Astro-statistics and Cosmology - Assignment 3

Matteo Scialpi

November 23th, 2022

1 Introducing doubt in Bayesian model comparison

1.1 Main question: what is the main message of the paper?

The paper takes up the challenge to introduce the *doubt* concept to Bayesian analysis, to create a useful procedure in order to understand if, for a given set of data, between a set of known models there is the one that better describes the data or not.

In a frequentist analysis the main tool to estimate the quality of a fit for one model is the χ^2-per-degree-of-freedom estimator (reduced χ^2; in the following it will called χ^2_{red}), which will be better described in Q1, in sec. 1.2. The basic χ^2_{red} idea is to estimate the distance between each data point and the interpolant curve's correspondent point, divide it by the number of degrees of freedom and so have a best-fit analysis directly with one model on the data set.

The concept of doubt $\mathcal{D}$ is instead introduced in Bayesian analysis as the "degree of disbelief in the in the ability of any known model to describe the data". The Bayesian statistics' intrinsic problem behind this challenge is that this approach for model comparison gives fully information only about the relative goodness of a model with respect to another, while no information on the absolute goodness is given. The main message of the paper could be viewed in this scenario as the creation of a tool to understand the absolute quality of fit of the preferred model.

1.2 Q1: what is the χ^2-per-degree-of-freedom rule-of-thumb? How is the χ^2 distribution related to this test?

As we said in the previous section, the χ^2_{red} idea is to estimate the difference between measured and predicted value for each data (squeared), divided by the number of degrees of freedom, and so have a best-fit analysis directly with one model on the data set. Usually it is defined as, given the set of data d $(= d_1, d_2, \dots, d_N)$ normally distributed with σ_i^2 variance,

$$\chi^2_{red} = \frac{\chi^2}{dof} = \frac{1}{dof} \sum_{i=1}^{N} \frac{(d_i - d_{i,model})}{\sigma_i^2}\,, \tag{1}$$

where dof are the considered model's degrees of freedom. We see that, if data points are normal distributed, χ^2_{red} is distributed as a χ^2. The latter is also equal to twice the best-fit log-likelihood:

$$\mathcal{L}(\theta) \equiv P(d|\theta, M_j) \propto e^{-\chi^2/2}\,, \tag{2}$$

where θ_j are model M_j's parameters. This χ^2_{red} test is particularly suitable to understand the absolute quality of fit of one specific model to data: we have $\chi^2_{red} \gg 1$ if the model is unsatisfactory, $\chi^2_{red} = 1$ if the model is appropriate and $\chi^2_{red} \ll 1$ if the model is overspecified.

1.3 Q2: what is the Jeffreys' scale?

Jeffreys' scale is introduced in this paper during the review of Bayesian model selection, in order to understand how no information on the absolute fit quality is contained in Bayesian analysis.

For a single model M_j, the posterior probability, given the data d, is

$$P(M_j|d) = \frac{P(d|M_j)\,P(M_j)}{P(d)}\,, \tag{3}$$

where $P(M_j)$ is the prior belief in model M_j and $P(d)$ is the *evidence* normalisation constant and $P(d|M_j)$ is the Bayesian evidence, linked to the likelihood via the extended sum rule on M_j's parameter set θ_j:

$$P(d|M_j) = \int d\theta\, P(d|\theta_j, M_j)\, P(\theta_j|M_j)\,, \tag{4}$$

$\lvert\ln B_{01}\rvert$	*Odds*	*Strenght of evidence*
< 1.0	$\lesssim 3:1$	Inconclusive
1.0	$\sim 3:1$	Weak evidence
2.5	$\sim 12:1$	Moderate evidence
5.0	$\sim 150:1$	Strong evidence

Tabella 1: *Jeffreys' scale.*

where $P(d|\theta_j, M_j)$ is the likelihood and $P(\theta_j|M_j)$ is the prior on parameters.

Given two competing models (M_0 and M_1) we can introduce the Bayes factor B_{01}, ratio of model evidences,

$$B_{01} = \frac{P(d|M_0)}{P(d|M_1)}\,. \tag{5}$$

This B_{01} can be used to understand the strength of belief of one model with respect to the other. In this context the Jeffreys' scale is introduced: it is an empirical prescription for translating the value of B_{01} into strengths of belief. It can be seen in Tab. 1. So, B_{01} allows to select one (or few) model(s) for a set of known models, but gives no information about whether the selected model is a good explanation for the data (absolute goodness of fit).

1.4 Q3: how it is used and what is the Bayesian Information Criterion (BIC)?

The concept of doubt $\mathcal{D}$ is formalised in the paper as the posterior probability of the unknown model $\mathfrak{X}$, $P(\mathfrak{X}|d)$. This $\mathfrak{X}$ is an expansion of our space of models M_j including the possibility that the latter is in incomplete and there might be a better model that we haven't yet identified. The doubt is so defined as

$$\mathcal{D} \equiv P(\mathfrak{X}|d) = \frac{P(d|\mathfrak{X})\,P(\mathfrak{X})}{P(d)} = \left[1 + \sum_i \frac{P(d|M_i)\,P(M_i)}{P(d|\mathfrak{X})\,P(\mathfrak{X})}\right]^{-1}, \tag{6}$$

where last equality is done to make evident the odds ratio. For the prior we can say that simply $P(\mathfrak{X})$ and $P(M_i)$, with $i = 1, \ldots, N$, must sum to 1. The crucial step is to calculate $P(d|\mathfrak{X})$, which can't be computed via (4) because $\theta_{\mathfrak{X}}$ aren't fully known. Anyway, we can say that we want a Bayes factor like

$$B_{xi} = \frac{P(d|\mathfrak{X})}{P(d|M_i)} \ll 1\,, \tag{7}$$

looking at Tab. 1. This translates in a calibration of the absolute value of the Bayesian evidence for $\mathfrak{X}$.

Here comes the BIC, as an approximation tool for $P(d|\mathfrak{X})$. Denoting the likelihood of $\mathfrak{X}$ as $\mathcal{L}(\theta)$ ($\equiv P(d|\theta, \mathfrak{X})$) and the prior as $P(\theta|\mathfrak{X})$, we can define

$$g(\theta) = \ln[\mathcal{L}(\theta)\,P(\theta|\mathfrak{X})]\,. \tag{8}$$

Expanding at the second order around the maximum likelihood value θ_{max}, we have

$$g(\theta) \approx g(\theta_{max}) - \frac{1}{2}(\theta - \theta_{max})^T\,H\,(\theta - \theta_{max})\,, \tag{9}$$

where H is the Hessian matrix and k is the number of parameters in $\mathfrak{X}$. Using this in (4), we have

$$\ln P(d|\mathfrak{X}) = \ln \mathcal{L}_{max} + \ln P(\theta_{max}|\mathfrak{X}) + \frac{k}{2}\ln 2\pi - \frac{1}{2}\ln|H| + o(n^{-1})\,. \tag{10}$$

Approximating $H \approx nI$, where I is the expected Fisher matrix, and that the prior $P(\theta|\mathfrak{X})$ is a multivariate Gaussian centered in θ_{max} with Fisher matrix I, we can say

$$\ln P(\theta_{max}|\mathfrak{X}) = -\frac{k}{2}\ln 2\pi + \frac{1}{2}\ln|I|\,, \tag{11}$$

while substituting this in (10), we have

$$\ln P(d|\mathfrak{X}) = \ln \mathcal{L}_{max} - \frac{k}{2}\ln n + o(n^{-1/2})\,, \tag{12}$$

main formula for the BIC.

It always requires an estimate of the best-fit likelihood $\mathcal{L}_{max}$ and of the number of parameters k. The first one can be computed in relation with the estimator $\widehat{L_{max}} \equiv -2\ln\mathcal{L}_{max}$. This can be seen as the α quantile for $\chi^2_{m,(\alpha)}$ (m dofs), to avoid unjustified doubt as a consequence of harmless statistical fluctuations. This means that the BIC equation (12) becomes

$$\ln P(d|\mathcal{X},k,\alpha) = -\frac{\chi^2_{m,(\alpha)}}{2} - \frac{k}{2}\ln n\,, \tag{13}$$

where now we have k (number of parameters for $\mathcal{X}$) and α to study.

In conclusion, the BIC is an useful tool to approximate $P(d|\mathcal{X})$, which cannot be calculated via marginalization because we don't know the entire ser of parameters $\theta_{\mathcal{X}}$. All the long calculation in the previous question is reported here to give a context to Q4 and Q5.

1.5 Q4: is it always possible to Taylor expand the likelihood around the maximum? What does it represent the curvature term of this expansion?

The curvature term for $g(\theta)$ in (9) represents the distance of parameters' value from their maximum likelihood values. An high value for this curvature term represents a sharp maximum in the maximum likelihood, while a low value for this term represents a flat maximum. This allows us to understand if θ_{max} is a sharp maximum or not. In the first case, it will be easier to calculate the Bayesian evidence $P(d|\mathcal{X})$, since dominant values are peaked around θ_{max}. The limiting, unphysical, case in which we can't expand the likelihood around θ_{max} is when the first or the second derivatives goes to infinity, for example in a δ-like distribution.

1.6 Q5: what is the Fisher information matrix and how is related to the likelihood?

The Fisher information matrix is a $k\times k$ matrix with elements equal to the expectation value of the second derivative of the log likelihhod. In other words, its elements are the expectation value for the the corresponding H element. The Hessian matrix is defined as

$$H_{ij} \equiv -\frac{\partial^2 g(\theta)}{\partial\theta_i\,\partial\theta_j}\bigg|_{\theta=\theta_{max}}\,, \tag{14}$$

while the Fisher matrix is defined as

$$I_{ij} = -E\left[\frac{\partial^2 g(\theta)}{\partial\theta_i\,\partial\theta_j}\right]. \tag{15}$$

In our study, for large samples we have that $H \approx nI$ ($\mathcal{O}(n^{-1/2})$).

1.7 Q6: are you convinced by this paper? Report some idea to improve the use of doubt or to avoid it if deemed unnecessary.

This paper could be a milestone in Bayesian analysis, giving a specific tool to deal with absolute goodness of fit, exactly as χ^2_{red} in frequentist analysis. In my opinion it is of huge importance to understand if there is space for improving the model describing a precise data set, both for a theoretical and a data analysis point of view. Besides that, one needs anyway to enlarge the data set improve the knowledge of the evidence $P(d)$.

Besides that, two particular things could be studied in more detail: how can we know the number of parameters k for the unknown model $\mathcal{X}$ (value to be inserted in (13)) and how can we estimate the prior $P(\mathcal{X})$?

2 Should we doubt the cosmological constant?

2.1 Main question: what is the main message of the paper?

The aim of this paper is to apply the previous paper methodology of Bayesian model selection to dark energy cosmological problem. In particular this is done via the same definition of *doubt* $\mathcal{D}$ given in (6). The base model from which one can create the model set M_i and from which one can extend this set to the unknown model $\mathcal{X}$ is the ΛCDM model. As a consequence, we can define the generic Bayes factor B_{ij} as

$$B_{ij} \equiv \frac{P(d|M_i)}{P(d|M_j)} \tag{16}$$

and the average Bayes factor between ΛCDM model and each of the known models as

$$\langle B_{i\Lambda} \rangle \equiv \frac{1}{N} \sum_{j=1}^{N} B_{i\Lambda} \,. \tag{17}$$

This last quantity can estimate the goodness of ΛCDM model with respect to other known models. Furthermore, to avoid a large number of free parameters, we can define the evidence of the unknown model $\mathcal{X}$ via the following absolute upper bound

$$B_{\mathcal{X}\Lambda} < \overline{B}_{\mathcal{X}\Lambda} = e^{-(\chi^2_{\mathcal{X}} - \chi^2_{\Lambda})/2} \,. \tag{18}$$

Via these $\langle B_{i\Lambda} \rangle$ and $\overline{B}_{\mathcal{X}\Lambda}$ we can understand the goodness of ΛCDM both whit respect to other known models and to the new model.

Using a uniform prior only on the dark energy equation of state (EoS) w, only on the curvature density parameter Ω_k and on both of them, the model set was built, while for the unknown model, they supposed that a model $\mathcal{X}$ is fully specified once one gives the expression for $w(z)$, where z is the redshift. With this choice for M_i and $\mathcal{X}$, ΛCDM model is proven to be the best to fit present data, with only a little space remaining for improvement.

2.2 Q1: what is the purpose of MultiNest algorithm?

From (Feroz et al. 2009)[1]'s abstract: "This Bayesian inference tool calculates the evidence, with an associated error estimate, and produces posterior samples from distributions that may contain multiple modes and pronounced (curving) degeneracies in high dimensions". The March (et al.)'s approach is based on our first paper's theoretical work: they wants to calibrate the value of the Bayesian evidence $P(d|\mathcal{X})$ on simulated data sets from the best of known models M_i. However, for the aim of the paper we want to analyze, MultiNest is too much computationally expensive to implement. It is exactly for this reason that March (et al.) changed the approach to look at an upper bound for the evidence.

2.3 Q2: what is the Savage-Dickey density ratio?

From (Trotta 2007)[2] abstract: "I introduce the Savage–Dickey density ratio, a computationally quick method to determine the Bayes factor of two nested models and hence perform model selection". As the MultiNest algorithm, Savage-Dickey density ratio (SSDR) deals with model comparison in Bayesian analysis, but this time the calculated quantity is the Bayes factor itself. In particular SSDR is really good to deal with nested models because of the usage of non-informative priors. Anyway, again it is too much computationally expensive to implement for the work done by March (et al.).

2.4 Q3: what are CosmoMC and CAMB, and what are their differences?

CosmoMC (Lewis & Bridle, 2002)[3] is an algorithm that allows to span, with a Markov Chain Monte Carlo engine, cosmological parameter space. In this work that we want to analyse, the parameter estimation package was modified to sample $w_i \equiv w(z_i)$, too, where z_i are 10 values of redshift uniformly spaced from $z = 0$ to $z = 1.5$. To do some progress in the concept of doubt, some simplification were done for the $\mathcal{X}$ model, as for example the consideration of changes only in the stress-energy tensor $T_{\mu\nu}$ in Einstein equations.

To interpolate those w, March et al. use a plugin to CAMB (Lewis et al, 2000)[4] implementing the Parameterized Post Friedman (PPF) prescription, which uses cubic spline.

Differences between the two are well suited to work together: while CosmoMC spans the parameter space, CAMB interpolates the physics behind those parameters.

2.5 Q4: in your opinion this work added something new to the problem of Dark Energy? Please explain.

In my opinion, this work didn't update so much our Universe understanding in a theoretical way, while those results are of huge importance for data analysis: they are coherent with ΛCDM model. This remains the best known model to fit data from different sets, with a very little room left for improvement. Those statements anyway are robust considering only data available at the time of the paper commission (2010). Nothing prevents things to change if new or better data emerge. An example could be the streamlined distance ladder constructed

[1] https://academic.oup.com/mnras/article/398/4/1601/981502.
[2] https://arxiv.org/abs/astro-ph/0504022.
[3] https://journals.aps.org/prd/abstract/10.1103/PhysRevD.66.103511.
[4] https://arxiv.org/abs/astro-ph/9911177

from infrared observations of Cepheids and type Ia supernovae with ruthless attention to systematics. Improving the precision and accuracy on the Hubble costant, we now see evidence for 5σ deviations from ΛCDM[5].

One particular thing of this paper that doesn't convince me is the complete neglecting of the left hand side of Einstein's equations, despite the great range for Ω_k and $w(z)$ when able to vary, which must affect locally and globally the Einstein tensor $G_{\mu\nu}$. If the left hand side is considered instead, this will affect some data sets interpretation. For example gravitational lensing statistics, used to reconstruct cosmic voids and cosmic strings, and distance ladders, as supernovae Ia data, both based on light propagation through space-time, will be affected, giving a wrong bias to corresponding data sets.

[5]I was not able to find the related paper, anyway this was the topic of the last Nobel Lecture offered by the physics' department from Adam Riess on November 16th, 2022.