *BIC* to capture specificity

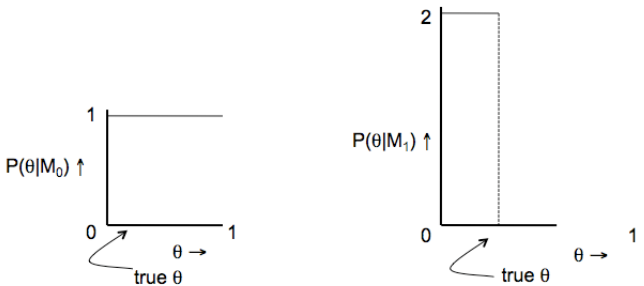
We can adapt the *IC*s to incorporate this aspect of specificity. We illustrate this for the *BIC*.

$$P(H_\theta|M_1)d\theta = \frac{1}{P(M_1|M_0)}P(H_\theta|M_0)d\theta.$$

To derive the *BIC*, we assume a prior over $M_1$ that is the truncated version of the prior over $M_0$. The normalisation is the relative size of the theory $M_1$ to $M_0$.

**Marginal likelihood for increasing sample size**

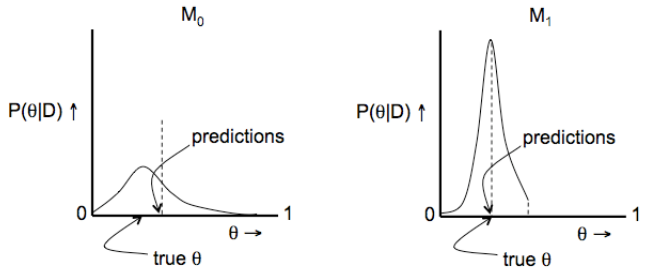
The constrained theory $M_1$ starts off with its prior probability closer to the maximum likelihood point than the encompassing model $M_0$.




At any point the ratio of the priors expresses the ratio in the size of the two theories.

**Marginal likelihood for increasing sample size**

With increasing sample size $n$ the posterior probability accumulates around the maximum likelihood point. But the theory $M_1$ begins with a head start.




Therefore the predictions from theory $M_1$ are more accurate than those of theory $M_0$, and its marginal likelihood is higher.

**Limiting marginal likelihood ratio**

The marginal likelihoods of $M_0$ and $M_1$ are dominated by what goes on in its immediate vicinity. If $\hat{\theta}$ lies within $M_1$, we find that

$$\lim_{n\to\infty} \frac{P(D_n|M_1)}{P(D_n|M_0)} = \frac{1}{P(M_1|M_0)}.$$

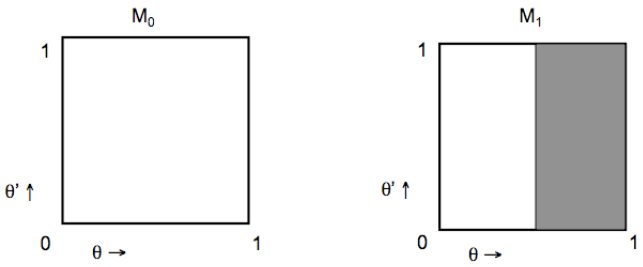
And if $\hat{\theta}$ lies outside $M_1$, we have

$$\lim_{n\to\infty} \frac{P(D_n|M_1)}{P(D_n|M_0)} = 0.$$

So the ratio of marginal likelihoods tends to the ratio of priors at the maximum likelihood point.

**Evaluation of the result**

The above reveals what must come out of an improved *BIC* approximation. Rather than deriving this so-called prior-adapted *BIC* in full, we concentrate on the interpretation of the additional term.




The term by which the marginal likelihoods differ is actually the relative size of the constrained model: it expresses a difference in specificity.

## Derivation of the prior-adapted *BIC*

In the original derivation of Schwarz (1978), it is shown that

$$\begin{aligned}\log P(D_n|M_i) &= \log P(D_n|H_{\hat{\theta}} \cap M_i) - \frac{d_i}{2}\log(n) + \log P(H_{\hat{\theta}}|M_i) \\ &\quad +(d/2)\log(2\pi) - \frac{1}{2}\log|I| + O\left(\frac{1}{\sqrt{n}}\right),\end{aligned}$$

where $I$ is the expected Fisher information matrix for a single observation. Following Kass and Wasserman (1992), we can eliminate the terms of order $O(1)$ by a clever choice of prior:

$$\log P(H_{\hat{\theta}}) = \frac{1}{2}\log|I| - \frac{d_i}{2}\log(2\pi).$$

This prior can be justified independently: it expresses that we have a roughly correct idea of where the maximum likelihood point will be.

**Retaining the prior term**

The key idea of the prior-adapted BIC is that this last step in the original derivation must be omitted. The effect of the truncated prior can be found back in the prior probability density. This motivates the proposal of the prior-adapted *BIC*:

$$PBIC(M_i) = -2 \log P(D_n | H_{\hat{\theta}} \cap M_i) + d \log(n) - 2 \log P(H_{\hat{\theta}} | M_i).$$

Because the priors over $M_0$ and $M_1$ differ by a factor $P(M_1|M_0)$, we find for $H_{\hat{\theta}}$ in $M_1$ that

$$PBIC(M_0) - PBIC(M_1) = -2 \log P(M_1|M_0) > 0.$$

The terms pertaining to likelihood and dimensionality do not differ. If $H_{\hat{\theta}}$ lies outside $M_1$, then the difference in likelihood terms dominates the comparison of *PBIC*.

**And the other $O(1)$ terms?**
While the other terms of this order in the derivation of Schwarz do not disappear, they are both equal for models that differ by constraints.

- The term $\frac{d_i}{2}\log(2\pi)$ is clearly the same, as it only depends on the dimension which is equal for the encompassing and constrained model.
- The term $\frac{1}{2}\log|I|$ is also the same. It is the expectation of the second order derivative of the likelihood of a single observation, evaluated at the maximum likelihood point. But the models have exactly the same likelihood function.

One worry may be that the accuracy of Schwarz's approximation is different for the models. But nothing in that approximation hinges on the exact region of admissible parameter values.

**Some more detail on the derivation**

The original derivation employs the so-called Laplacian method for integrals on a Taylor expansion of the function $g(\theta) = P(H_\theta|M_i)P(D_n|H_\theta \cap M_i)$, as it appears in the marginal likelihood. This leads to

$$P(D|M_i) = \exp\left[g(\tilde{\theta})\right](2\pi)^{\frac{d_i}{2}}|A|^{-\frac{1}{2}} + O\left(\frac{1}{n}\right),$$

with $\tilde{\theta}$ the value where the function $g(\theta)$ is maximal. It is assumed that $g(\tilde{\theta})$ can be approximated by $g(\hat{\theta})$. The remaining terms $-\frac{d_i}{2}\log(n) - \frac{1}{2}\log|I|$ result from

$$|A| = n^d|I| + O\left(\frac{1}{\sqrt{n}}\right).$$

This approximation is based on two further assumptions: the observations in $D_n$ are independent and identically distributed, and the second derivative of $g(\theta)$ is dominated by the likelihood factor, so that we can omit $P(H_\theta|M_i)$ from $g(\theta)$.

# 4 Specificity and model selection

There are a number of model selection tools available, each with their own motivation:

**BIC** We choose the theory with the largest approximated marginal likelihood.

**AIC** We choose the theory whose approximated distance to the hypothesized truth is minimal.

**DIC** We choose the theory that has the best expected predictive performance under a particular loss function.

A very attractive feature of the information criteria is that they independently arrive at very similar expressions:

$$IC \sim -\text{Fit}[P(D_n|H_{\hat{\theta}})] - \text{Specificity}[-d_i \log n]$$

**Dimension and size as specificity**

The dependence on the dimension $d_i$ drops out of the approximation methods for all the $IC$s. It is not put in to express specificity, but interpreted as an expression of specificity afterwards.

$$-d_i \log n \sim \frac{1}{\#\text{ theoretical possibilities}} = \text{Specificity}(M_i)$$

For theories differing merely in size and not in dimension, the very same intuition can be applied. For the $BIC$ we found:

$$P(M_i|M_0) \sim \frac{1}{\#\text{ theoretical possibilities}} = \text{Specificity}^\star(M_i)$$

Both specificity terms concern differences of theory size, albeit at different orders of magnitude.

**Simplicity, size, specificity**

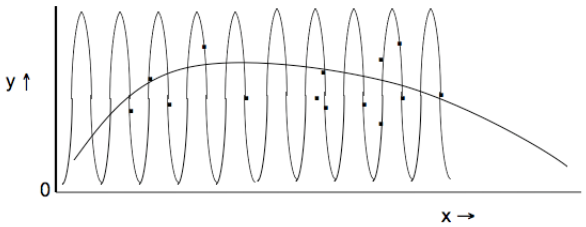
Similar refinements are available, or in the making, for the *AIC* and the *DIC*. We see that specificity concerns size and not just dimensionality.

- This answers to the problem of specificity. Probabilistic accounts of scientific inference can accommodate our preference for high specificity as well as high probability.
- It also throws new light on the case of Linda the bank teller: people prefer the feminist bank teller because it is more specific, or in another word simpler; it therefore has a higher marginal likelihood.

Recall that we can define theories $M_0$ and $M_1$ so that they are disjunct sets. So specificity can be expressed in high posterior probability.

# 5 The sub-family problem

Some things remain awkward about model selection: we can gerrymander the parameterisation of the theory in order to improve our fit while keeping the number of parameters low.




This is the problem of accommodation, or the sub-family problem in the context of curve-fitting: we can always come up with a smart parameterisation of the space of possible curves that renders a good fit at little cost.

**The sensitivity of the estimations**

The solution for this problem lies in testing the estimation for sensitivity to slight changes in the data: if for small changes to the data the estimations vary wildly, this tells against the family of functions used to fit the curve.

$$\begin{aligned} \log P(D_n|M_i) &= \log P(D_n|H_{\hat{\theta}} \cap M_i) - \frac{d_i}{2}\log(n) + \log P(H_{\hat{\theta}}|M_i) \\ &\quad +(d/2)\log(2\pi) - \frac{1}{2}\log|I| + O\left(\frac{1}{\sqrt{n}}\right), \end{aligned}$$

As it turns out, a measure of sensitivity is already present in the *BIC* approximation, as the so-called Fisher information $\log|I|$.

**Solving the problem?**
Parallel to existing tools employing minimum description length (*MDL*), one might develop adapted *IC*s that compare different parameterisations of the same model.

$$\begin{aligned} IC^{+} \quad \sim \quad & -\text{Fit}[P(D_n|H_{\hat{\theta}})] - \text{Specificity}\,[-d_i \log n] \\ & -\text{Specificity}^{\star}\,[P(M_i|M_0)] + \text{Sensitivity}[\log |I|] \end{aligned}$$

This solves the problem of accommodation if we indeed have an independent ground for the way we label and structure our data: the latter determines the sensitivity of the theory.

## 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.