BAYESIAN MODEL COMPARISON

ICIC DATA ANALYSIS WORKSHOP, 11-13 SEPT 2013

DR ROBERTO TROTTA

This is a summary of the notes for the 3rd day of the "ICIC Data Analysis Workshop" held in Sept 2013 at Imperial College London.

Notice that the list of references has been kept very minimal. Please refer to [13] for a far more complete overview of the literature. Useful reference textbooks are Refs. [1, 2, 3, 4, 5, 6, 7, 8]; applications to cosmology and astrophysics can be found in e.g. [9, 10, 11, 12, 13].

Comments, suggestions and corrections are very welcome! `r.trotta@imperial.ac.uk`.

CONTENTS

Guildenstern — The law of probability, as it has been oddly asserted, is something to do with the proposition that [...], if I have got this right, if six monkeys were thrown up in the air for long enough they would land on their tails about as often as they would land on their...

Rosencratz — Heads.

Tom Stoppard, *Rosencratz and Guildenstern are dead* (1966)

1. FOUNDATIONAL ASPECTS

- Any physical theory must be validated against observations. The cycle theory → prediction → observation → theory forms the pillar of the scientific process.
- The comparison between theory and observation is the key phase when statistical methods are needed: *how do we learn about the world from a collection of noisy observations?*
- Statistics is at the heart of the scientific process, not merely an optional nuisance. Ernest Rutherford is reported to have said: "If you need statistics, you did the wrong experiment". This completely misses the point: if you do not need statistics, it's because you are doing the wrong kind of physics! Five reasons why you need statistics:
  - (i) The complexity of the modelling of both our theories and observations will always increase, thus requiring correspondingly more refined statistical and data analysis skills. In fact, the scientific return of the next generation of surveys will be limited by the level of sophistication and efficiency of our inference tools.
  - (ii) The discovery zone for new physics is when a potentially new effect is seen at the 3–4 σ level. This is when tantalizing suggestions for an effect start to accumulate but there is no firm evidence yet. In this potential discovery region a careful application of statistics can make the difference between claiming or missing a new discovery.
  - (iii) If you are a theoretician, you do not want to waste your time trying to explain an effect that is not there in the first place. A better appreciation of the interpretation of statistical statements might help in identifying robust claims from spurious ones.
  - (iv) Limited resources mean that we need to focus our efforts on the most promising avenues. Experiment forecast and optimization will increasingly become prominent as we need to use all of our current knowledge (*and* the associated uncertainty) to identify the observations and strategies that are likely to give the highest scientific return in a given field.
  - (v) Sometimes there will be no better data! This is the case for the many problems associated with cosmic variance limited measurements on large scales, for example in the cosmic background radiation, where the small number of independent directions on the sky makes it impossible to reduce the error below a certain level.

1.1. **Revision – Bayes theorem and notation.**

- Bayes theorem, Eq. (4), encapsulates the notion of *probability as degree of belief.* The Bayesian outlook on probability is more general than the frequentist one, as the former can deal with unrepeatable situations that the latter cannot address. We begin with some simple definitions and consequences.
- The *joint probability* of A and B is the probability of A and B happening together, and is denoted by $P(A,B)$.

  The *conditional probability* of A given B is the probability of A happening given that B has happened, and is denoted by $P(A|B)$.
- The sum rule:

$$P(A) + P(\overline{A}) = 1, \tag{1}$$

where $\overline{A}$ denotes the proposition "not A".

- The product rule:

(2) $$P(A,B) = P(A|B)P(B).$$

By inverting the order of A and B we obtain that

(3) $$P(B,A) = P(B|A)P(A)$$

and because $P(A,B) = P(B,A)$, we obtain, by equating Eqs. (2) and (3):

- *Bayes theorem*

(4) $$P(A|B) = \frac{P(B|A)P(A)}{P(B)}.$$

- The marginalisation rule (follows from the two rules above):

(5) $$P(A) = P(A,B_1) + P(A,B_2) + \cdots = \sum_i P(A,B_i) = \sum_i P(A|B_i)P(B_i),$$

where the sum is over all possible outcomes for proposition B.

- Two propositions (or events) are said to be *independent* if and only if

(6) $$P(A,B) = P(A)P(B).$$

- We replace in Bayes theorem, Eq. (4), $A \to \theta$ (the parameters) and $B \to d$ (the observed data, or samples), obtaining

(7) $$P(\theta|d) = \frac{P(d|\theta)P(\theta)}{P(d)}.$$

On the LHS, $P(\theta|d)$ is *the posterior probability for* θ (or "posterior" for short), and it represents our degree of belief about the value of θ after we have seen the data d.

On the RHS, $P(d|\theta) = \mathcal{L}(\theta)$ is the likelihood we already encountered. It is the probability of the data given a certain value of the parameters. The quantity $P(\theta)$ is *the prior probability distribution* (or "prior" for short). It representes our degree of belief in the value of θ before we see the data. This is an essential ingredient of Bayesian statistics. In the denominator, $P(d)$ is a normalizing constant (often called "the evidence"), which ensures that the posterior is normalized to unity:

(8) $$P(d) = \int d\theta P(d|\theta)P(\theta).$$

The evidence is important for Bayesian model selection (see section 2).

- **Interpretation:** Bayes theorem relates the posterior probability for θ (i.e., what we know about the parameters after seeing the data) to the likelihood. It can be thought of as a general rule to update our knowledge about a quantity (here, θ) from the prior to the posterior. A result known as Cox theorem shows that Bayes theorem is the unique generalization of boolean algebra in the presence of uncertainty.
- Remember that in general $P(\theta|d) \neq P(d|\theta)$ (see ex. of pregnant woman!), i.e. the posterior and the likelihood are two different quantities with different meaning!

2. Bayesian model selection

2.1. The three levels of inference.

- It is convenient to divide Bayesian inference in three different levels:
  (i) **Level 1:** We have chosen a model $\mathcal{M}_0$, assumed true, and we want to learn about its parameters, θ_0. E.g.: we assume ΛCDM to be the true model for the Universe and try to constrain its parameters. This is the usual parameter inference step.
  (ii) **Level 2:** We have a series of alternative models on the table ($\mathcal{M}_1, \mathcal{M}_2, \ldots$) and we want to determine which of those is in best agreement with the data. This is a problem of model selection, or model criticism. For example, we might want to decide whether a dark energy equation of state $w = -1$ is a sufficient description of the available observations or whether we need an evolving dark energy model, $w = w(z)$.

(iii) **Level 3:** Of the N models considered in Level 2, there is no clear "best" model. We want to report inferences on parameters that account for this model uncertainty. This is the subject of Bayesian model averaging (not covered in these lectures). For example, we want to determine Ω_m independently of the assumed dark energy model.

- The Frequentist approach to model criticism is in the form of hypothesis testing (e.g., "chi-squared-per-degree-of-freedom" type of tests). One ends up rejecting (or not) a null hypothesis H_0 based on the p-value, i.e., the probability of getting data as extreme or more extreme than what has been observed if one assumes that H_0 is true. Notice that this is *not* the probability for the hypothesis! Classical hypothesis testing assumes the hypothesis to be true and determines how unlikely are our observations given this assumption. Studying Ref. [20] is highly recommended.
- The Bayesian approach takes the view that there is no point in rejecting a model unless there are specific alternatives available: it takes therefore the form of model *comparison*. The key quantity for model comparison is the Bayesian evidence, which automatically implements a quantitative version of Occam's razor (i.e., the notion that simpler models ought to be preferred if they can explain the data sufficiently well).

2.2. The Bayesian evidence.

- The evaluation of a model's performance in the light of the data is based on the *Bayesian evidence*. As seen above, this is the normalization integral on the right–hand–side of Bayes' theorem, Eq. (8), which we rewrite here for a continuous parameter space $\Omega_{\mathcal{M}}$ and conditioning explicitly on the model under consideration, $\mathcal{M}$:

$$p(d|\mathcal{M}) \equiv \int_{\Omega_{\mathcal{M}}} p(d|\theta,\mathcal{M})p(\theta|\mathcal{M})\mathrm{d}\theta \quad \text{(Bayesian evidence)}. \tag{9}$$

- The Bayesian evidence is the average of the likelihood under the prior for a specific model choice. From the evidence, the model posterior probability given the data is obtained by using Bayes' Theorem to invert the order of conditioning:

$$p(\mathcal{M}|d) \propto p(\mathcal{M})p(d|\mathcal{M}), \tag{10}$$

where we have dropped an irrelevant normalization constant that depends only on the data and $p(\mathcal{M})$ is the prior probability assigned to the model itself. Usually this is taken to be non–committal and equal to $1/N_m$ if one considers N_m different models.

- When comparing two models, $\mathcal{M}_0$ versus $\mathcal{M}_1$, one is interested in the ratio of the posterior probabilities, or *posterior odds*, given by

$$\frac{p(\mathcal{M}_0|d)}{p(\mathcal{M}_1|d)} = B_{01}\frac{p(\mathcal{M}_0)}{p(\mathcal{M}_1)}. \tag{11}$$

- The *Bayes factor* B_{01} is the ratio of the models' evidences:

$$B_{01} \equiv \frac{p(d|\mathcal{M}_0)}{p(d|\mathcal{M}_1)} \quad \text{(Bayes factor)}. \tag{12}$$

A value $B_{01} > (<)\ 1$ represents an increase (decrease) of the support in favour of model 0 versus model 1 given the observed data (see [16] for more details on Bayes factors).

- Bayes factors are usually interpreted against the Jeffreys' scale [2] for the strength of evidence, given in Table 1. This is an empirically calibrated scale, with thresholds at values of the odds of about 3 : 1, 12 : 1 and 150 : 1, representing weak, moderate and strong evidence, respectively.

2.3. The Occam's razor effect.

- **Example (nested models)**: Consider two competing models: $\mathcal{M}_0$ predicting that a quantity $\theta = 0$ with no free parameters, and $\mathcal{M}_1$ which assigns θ a Gaussian prior distribution with 0 mean and variance Σ^2. Assume we perform a measurement of θ described by a normal likelihood of standard

$\|\ln B_{01}\|$	Odds	Probability	Strength of evidence
< 1.0	$\lesssim 3:1$	< 0.750	Inconclusive
1.0	$\sim 3:1$	0.750	Weak evidence
2.5	$\sim 12:1$	0.923	Moderate evidence
5.0	$\sim 150:1$	0.993	Strong evidence

TABLE 1. Empirical scale for evaluating the strength of evidence when comparing two models, $\mathcal{M}_0$ versus $\mathcal{M}_1$ (so–called "Jeffreys' scale"). Threshold values are empirically set, and they occur for values of the logarithm of the Bayes factor of $|\ln B_{01}| = 1.0$, 2.5 and 5.0. The right–most column gives our convention for denoting the different levels of evidence above these thresholds. The probability column refers to the posterior probability of the favoured model, assuming non–committal priors on the two competing models, i.e., $p(\mathcal{M}_0) = p(\mathcal{M}_1) = 1/2$ and that the two models exhaust the model space, $p(\mathcal{M}_0|d) + p(\mathcal{M}_1|d) = 1$.

deviation σ, and with the maximum likelihood value lying λ standard deviations away from 0, i.e. $|\theta_{\max}/\sigma| = \lambda$. Then the Bayes factor between the two models is given by, from Eq. (12)

$$B_{01} = \sqrt{1 + (\sigma/\Sigma)^{-2}} \exp\left(-\frac{\lambda^2}{2(1 + (\sigma/\Sigma)^2)}\right). \tag{13}$$

For $\lambda \gg 1$, corresponding to a detection of the new parameter at many sigma, the exponential term dominates and $B_{01} \ll 1$, favouring the more complex model with a non–zero extra parameter, in agreement with the usual conclusion. But if $\lambda \lesssim 1$ and $\sigma/\Sigma \ll 1$ (i.e., the likelihood is much more sharply peaked than the prior and in the vicinity of 0), then the prediction of the simpler model that $\theta = 0$ has been confirmed. This leads to the Bayes factor being dominated by the Occam's razor term, and $B_{01} \approx \Sigma/\sigma$, i.e. evidence accumulates in favour of the simpler model proportionally to the volume of "wasted" parameter space. If however $\sigma/\Sigma \gg 1$ then the likelihood is less informative than the prior and $B_{01} \to 1$, i.e. the data have not changed our relative belief in the two models.

- In the above example, if the data are informative with respect to the prior on the extra parameter (i.e., for $\sigma/\Sigma \ll 1$) the logarithm of the Bayes factor is given approximately by

$$\ln B_{01} \approx \ln(\Sigma/\sigma) - \lambda^2/2, \tag{14}$$

where as before λ gives the number of sigma away from a null result (the "significance" of the measurement). The first term on the right–hand–side is approximately the logarithm of the ratio of the prior to posterior volume. We can interpret it as the information content of the data, as it gives the factor by which the parameter space has been reduced in going from the prior to the posterior. This term is positive for informative data, i.e. if the likelihood is more sharply peaked than the prior. The second term is always negative, and it favours the more complex model if the measurement gives a result many sigma away from the prediction of the simpler model (i.e., for $\lambda \gg 0$). We are free to measure the information content in base–10 logarithm (as this quantity is closer to our intuition, being the order of magnitude of our information increase), and we define the quantity $I_{10} \equiv \log_{10}(\Sigma/\sigma)$. Figure 1 shows contours of $|\ln B_{01}| =$ const for const $= 1.0, 2.5, 5.0$ in the (I_{10}, λ) plane, as computed from Eq. (14). The contours delimit significative levels for the strength of evidence, according to the Jeffreys' scale (Table 1). For moderately informative data ($I_{10} \approx 1-2$) the measured mean has to lie at least about 4σ away from 0 in order to robustly disfavor the simpler model (i.e., $\lambda \gtrsim 4$). Conversely, for $\lambda \lesssim 3$ highly informative data ($I_{10} \gtrsim 2$) do favor the conclusion that the extra parameter is indeed 0. In general, a large information content favors the simpler model, because Occam's razor penalizes the large volume of "wasted" parameter space of the extended model.

- An useful properties of Figure 1 is that the impact of a change of prior can be easily quantified. A different choice of prior width (i.e., Σ) amounts to a *horizontal shift* across Figure 1, at least as long as $I_{10} > 0$ (i.e., the posterior is dominated by the likelihood). Picking more restrictive priors (reflecting more predictive theoretical models) corresponds to shifting the result of the model comparison

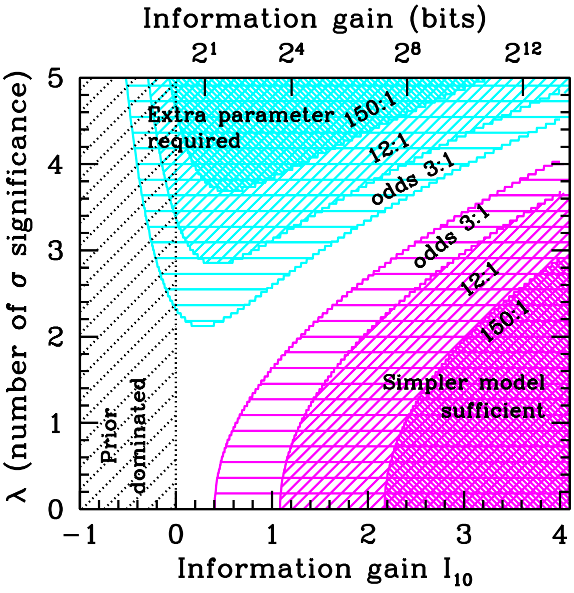

FIGURE 1. Illustration of Bayesian model comparison for two nested models, where the more complex model has one extra parameter. The outcome of the model comparison depends both on the information content of the data with respect to the *a priori* available parameter space, I_{10} (horizontal axis) and on the quality of fit (vertical axis, λ, which gives the number of sigma significance of the measurement for the extra parameter). Adapted from [14].

to the left of Figure 1, returning an inconclusive result (white region) or a prior–dominated outcome (hatched region). Notice that results in the 2–3 sigma range, which are fairly typical in cosmology, can only support the more complex model in a very mild way at best (odds of 3 : 1 at best), while actually being most of the time either inconclusive or in favour of the simpler hypothesis (pink shaded region in the bottom right corner).

- Notice that Bayesian model comparison is usually *conservative* when it comes to admitting a new quantity in our model, even in the case when the prior width is chosen "incorrectly" (whatever that means!). In general the result of the model comparison will eventually override the "wrong" prior choice (although it might take a long time to do so), exactly as it happens for parameter inference.
- Bayesian model selection does not penalize parameters which are unconstrained by the data. This is easily seen from Eq. (14): if a parameter is unconstrained, its posterior width σ is approximately equal to the prior width, Σ, and the Occam's razor penalty term goes to zero. In such a case, consideration of the Bayesian model complexity might help in judging model performance, see [25] for details.

2.4. Computation of the evidence.

- **Nested sampling.** A powerful and efficient alternative to classical MCMC methods has emerged in the last few years in the form of the so–called "nested sampling" algorithm, invented by John Skilling [18]. Although the original motivation for nested sampling was to compute the evidence integral of Eq. (9), the recent development of the multi–modal nested sampling technique [17] has delivered an extremely powerful and versatile algorithm which has been demonstrated to be able to deal with extremely complex likelihood surfaces, see Fig. 2.

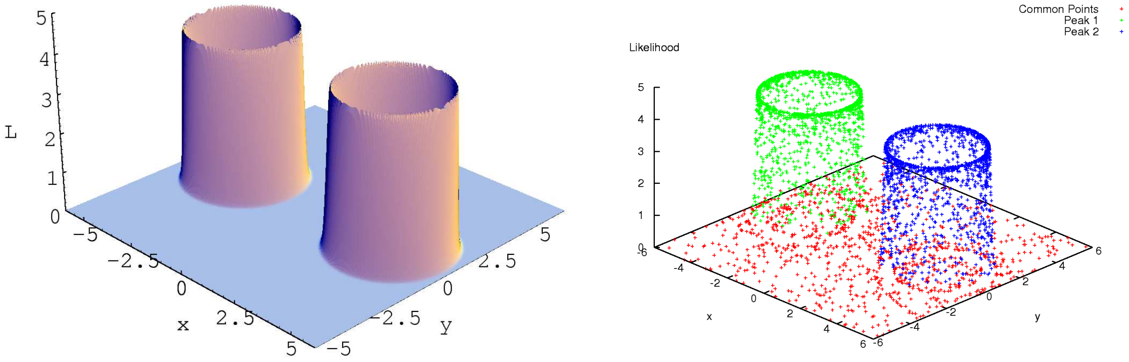

FIGURE 2. Example of posterior reconstruction using Nested Sampling. Left panel: target likelihood in a 2D parameter space (x,y). Right panel: reconstructed posterior (with flat priors) using Nested Sampling. From Ref. [17].

- The gist of nested sampling is that the multi–dimensional evidence integral is recast into a one–dimensional integral, by defining the prior volume X as $\mathrm{d}X \equiv p(\theta|\mathcal{M})\mathrm{d}\theta$ so that

$$X(\lambda) = \int_{\mathcal{L}(\theta)>\lambda} p(\theta|\mathcal{M})\mathrm{d}\theta \tag{15}$$

  where $\mathcal{L}(\theta) \equiv p(d|\theta,\mathcal{M})$ is the likelihood function and the integral is over the parameter space enclosed by the iso–likelihood contour $\mathcal{L}(\theta) = \lambda$. So $X(\lambda)$ gives the volume of parameter space above a certain level λ of the likelihood.
- The Bayesian evidence, Eq. (9), can be written as

$$p(d|\mathcal{M}) = \int_0^1 \mathcal{L}(X)\mathrm{d}X, \tag{16}$$

  where $\mathcal{L}(X)$ is the inverse of Eq. (15). Samples from $\mathcal{L}(X)$ can be obtained by drawing uniformly samples from the likelihood volume within the iso–contour surface defined by λ. This is the difficult part of the algorithm.
- Finally, the 1–dimensional integral of Eq. (16) can be obtained by simple quadrature, thus

$$p(d|\mathcal{M}) \approx \sum_i \mathcal{L}(X_i)W_i, \tag{17}$$

  where the weights are $W_i = \frac{1}{2}(X_{i-1} - X_{i+1})$, see [18, 19] for details[1].

▸ **Thermodyamic integration.** The numerical method of choice until recently has been thermodynamic integration, which computes the evidence integral by definining

$$E(\mu) \equiv \int_{\Omega_{\mathcal{M}}} \mathcal{L}(\theta)^{\mu} p(\theta|\mathcal{M})\mathrm{d}\theta, \tag{18}$$

where μ is an annealing parameter and $\mathcal{L}(\theta) \equiv p(d|\theta,\mathcal{M})$. Obviously the desired evidence correponds to $E(1)$. One starts by performing a standard MCMC sampling with $\mu = 0$ (i.e., sampling from the prior), then gradually increases μ to 1 according to some annealing schedule. The log of the evidence is then given by

$$\ln E(1) = \ln E(0) + \int_0^1 \frac{d\ln E}{d\mu}\mathrm{d}\mu = \int_0^1 \langle \ln \mathcal{L}\rangle_{\mu}\mathrm{d}\mu, \tag{19}$$

[1] Publicly available software implementing nested sampling can be found at cosmonest.org and http://www.mrao.cam.ac.uk/software/cosmoclust/. The latest release of the SUSY constraints package SuperBayeS also implements the MultiNest algorithm, see http://superbayes.org.

where the average log-likelihood is taken over the posterior with annealing parameter μ, i.e.

(20) $$\langle \ln \mathcal{L} \rangle_\mu = \frac{\int_{\Omega_\mathcal{M}} (\ln \mathcal{L}) \mathcal{L}(\theta)^\mu p(\theta|\mathcal{M}) \mathrm{d}\theta}{\int_{\Omega_\mathcal{M}} \mathcal{L}(\theta)^\mu p(\theta|\mathcal{M}) \mathrm{d}\theta}.$$

The drawback is that the end result might depend on the annealing schedule used and that typically this methods takes 10 times as many likelihood evaluations as parameter estimation.

- **Laplace approximation.** An approximation to the Bayesian evidence can be obtained when the likelihood function is unimodal and approximately Gaussian in the parameters. Expanding the likelihood around its peak to second order one obtains the Laplace approximation

(21) $$p(d|\theta,\mathcal{M}) \approx \mathcal{L}_{\mathrm{max}} \exp\left[-\frac{1}{2}(\theta-\theta_{\mathrm{max}})^t L(\theta-\theta_{\mathrm{max}})\right],$$

where θ_{max} is the maximum–likelihood point, $\mathcal{L}_{\mathrm{max}}$ the maximum likelihood value and L the likelihood Fisher matrix (which is the inverse of the covariance matrix for the parameters). Assuming as a prior a multinormal Gaussian distribution with zero mean and Fisher information matrix P one obtains for the evidence, Eq. (9)

(22) $$p(d|\mathcal{M}) = \mathcal{L}_{\mathrm{max}} \frac{|F|^{-1/2}}{|P|^{-1/2}} \exp\left[-\frac{1}{2}(\theta_{\mathrm{max}}{}^t L \theta_{\mathrm{max}} - \overline{\theta}^t F \overline{\theta})\right],$$

where the posterior Fisher matrix is $F = L + P$ and the posterior mean is given by $\overline{\theta} = F^{-1} L \theta_{\mathrm{max}}$.

- **The Savage-Dickey density ratio.** A useful approximation to the Bayes factor, Eq. (12), is available for situations in which the models being compared are *nested* into each other, i.e. the more complex model ($\mathcal{M}_1$) reduces to the original model ($\mathcal{M}_0$) for specific values of the new parameters. This is a fairly common scenario in cosmology, where one wishes to evaluate whether the inclusion of the new parameters is supported by the data (e.g., do we need isocurvature contributions to the initial conditions for cosmological perturbations, or whether a curvature term in Einstein's equation is needed, or whether a non–scale invariant distribution of the primordial fluctuation is preferred). Writing for the extended model parameters $\theta = (\phi,\psi)$, where the simpler model $\mathcal{M}_0$ is obtained by setting $\psi = 0$, and assuming further that the prior is separable (which is usually the case in cosmology), i.e. that

(23) $$p(\phi,\psi|\mathcal{M}_1) = p(\psi|\mathcal{M}_1) p(\phi|\mathcal{M}_0),$$

the Bayes factor can be written in all generality as

(24) $$B_{01} = \left.\frac{p(\psi|d,\mathcal{M}_1)}{p(\psi|\mathcal{M}_1)}\right|_{\psi=0}.$$

This expression is known as the Savage–Dickey density ratio (see [14] and references therein). The numerator is simply the marginal posterior under the more complex model evaluated at the simpler model's parameter value, while the denominator is the prior density of the more complex model evaluated at the same point. This technique is particularly useful when testing for one extra parameter at the time, because then the marginal posterior $p(\psi|d,\mathcal{M}_1)$ is a 1–dimensional function and normalizing it to unity probability content only requires a 1–dimensional integral, which is simple to do using for example the trapezoidal rule.

- **Information criteria.** Sometimes it might be useful to employ methods that aim at an approximate model selection under some simplifying assumptions that give a default penalty term for more complex models, which replaces the Occam's razor term coming from the different prior volumes in the Bayesian evidence [21].

  (i) **Akaike Information Criterion (AIC):** the AIC is an essentially frequentist criterion that sets the penalty term equal to twice the number of free parameters in the model, k:

(25) $$\mathrm{AIC} \equiv -2\ln \mathcal{L}_{\mathrm{max}} + 2k$$

where $\mathcal{L}_{\mathrm{max}} \equiv p(d|\theta_{\mathrm{max}},\mathcal{M})$ is the maximum likelihood value.

(ii) **Bayesian Information Criterion (BIC):** the BIC follows from a Gaussian approximation to the Bayesian evidence in the limit of large sample size:

$$\text{BIC} \equiv -2\ln\mathcal{L}_{\max} + k\ln N \tag{26}$$

where k is the number of fitted parameters as before and N is the number of data points. The best model is again the one that minimizes the BIC.

(iii) **Deviance Information Criterion (DIC):** the DIC can be written as

$$\text{DIC} \equiv -2D_{\text{KL}} + 2\mathcal{C}_b. \tag{27}$$

In this form, the DIC is reminiscent of the AIC, with the $\ln\mathcal{L}_{\max}$ term replaced by the estimated KL divergence D_{KL} and the number of free parameters by the effective number of parameters, $\mathcal{C}_b$ (see [13] for definitions).

The information criteria ought to be interpreted with care when applied to real situations. Comparison of Eq. (26) with Eq. (25) shows that for $N > 7$ the BIC penalizes models with more free parameters more harshly than the AIC. Furthermore, both criteria penalize extra parameters regardless of whether they are constrained by the data or not, unlike the Bayesian evidence. In conclusion, what makes the information criteria attractive, namely the absence of an explicit prior specification, represents also their intrinsic limitation.

References

[1] G. E. P. Box and G. C. Tiao, *Bayesian Inference in Statistical Analysis* (John Wiley & Sons, Chicester, UK, 1992).

[2] H. Jeffreys, *Theory of probability*, 3rd edn , Oxford Classics series (reprinted 1998) (Oxford University Press, Oxford, UK, 1961).

[3] E. T. Jaynes, *Probability Theory. The logic of science* (Cambridge University Press, Cambridge, UK, 2003).

[4] J. M. Marin and C. P. Robert, *Bayesian Core. A Practical Approach to Computational Bayesian Statistics* (Springer, New York, 2007).

[5] D. MacKay, *Information theory, inference, and learning algorithms* (Cambridge University Press, Cambridge, UK, 2003).

[6] D. Sivia, *Data Analysis: A Bayesian tutorial* (Oxford University Press, Oxford, UK, 1996).

[7] P. Gregory, *Bayesian logical data analysis for the physical sciences* (Cambridge University Press, Cambridge, UK, 2003).

[8] G.A. Young & R.L. Smith, *Essentials of Statistical Inference*, (Cambridge University Press, Cambridge, UK, 2005).

[9] G. D'Agostini, Probability and Measurement Uncertainty in Physics - a Bayesian Primer, (hep-ph:/9512295) (1995).

[10] T. J. Loredo, From Laplace to Supernova SN 1987A: Bayesian Inference in Astrophysics, in T. Fougere (Editor) *Maximum-Entropy and Bayesian Methods*, Available from: http://bayes.wustl.edu/gregory/articles.pdf (accessed Jan 15 2008) (Kluwer Academic Publishers, Dordrecht, The Netherlands, 1990), pp. 81–142.

[11] T. J. Loredo, The promise of Bayesian inference for astrophysics, in E. D. Feigelson and G. J. Babu (Eds) *Statistical Challenges in Modern Astronomy*, Available from: http://www.astro.cornell.edu/staff/loredo/bayes/promise.pdf (accessed Jan 15 2008) (Springer, New York, 1992), pp. 275–297.

[12] M. Hobson, A. Jaffe, A. Liddle, P. Mukherjee, *et al.* (Eds) *Bayesian Methods in Cosmology* (Cambridge University Press, Cambridge, UK, 2010).

[13] R. Trotta, Contemp. Phys. **49**, 71 (2008) [arXiv:0803.4089 [astro-ph]].

[14] R. Trotta, Mon. Not. Roy. Astron. Soc. **378**, 72 (2007) [arXiv:astro-ph/0504022].

[15] R. E. Kass & L. Wasserman, J. Am. Stat. Ass., **91**, 435, 1343–1370 (1996).

[16] R. E. Kass and A. E. Raftery, J. Am. Stat. Ass. **90**, 430, 773–795 (1995).

[17] F. Feroz and M. P. Hobson, Mon. Not. Roy. Astron. Soc., 384, 2, 449-463 (2008) arXiv:0704.3704 [astro-ph].

[18] J. Skilling, Nested sampling, in R. Fischer, R. Preuss and U. von Toussaint (Eds) *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, 735 (Amer. Inst. Phys. conf. proc. 2004), pp. 395–405.

[19] P. Mukherjee, D. Parkinson and A. R. Liddle, Astrophys. J. **638** L51–L54 (2006).

[20] T. Sellke, M. Bayarri and J. O. Berger, American Statistician **55** 62–71 (2001).

[21] A. R. Liddle, Mon. Not. Roy. Astron. Soc. **351** L49–L53 (2004).

[22] A. Raftery, Sociological Methodology **25** 111–163 (1995).

[23] A. Gelman and D.B. Rubin, Statistical Science, 7, 457-511 (1992).

[24] J. Dunkley, M. Bucher, P. G. Ferreira, K. Moodley and C. Skordis, Mon. Not. Roy. Astron. Soc. **356**, 925 (2005) [arXiv:astro-ph/0405462].

[25] M. Kunz, R. Trotta and D. Parkinson, Phys. Rev. **D74** 023503 (2006).