Discussion of "Cosmological Bayesian Model Selection: Recent Advances and Open Challenges"

David A. van Dyk

1 Introduction

Doctor Trotta is to be congratulated for his lucid summary of recent advances in Bayesian fitting of cosmological models and of the outstanding challenges in the more difficult problem of model selection. This situation is not unique to cosmology. Differences among statistical paradigms such as frequency-based or Bayesian methods are generally much more pronounced in model checking and selection than in fitting. Indeed no consensus exists even among Bayesians or among frequentists as to the best way forward in model selection. As such this remains an active area of statistical research where the experience of cosmologists may lead to insight with impact on more general statistical methodology. It is also a subtle area where one must be wary of all-purpose solutions. As Doctor Trotta points out, model selection in cosmology is not confined to nested models (e.g., adding "extra parameters in the ΛCDM beyond the vanilla ones") but includes the more technically challenging case of comparing non-nested models "whose parameter spaces are largely or completely disjoint". Such seemingly innocuous differences may be highly consequential and lead to subtle technical issues. I hope to illustrate some of the subtleties involved and the advantages of a mixed approach that considers and compares various methods in the context of a specific model selection problem.

2 Methods for Model Selection and Checking

Model checking problems often begin with a default or presumed model,

Null Hypothesis: E.g., the Universe is "Flat".

David A. van Dyk
Statistics Section, Department of Mathematics, Imperial College London, SW7 2AZ, UK
e-mail: dvandyk@imperial.ac.uk

The scientist asks whether the model is consistent with the data or if it is plausible that the data were generated under the model. If not, we aim to characterize the inconsistency, improve the model, and recheck the improved model. In principle this cycle of model improvement can be iterated, perhaps with the acquisition of new data, until a satisfactory model is obtained.

We may also have a model that we suspect or hope is better than the null model,

Alternative Hypothesis: E.g., the Universe is "Hyperbolic".

With a competing model in hand, we typically aim to decide between or weigh the evidence for the two (or more) models. These procedures are known as *model selection* and *model comparison*. In some situations we may wish to assume the null hypothesis until we have substantial evidence it is implausible. This is analogous to assuming a defendant is innocent, until proven guilty in a court of law. Similarly we may not wish to overturn a long standing standard model without truly solid evidence. In other situations we may not have any particular reason to favor one model over another and may wish to simply weigh the relative evidence for each.

These are surprisingly subtle problems and despite decades of research, discussion, and sometimes heated arguments, little consensus exists among statisticians as how to best tackle them. This is especially concerning because competing methods may lead to very different conclusions. Part of the difficulty is that model selection is somewhat ill-posed. Statisticians view models as parsimonious mathematical summaries of complex phenomena. They are not meant to capture the full complexity of that which they summarize. As such different models can be viewed as approximations with various tradeoffs between simplicity and detail, no one of which may be 'true' or even better than the others; they are simply different. Nonetheless we may wish to investigate how a particular model differs from reality (i.e., model checking) or which of a set of models better approximates a particular aspect of reality (i.e., model comparison). Remembering that models are not meant to be perfect, however, it is no surprise that there is no completely general theory for model selection nor is there always a clear cut answer to model selection problems. Model checking, comparison, and selection are nuanced endeavors into the shades of grey.

Frequency-Based Methods. The standard frequency based method begins with a statement of a null and an alternative hypothesis,

H_0: E.g., the Universe is Flat: $\Omega_\kappa = 0$, and
H_1: E.g., the Universe is not Flat: $\Omega_\kappa \neq 0$,

and computes a test statistic, T, with known distribution under H_0. A threshold, $T^\star$ is then computed as, e.g., the smallest value such that $\text{Prob}(T > T^\star | \Omega_\kappa = 0, \text{other parameters}) \leq \alpha$,, where α is the *significance level* of the test. Under the assumption that H_0 holds, T is greater than $T^\star$ with probability less than α. This is an infrequent occurrence if α is small. Thus, we typically choose a small value of α and if we observe $T > T^\star$ conclude that there is sufficient evidence to declare H_0 implausible. In this example, we would conclude that the Universe is not flat.

This paradigm is generally advocated on the basis of its control of the probability of *false positive*. That is, we will wrongly conclude that H_0 is implausible with

probability less than α, when H_0 actually holds. On the other hand the method offers no characterization of the strength of evidence, a task left to the notorious p-value, see below. Another important sticking point lies in the derivation of a test statistic with known distribution under H_0. This can be a difficult if not impossible task in complex models that have numerous unknown parameters.

Bayesian Methods. Because Bayesian methods treat parameters as random quantities there is no problem in principle with unknown parameters under either H_0 or H_A. In particular the *prior predictive distribution*, given in Trotta's equation (2) specifies how likely the data, d, is under model $i \in \{0,1\}$. In a Bayesian paradigm the model consists of a specification of both the likelihood and the prior distribution and both are compared together. The typical method for comparing two models involves the *Bayes Factor*, or the *posterior probability of* H_0. Unlike standard frequency-based methods, both the Bayes Factor and $p(H_0|\mathscr{M})$ treat H_0 and H_1 essentially symmetrically. There is no need to treat H_0 as the default or *a priori* assumed model.

A typical criticism of Bayesian methods in general is their requirement that one specifies a prior distribution. Of course, when informative prior information is available, Bayesian methods offer a principled method of combining this information with the current data. In many situations, these concerns are of little practical importance because the posterior distribution, parameter estimates, error bars, and interval estimates are quite insensitive to the choice of prior distribution. Unfortunately, the same is not true of prior predictive distributions and Bayes factors which can be quite sensitive to the choice of prior distribution. As an example, suppose we observe a Gaussian variable with mean μ and variance one, use a Gaussian prior distribution on μ with mean zero and variance τ^2, and are interested in testing $H_0 : \mu = 0$ against $H_1 : \mu \neq 0$. Using (2) we can compute the prior predictive distribution of d which is a Gaussian distribution with mean zero and variance $1+\tau^2$ and is plotted in the lefthand panel of Figure 1 for several values of τ^2. The prior predictive distribution is highly dependent on the prior distribution and $p(d|\mathscr{M})$ can be made arbitrarily small for any value of d by choosing τ^2 large enough. The righthand panel of Figure 1 illustrates the effect on the log Bayes factor, which varies from indifference between H_0 and H_1 to strong support for H_1 to strong support for H_0 as τ^2 increases.

Reflecting on Figure 1, it is clear that we must think carefully about our choice of prior distribution and it is critical that the prior distribution accurately summarizes available prior information. The typical strategy of using "non-informative" prior distributions with large variances clearly effects the Bayes factor. In fact "improper" prior distributions (e.g., with infinite variance) result in improper prior predictive distributions and undefined Bayes factors. There is no simple default prior distribution available when computing Bayes Factors. This is especially problematic when the parameter space is large and in particular when the H_A and H_0 are either not nested or the dimension of the parameter space under H_A is much larger than that under H_0. Specifying subjective prior distributions in large multivariate spaces involves careful consideration of the correlations and likely relationships among the parameters. In model selection problems, the hypothesized models may be rather speculative and little prior information about the values of the parameters may be

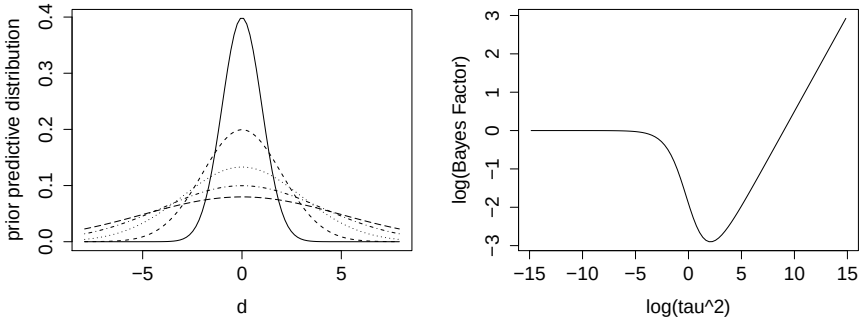

Fig. 1 The dependence of the prior predictive distribution and the Bayes factor on the choice of prior distribution. The lefthand panel plots the prior predictive distribution for the Gaussian example describe in the text with five choices of the prior distribution. The righthand plot shows the effect of the prior variance, τ^2, on the Bayes factor. Results are highly dependent on the prior distribution and the prior must be chosen with care to accurately represent available prior information.

forthcoming. Thus, we may have little information for a the choice of prior and the prior may heavily influence results. In such situations, it is absolutely critical that the choice of prior distribution be reported along with the Bayes factor.

I worry about the application of Bayes factors in cosmology, just as I generally worry about their use by scientists and statistician alike. Doctor Trotta mentions the "Astronomer's Prior" ($\Omega_\kappa \sim \text{Unif}(-1,1)$) and the "Curvature Scale Prior" ($\log|\Omega_\kappa| \sim \text{Unif}(-5,0)$). In the inflationary model he notes that "little if anything is known *a priori* about the free parameter Ψ..." and that "non-linear transformations ... in general change ... the model comparison results." Understandably convenient prior distributions are used in the absence of well quantified substantive prior knowledge. Unfortunately, Bayes factors based on such priors lead to questionable results.

P-values. In the context of frequency based hypothesis testing, the *p-value* is often reported to quantify the degree of evidence, p-value $=$ prob$(T > T^\star|\Omega_\kappa = 0, \text{other parameters})$. Although they are endemic in data analysis, there is a large literature on the difficulties of interpreting p-values, especially when testing precise null hypotheses (e.g., Berger and Delampady, 1987). When compared to Bayes factors and the posterior probability of H_0, p-values *vastly overstate the evidence for* H_1, even when compared to Bayesian methods that use the prior most favorable to H_1 from a large class of priors. This is because p-values are computed given data *as extreme or more extreme* than d. This is *much stronger evidence* for H_1 than d. (In some cases p-values agree with Bayesian measures computed with "as extreme or more extreme' data (Berger and Sellke, 1987)). P-values cannot be simply recalibrated to agree with Bayesian measures because the magnitude of the discrepancy depends on the sample size, the model, and the precision of H_0. In short p-values should be avoided because they are difficult to interpret, have questionable frequency properties, and bias inference in the direction of false discovery.

3 The Bottom Line

There are other statistical paradigms and hybrid methods that aim to evaluate models and decide between them, e.g., posterior-predictive-p-values (Gelman *et al.*, 1996), conditional error probabilities (Berger *et al.*, 1997), decision theory (e.g., Casella and Berger, 1990; van Dyk, 2011), etc. Still there are no silver bullets. Most statisticians agree that model selection should be rephrased into model fitting problems whenever possible. In the case of nested models, it is often possible to fit the larger model and report interval for the parameters that are free in the larger but not the smaller model. The added value of the larger model can be assessed by examining the likely values of these parameters. This avoids the problem of model selection, but may not adequately address the scientific question. In such cases, I agree with Doctor Trotta that Bayesian methods are most promising. Despite their dependence on the choice of prior distribution, Bayes factors represent a principled probability-based assessment of the relative evidence for H_0 and H_1. Unlike p-vales, they aim to answer the right questions and like other Bayesian methods, they have no problem with nuisance parameters. Various strategies exist for mitigating their dependence on the prior distribution. For example, Berger and Delampady (1987) recommends optimizing the Bayes factor over a class of priors and Berger and Pericchi (2001) review methods that use a subset of the data to construct an informative prior distribution and the remainder to compute the Bayes factor. Overall, I view Bayes factors as the most promising method for model selection. Clearly care must be taken when selecting prior distributions and sensitivity analyses must be conducted. But at a fundamental level Bayes factors answer the questions of most interest to scientists.

References

Berger, J. O., Boukai, B., and Wang, Y. (1997). Unified frequentist and Bayesian testing of a precise hypothesis (with discussion). *Statistical Science* **12**, 133–160.

Berger, J. O. and Delampady, M. (1987). Testing precise hypotheses (with discusion). *Statistical Science* **2**, 317–352.

Berger, J. O. and Pericchi, L. R. (2001). Objective Bayesian methods for model selection: Introduction and comparision (with discussion). In *Model Selection* (Editor: P. Lahiri), 135–207. IMS, Beachwood, Ohio.

Berger, J. O. and Sellke, T. (1987). Testing a point null hypothesis: The irreconcilability of P-values and evidence (with discussion). *Journal of the American Statistical Association* **82**, 112–139.

Casella, G. and Berger, R. L. (1990). *Statistical Inference*. Wadsworth & Brooks/Cole, Pacific Grove, CA.

Gelman, A., Meng, X.-L., and Stern, H. (1996). Posterior predictive assessment of model fitness (with discussion). *Statistica Sinica* **6**, 733–807.

van Dyk, D. A. (2011). Setting Limits, Computing Intervals, and Detiction. In *Phystat 2011 Proc.* (Eds: H. Prosper and L. Lyons), in press. CERN Yellow Report.