university of groningen

NWO
Netherlands Organisation for Scientific Research

Simplicity seminar
October 17, 2011

One size does not fit all: a prior-adapted BIC

Jan-Willem Romeijn
University of Groningen

To appear in an edited volume on *Plurality in Statistics*

Model selection

Statistical inference concerns the comparison of statistical hypotheses from a given set of hypotheses, or model, M_i. For instance,

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

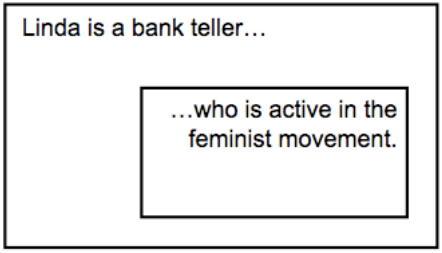
Model selection tools facilitate the comparison of statistical models on their ability to accommodate the data. The Bayesian information criterion (*BIC*) compares models by their approximated marginal likelihood:

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

When choosing among models by means of the *BIC*, we make a trade-off between simplicity and fit. They are expressed in the maximum likelihood term $P(D_n|H_{\hat{\theta}} \cap M_i)$ and in the dimensionality term $d_i \log(n)$ respectively.

Specificity vs probability

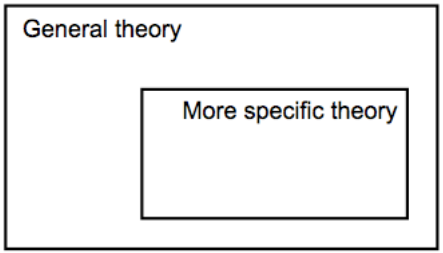
One driving intuition is that we prefer true scientific theories that allow for few possibilities over true ones that allow for many. But recall the paradox of Linda.



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

Models with constraints

In practical applications, scientists often compare models that only differ in terms of a set of constraints, and not in dimensionality. We can avoid trivialising the comparison by carefully defining the models involved.



In this talk, I avoid the issue by restricting attention to likelihoods, and I show how they can used for a meaningful comparison of the models.

Model selection for constrained models

We define a probability space $\langle W, \mathcal{F}, P \rangle$, with W a set of worlds w, and $\mathcal{F} = \mathcal{H} \times \mathcal{D}$, in which $\mathcal{D}$ is the sample space and $\mathcal{H}$ is an algebra based on a partition of hypotheses H_θ. We will consider a comparison between the following two models:

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

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

where M_0 is a so-called encompassing model, and M_1 constrained. In this setup the hypotheses H_θ for $\theta < \frac{1}{2}$ are included in both models. In this setup the models M_0 and M_1 partly overlap, so a comparison of posteriors is nonsensical.

BIC **vs marginal likelihood**

The BIC is an approximation of marginal likelihoods, and hence can be applied to such models. Say that the maximum likelihood hypothesis $H_{\hat{\theta}}$ is included in both models, for instance

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

Then the BIC of the two models is the same. The maximum likelihood terms are equal, as are the dimensions of the models and the number of observations:

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

But as argued below, the marginal likelihood of the two models differ. There is a discrepancy between the BIC and the marginal likelihood that it is supposed to approximate.

Contents

1 Truncated priors and marginal likelihoods **8**

2 The prior-adapted *BIC* **12**

3 A refinement of statistical simplicity **16**

4 The sub-family problem **19**

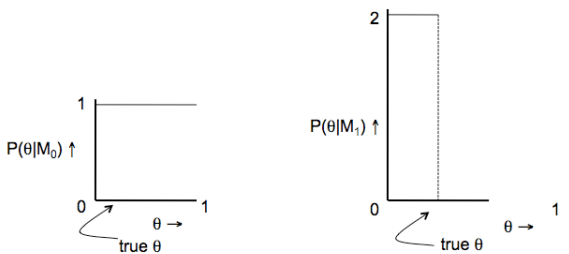
5 Conclusions **22**

1 Truncated priors and marginal likelihoods

The prior within the model M_1 is a truncated version of the prior within M_0:

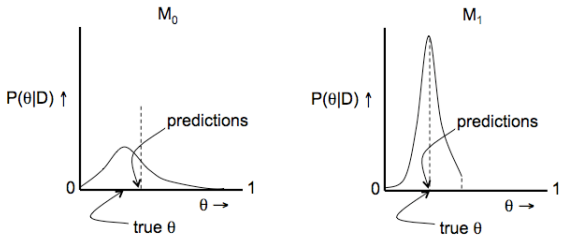
$$P(H_\theta|M_1)d\theta = \frac{1}{P(M_1|M_0)}P(H_\theta|M_0)d\theta.$$

In the example, the constrained model M_1 starts off with its prior probability closer to the maximum likelihood point than the encompassing model M_0.



Marginal likelihood for increasing sample size

For both models, with increasing sample size n the posterior probability accumulates around the maximum likelihood point. But the model M_1 begins with a head start.



As a result, the predictions from model M_1 are more accurate than those of model M_0, and its marginal likelihood is higher.

Limiting marginal likelihood ratio

Since all posterior probability collects around the maximum likelihood point, 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

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

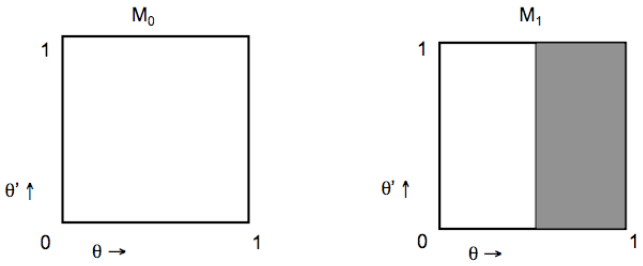
And if $\hat{\theta}$ lies outside M_1, we have

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

So the ratio of marginal likelihoods tends to the ratio of priors at the maximum likelihood point. The original *BIC* is supposed to replicate this limiting behaviour, but it does not.

Evaluation of the result

At this point we already know what should come out of the approximation. But there are independent reasons for aligning the BIC with this result on marginal likelihoods, to do with the interpretation of the simplicity term.



For the moment we just note that the term by which the marginal likelihoods differ is actually the size of the constrained model, in comparison to the encompassing model.

2 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:

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

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

Some more detail on the derivation

The original derivation employs the so-called Laplacian method for integrals on a Taylor expansion of the function $g(\theta) = \log P(H_\theta|M_i)P(D_n|H_\theta \cap M_i)$, 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 mode of the function $g(\theta)$. 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 $g(\theta)$.

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.

3 A refinement of statistical simplicity

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

BIC We choose the model with the largest approximated marginal likelihood.

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

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

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

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

Dimension and size as penalty

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 complexity, but interpreted as penalty for complexity afterwards and on independent grounds.

$$\text{Complexity}(M_i) = \#\text{ statistical possibilities} \sim d_i$$

The intuition is that complex models include more statistical possibilities and are therefore more versatile in adapting to observations. For models with truncated priors, the very same intuition can be applied to interpret the additional term in the *PBIC*:

$$\text{Complexity}(M_i) = \#\text{ statistical possibilities} \sim \frac{1}{P(M_i|M_0)}$$

Both penalty terms concern differences of models size, albeit at different orders of magnitude.

Simplicity, size, specificity

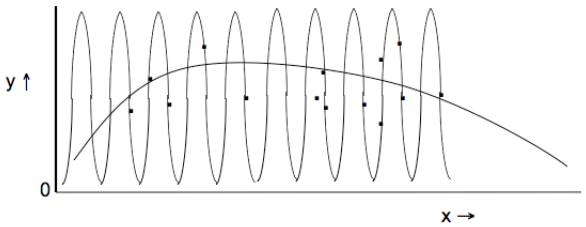
Similar refinements are available, or in the making, for the *AIC* and the *DIC*. There seems to be a basis for adapting the concept of statistical simplicity: it is about model size, not just about dimensionality.

- This role of size ties in with a well-known problem for Bayesian inference, namely its failure to accommodate that science strives 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, and therefore has a higher marginal likelihood.

Recall that we can define models M_0 and M_1 so that they are disjunct sets. A fully Bayesian solution of the Linda case, concerning posterior probabilities, may still be possible.

4 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 may lie 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 helps if we indeed have an independent ground for the way we label and structure our data: the latter determines the sensitivity of the theory. But then again, the order of the terms seems wrong, and we can still gerrymander a family of distributions ex post.

5 Conclusions

Some general claims I am happy to defend:

- The *PBIC* can replace the original *BIC* at no extra cost, thereby bringing the comparison of constrained models within the scope of model selection tools.
- The gain is not so much that we can apply the *PBIC* to such cases: we know the results of such comparisons already. The gain is rather that the *PBIC* motivates a refinement of the notion of simplicity.
- The new notion of simplicity runs parallel to that of specificity.
- Perhaps there is, after all, a probabilistic account for our preference towards logically stronger theories: if they are right, they have higher marginal likelihood.

And future work. . .

I think the notion of simplicity at work in model selection can be supplemented with a number of further features:

- The parameterisation of the model at the maximum likelihood point comes back in the $\log|I|$ term. One might argue that some *PBIC'* can thus compare different parameterisations of the same model as well as various truncated models.
- More generally, the marginal likelihoods depend on the full prior over the models. Studying the behaviour of the marginals in the short and medium term leads to some surprising results. Approach to the limit should perhaps play a more important role in model selection.
- The sub-family problem keeps bothering us: we can still gerrymander the dimensionality of the model d_i and the number of parameters n. Does this matter at all? Is there a principled way of fixing these numbers?

Thank you

The slides for this talk will be available at Kevin’s course website and at

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

For comments and questions, email j.w.romeijn@rug.nl.