When is tension just a fluctuation?
How noisy data affects model comparison

B. Joachimi\textsuperscript{1}, F. Köhlinger\textsuperscript{2}, W. Handley\textsuperscript{3,4}, and P. Lemos\textsuperscript{1}

1 Department of Physics and Astronomy, University College London, Gower Street, London WC1E 6BT, UK
e-mail: b.joachimi@ucl.ac.uk
2 Kavli IPMU (WPI), UTIAS, The University of Tokyo, Kashiwa, Chiba 277-8583, Japan
3 Astrophysics Group, Cavendish Laboratory, J.J. Thomson Avenue, Cambridge CB3 0HE, UK
4 Kavli Institute for Cosmology, Madingley Road, Cambridge CB3 0HA, UK

Received 2020 September 30; accepted 2021 February 10

ABSTRACT

Summary statistics of the likelihood, such as the Bayesian evidence, offer a principled way of comparing models and assessing tension between, or within, the results of physical experiments. Noisy realisations of the data induce scatter in these model comparison statistics. For a realistic case of cosmological inference from large-scale structure we show that the logarithm of the Bayes factor attains scatter of order unity, increasing significantly with stronger tension between the models under comparison. We develop an approximate procedure that quantifies the sampling distribution of the evidence at small additional computational cost and apply it to real data to demonstrate the impact of the scatter, which acts to reduce the significance of any model discrepancies. Data compression is highlighted as a potential avenue to suppressing noise in the evidence to negligible levels, with a proof of concept on Planck cosmic microwave background data.

Key words. Methods: data analysis – Methods: statistical – Cosmology: observations – cosmic background radiation – gravitational lensing: weak

1. Introduction

Binary decisions inevitably have to be made at the conclusion of a physical experiment. Has a feature been detected significantly? Which model describes the data better? Are data sets (or subsets thereof) consistent with each other, or are they in ‘tension’, a potential indicator for new physics not incorporated in the model?

Traditionally, hypothesis tests were the statistical tools of choice to answer these questions. With the advent of high-performance computing, Bayesian techniques building on the prior model parameters. The Bayesian evidence, or marginal likelihood, is the normalisation given by

$$Z_i \equiv \Pr(d|M_i) = \int \text{d}^m p \Pr(d|p, M_i) \Pr(p|M_i),$$

where $\text{d}^m$ denotes the number of parameters. For a given data set, $Z_i$ reduces to a non-stochastic scalar that attains larger values the more likely the model is (as a more flexible model could accommodate many possible forms of the data).

However, a physical experiment does generally not take acquired data as a given, but rather interprets it as a stochastic realisation of an underlying truth that we wish to approximate by our model. A different realisation of the data leads to a different value for $Z_i$, which could alter our decision on tension or consistency. In this view, statistical uncertainty in the data turns the evidence (or any related tension and model comparison measure) into a noisy statistic\cite{Jenkins2011}. Jenkins & Peacock (2011) argued, based on toy experiments and analytical arguments, that

$$\Pr(p|d, M_i) = \frac{1}{Z_i} \Pr(d|p, M_i) \Pr(p|M_i) ,$$

which is then compared against a pre-defined scale to judge significance. In Bayesian statistics the tension measure is conditioned on the observed data. The posterior probability of the parameters $p$ of a model $M_i$, for measured data, $d$, is

1. This viewpoint will require us to go beyond a purely Bayesian approach. However, hybrid Bayesian-frequentist methods are commonplace in statistics; see \cite{Good1992} for an overview, and also \cite{Jenkins2014}.
the thus-inherited statistical uncertainty in $\mathcal{Z}_i$ is substantial. Ignoring this scatter will therefore lead to over-confident or incorrect decisions in model comparison.

In this work we quantify the scatter in the Bayesian evidence and some of its derived tension/model comparison statistics, affirming the findings of Jenkins & Peacock (2011) in a realistic cosmological experiment. We devise a computationally efficient procedure to calculate statistical errors on the evidence, apply it to an analysis of internal consistency in Kilo Degree Survey (KiDS) weak lensing data, and explore the impact of data compression on evidence scatter for the example of Planck CMB data.

2. Noisy model comparison

Figure 1 illustrates the notion of evidence and its associated scatter using a Gaussian toy model that is one-dimensional in both data and parameter space. It builds on Fig. 28.6 of MacKay (2003). While at the observed data Model 1 has higher evidence in this example, it is not unambiguously superior because alternative realisations of the data under the more probable Model 1 could result in equal or reversed evidences of Models 1 and 2 instead (see the boxes in blue and red shading). We seek to quantify this statistical uncertainty of the evidence. See also Appendix A for a closed-form analytic calculation in the Gaussian case analogously to Fig. 1.

2 For ease of illustration, the toy model considers the likelihood of the data conditioned on the best fit parameter $p_{\text{best}}$; in our implementation we take the full posterior into account when drawing new data realisations.

2.1. Scatter in the evidence and the Bayes factor

The standard statistic to compare two models $i$ and $j$ is the Bayes factor (see Kass & Raftery 1995 for a review),

$$R_{ij} \equiv \frac{\Pr(M_i|d)}{\Pr(M_j|d)} = \frac{\mathcal{Z}_i \Pr(M_i)}{\mathcal{Z}_j \Pr(M_j)} = \frac{\mathcal{Z}_i}{\mathcal{Z}_j},$$

which, for equal prior probabilities of the models themselves, is given by the ratio of the model evidences. $R_{ij}$ has the intuitive interpretation of betting odds in favour of model $i$ over $j$. Let us assume that we know the true underlying model $M_{\text{true}}$, including its parameters $p_{\text{true}}$, that generates the data we observe, which need not coincide with either $M_i$ or $M_j$. Then the probability density of the Bayes factor is given by

$$\Pr(R_{ij}|M_{\text{true}}) = \int d^n d' \Pr(R_{ij}|d') \Pr(d'|M_{\text{true}}),$$

where $n$ is the dimension of the data vector, $\Pr(d|M_{\text{true}})$ is the true likelihood of the data, and $\Pr(R_{ij}|d)$ the distribution of $R_{ij}$ for a given data set which we shall assume to be deterministic. Hence, if the true likelihood is known, we can proceed as follows to create a distribution of $R_{ij}$: (i) generate samples of the data from the true likelihood; (ii) for each sample calculate the Bayes factor according to Eqs. (2) and (3).

As a realistic example we choose a recent cosmological analysis of tomographic weak lensing measurements by the KiDS survey (KiDS-450, Kuijken et al. 2015; Hildebrandt et al. 2017). We work with a simulated data vector that, like the real data, has size $n = 130$ and depends in a highly non-linear way on 7 model parameters (5 cosmological parameters of a flat $\Lambda$CDM model, plus two parameters describing astrophysical effects on the observables). It is assumed that the data have a Gaussian likelihood
with a known and fixed covariance. To perform an internal consistency test, we create two copies of the parameter set and assign one copy to the elements of the data vector dependent on tomographic bin no. 3, and the other copy to the remaining elements. The model comparison is then between the analysis with the original set of model parameters (Model 0) and that with the doubled parameter set (Model 1). For details of the methodology and analysis, see Köhlinger et al. (2019) and Appendix B.

We generate 100 realisations of the data vector from the true likelihood, evaluated at a fiducial choice of the parameters. For each simulated data vector we repeat a full nested sampling analysis of both models (0 and 1) and infer the evidences (see Appendix B for an assessment of the robustness of the sampling algorithms). By default we do not introduce any systematic shift into our simulated data, so that strong concordance is expected as the outcome of the tension analysis.

The resulting distribution of evidences is shown in Fig. 2. We compute the true value of the Bayes factor by re-running the analyses for a noise-free data vector generated for the fiducial parameter values. The two evidence distributions are each consistent with being lognormal in each with a standard deviation in the log of 7.9. The evidences are strongly correlated (Pearson correlation coefficient 0.99), which is plausible as the scatter derives from the same noisy data realisation, with both models yielding good fits.

Due to the strong correlation, the distribution of the Bayes factor is narrower, with $\sigma(\ln R_{01}) \approx 1.25$; see Fig. 3 for its distribution. We also observe skewness in $\ln R_{01}$ (already visible in Fig. 2), which causes the mean to be lower relative to the true value by $\sim 1\sigma$. We do not find evidence that the skewness is due to numerical issues, so ascribe it to the non-linearity of the models, which means this feature will be strongly dependent on the details of the analysis. A value of $\sigma(\ln R_{01}) \sim 1$ is in excellent agreement with the conclusions of Jenkins & Peacock (2011) although they predicted a normal distribution for $\ln R_{01}$ (see also Appendix A).

### 2.2. Impact on suspiciousness

By design, the Bayes factor depends on the parameter prior, which can be a hindrance for tension assessment, as demonstrated by Handley & Lemos (2019b). They proposed a modified statistic called suspiciousness, defined as

$$
S_{ij} \equiv \ln R_{ij} + D_{\text{KL}}^i - D_{\text{KL}}^j,
$$

where

$$
D_{\text{KL}}^i = \int d^m p_i \ln \frac{\Pr(p_i|d_i, M_i)}{\Pr(p_i|d_i, M_j)}
$$

is the Kullback-Leibler (KL) divergence (Kullback & Leibler 1951; see Lemos et al. 2020 for a generalisation to correlated data sets, which we consider here). This combination of evidence and KL divergence is independent of the prior widths, which means this statistic is appropriate for tension assessment.

Again assuming a Gaussian likelihood, $\ln S_{ij}$ is $\chi^2$-distributed with the degrees of freedom given by the difference in the effective dimension of the parameter space in the two models. Handley & Lemos (2019a) propose to calculate this effective dimension as

$$
m_{\text{eff}} = 2 \left\langle \left( \ln \Pr(p_i|d_i, M_j) \right)^2 \right\rangle_p - \left\langle \ln \Pr(p_i|d_i, M_j) \right\rangle_p^2,
$$

i.e. twice the variance of the log-likelihood evaluated over the posterior distribution (indicated by the subscript ‘$p$’).

We extract $\ln S_{ij}$ from the output of our nested sampling analysis and determine the scatter of $m_{\text{eff}}$ from the sub-sample variance computed on a posterior sample. The standard deviations of $\ln R_{01}$ and $\ln S_{ij}$ agree to better than 10%; see the following section for an argument why the distributions of $R$ and $S$ are expected to be very similar. The standard deviation of $m_{\text{eff}}$ is of order 10% and can be treated as uncorrelated with $S$ (correlation coefficient -0.17).

There are two obstacles to using the approach above on real data: (i) repeating full likelihood analyses including evidence
calculations many times to build a sample is prohibitively expensive, and (ii) we do not know the true likelihood to generate samples of the real data. We will address both points in Sects. 3 and 4.

2.3. Strong tension case

To investigate a case of strong tension, we insert a large shift in the mean redshift of tomographic bin no. 3 of \(dz = 0.3\) into the simulated data vectors and repeat the analysis\(^\text{[5]}\). In this case the alternative model 1 is clearly preferred (\(\ln R_{01} \approx -23\)). The impact on the distribution of \(\ln R_{01}\) is dramatic, as can be seen from Fig. 3. While the skewness and corresponding discrepancy between mean and true value persist, the standard deviation increases to 7.3, spanning more than three orders of magnitude in odds.

This result is driven by an increase in the scatter of the evidences for models 0 and 1, thereby reducing the difference between the model predictions at the respective best-fit parameters (as also shown in Appendix A). This difference is small in our concordant case with nested models, deviating from zero only through scatter in the data. In the \(dz = 0.3\) case the best fits of the models now lie far apart, enlarging the scatter and propagating the noise differently into the evidences for models 0 and 1, thereby reducing their correlation.

3. A fast approximate algorithm

Let us consider the Laplace approximation of the log-likelihood (we drop the explicit dependence on the model for simplicity)\(^\text{[6]}\):

\[
\ln Pr(d|p) \approx \ln \ln Pr(d|p_0) - \frac{1}{2} (p - p_0)^T F^{-1}(p_0)(p - p_0),
\]

where we expanded around the maximum of the log-likelihood at \(p_0\) and introduced the Fisher matrix

\[
F_{\alpha\beta} = -\left(\frac{\partial^2 \ln \Pr(d|p)}{\partial p_\alpha \partial p_\beta}\right|_{p_0},
\]

as the negative expectation of the Hessian of the log-likelihood at \(p_0\). With this approximation the evidence reads

\[
\ln Z \approx \frac{2n^{n/2} \Pr(d|p_0)}{\sqrt{\det F(p_0)}} V_{\text{prior}},
\]

where we additionally assumed that the prior is uninformative, i.e. the bulk of the likelihood lies well within the volume covered by the prior, denoted as \(V_{\text{prior}}\). Considering a Gaussian likelihood, so that \(\Pr(d|p_0) \propto e^{-\frac{1}{2} \chi^2(p_0)}\), one finds (cf. Handley & Lemos 2019b)

\[
\ln Z \approx \ln \ln V_{\text{prior}} - \frac{1}{2} \ln \det F(p_0) - \frac{1}{2} \chi^2(p_0); \tag{10}
\]

\[
D_{\text{KL}} \approx \ln V_{\text{prior}} - \frac{m}{2} (1 + \ln 2\pi) + \frac{1}{2} \ln \det F(p_0). \tag{11}
\]

Note that we employ the fast approximate algorithm detailed in Sect. 4.

We thank our referee for pointing out that this assumption has a more principled grounding in that it maximises the entropy in absence of further information on the form of the likelihood.

We see that the only source of scatter is due to the best-fit parameter set \(p_0\), which varies with the noise realisation of the data. If we further assume that the curvature of the likelihood does not vary strongly as the best-fit position moves, only the last term in Eq. (10) is relevant for the statistical uncertainty in \(\ln Z\), while \(D_{\text{KL}}\) is robust to the scatter. Since \(\ln Z + D_{\text{KL}} \approx \const - \chi^2(p_0)/2\), \(\ln S\) has identical noise properties to \(\ln Z\) under these assumptions.

Equipped with these considerations, we propose the following algorithm: (i) perform a single full likelihood analysis and determine fiducial evidence values, \(Z_{\text{fid}}\); (ii) generate samples of the data from the likelihood; (iii) for each sample determine the maximum of the likelihood or equivalently, \(\chi^2\); (iv) derive samples of the evidence via

\[
\ln Z_{\text{approx}} := \ln Z_{\text{fid}} - \frac{1}{2} \left( \chi^2_{\text{min}} - \chi^2_{\text{min,fid}} \right). \tag{12}
\]

Following this procedure with 100 samples results in the blue points shown in Fig. 2 and the blue distribution in Fig. 3. Apart from sampling noise in the tail, we recover the true distribution well, with mean and variance in agreement within \(\sim 10\%\). The change from a full exploration of the posterior, which typically runs in hours to days, to a maximisation of the likelihood, which will usually take minutes to hours, makes exploring the noise properties of the Bayesian evidence and its derived quantities feasible.

4. Evidence samples from real data

When analysing real data, the true likelihood \(\Pr(d'|M_{\text{true}})\) found in Eq. (4) from which to generate new copies of the data vector is unavailable. Our best guess for this truth is the best-fitting model, which itself carries uncertainty as it is inferred from the data. In this case we can make use of the posterior predictive distribution (PPD, Gelman et al. 1996), \(\Pr(d'|d, M_k)\), which yields new samples of the data \(d'\) for a given observation \(d\) assuming model \(M_k\) (see Trotta 2007 for a very similar application of the PPD). Averaging over all models using the posterior model probabilities \(\Pr(M_k|d)\) from Eq. (3) then yields

\[
\Pr(d'|M_{\text{true}}) = \sum_k \Pr(d'|d, M_k) \Pr(M_k|d). \tag{13}
\]

The algorithm presented in Sect. 3 therefore only needs to be adjusted in Step (ii), where instead of generating data realisations from the true likelihood in the mock scenario, these are now produced from the PPD by randomly selecting a subset of posterior samples and evaluating the likelihood at the parameter values corresponding to these samples.

In practice, we simplify the approach by choosing the model that yields the higher evidence to produce the PPD samples, rather than full model averaging. If both models have similar evidences, the choice should have little impact; if the evidence ratio is large, the model with higher evidence is more accurate and/or more predictive (cf. the solid and hatched regions in Fig. 1).

5. Application to KiDS-450 internal consistency

We now insert the real KiDS-450 data vector into our analysis and generate 10 PPD samples from the joint Model 0 as this yields a significantly higher evidence than the split model. The evidence is large, the model with higher evidence is more accurate and/or more predictive (cf. the solid and hatched regions in Fig. 1).

In practice, we obtain the minimum \(\chi^2\) within the wide prior ranges of the parameters.
derived standard deviations of $\ln R$ and $\ln S$ are shown in Fig. 4. These statistical errors far exceed the typically quoted ‘method’ errors which derive from the finite sampling of the posterior. The interpretation of the suspiciousness acquires an additional, albeit smaller, source of error through the effective model dimension $m_{\text{eff}}$ that determines the $\sigma$-levels.

The noise in the tension statistics leads to a more conservative evaluation of discrepancies in the data. While the point estimate suggests ‘tension’ at $1.6\sigma$, this reduces to $1.1\sigma$ if we require that all but $16\%$ (i.e. the one-sided tail beyond $1\sigma$ of a normal distribution) of possible realisations of the data are discrepant by at least that level. Visually, this corresponds to the upper $1\sigma$ error of $\ln S$ almost touching the lower limit of the $1\sigma$ band in Fig. 4.

6. Benefits of data compression

Planck CMB data are at the centre of both current major tension controversies in cosmology. A practical obstacle to applying our formalism is the complexity of the Planck temperature likelihood, which is assumed to be Gaussian only for $\ell > 30$ and builds on pixelised sky maps on larger scales (Planck Collaboration et al. 2020b). This makes drawing PPD samples challenging. However, Prince & Dunkley (2019) recently showed that the low-$\ell$ likelihood can be efficiently compressed into two Gaussian-distributed band powers. They proceeded to apply maximal, linear compression (using the MOPED scheme, Tegmark et al. 1997; Heavens et al. 2000) to the full temperature likelihood and demonstrated it to be nearly lossless. This is not unexpected since the cosmological sampling parameters in CMB analyses are chosen to be close to linear and Gaussian-distributed (Kosowsky et al. 2002).

There is an additional motivation to apply data compression: it can suppress scatter in the Bayesian evidence. Under the assumptions of Sect. 5 the statistical properties of $\ln S$ are driven by the distribution of $\chi^2(p_n)$, i.e. the minimum $\chi^2$ (cf. Eq. 10). If the data are approximately Gaussian and well fitted by a model whose parameters are close to linear, $\chi^2(p_n)$ follows a $\chi^2$-distribution with $N_{\text{dof}} = n - m$ degrees of freedom, so that $\text{Var}(\ln Z) = 2N_{\text{dof}}$. Data compression decreases $n$ and can yield $N_{\text{dof}} \approx 0$ in the maximal case, i.e. evidence becomes essentially noise-free because a good model with $n$ linear parameters perfectly fits $n$ compressed data. Appendix A demonstrates this explicitly for the Gaussian case.

This may seem paradoxical because compression cannot at best preserve information, so how can it facilitate a more precise determination of evidence? In the context of Fig. 1 compression reduces the scatter between the model parameter and the data, so that for a given parameter the data varies little and thus the evidence is known precisely. Conversely, a broad likelihood and/or a high-dimensional data vector lead to large variations in possible realisations of data. While this has no bearing on the posterior, and therefore information content, it increases the probability that a certain level of tension or model preference is owed to a particularly (un)lucky noise realisation of the data vector and does not reflect a physical trend.

As a proof of concept, we adopt the Prince & Dunkley (2019) approach using the provided software\footnote{https://github.com/heatherprince/planck-lite-py} and compress the Planck temperature (TT) power spectra into the six cosmological parameters of a spatially flat $\Lambda$CDM model (nuisance parameters are marginalised over pre-compression). We then determine the $\chi^2_{\text{min}}$ for the compressed real data, as well as for new data realisations generated from the compressed likelihood. We find an extremely small $\chi^2_{\text{min}} \approx 1.4 \times 10^{-8}$ for the real data and similar values for the noise realisations, with standard deviation of $4.4 \times 10^{-9}$. Hence, practically noise-free evidence measurements from Planck are indeed possible.

7. Conclusions

We studied the impact on model comparison statistics if these are based on the ensemble of possible observations, rather than a single observed realisation of the data. In this setting they become noisy quantities, which affects binary decisions on signal detection, model selection, or tension between experiments. Confirming earlier analytic arguments, we found standard deviations of order unity for the logarithm of the Bayes factor, and the suspiciousness statistic, with substantially broader distributions in case of strong discrepancies between the models under comparison. We expect these conclusions to apply to most, possibly all, informative tension metrics available in the literature as they typically depend on the maximum likelihood or $\chi^2$-like expressions.

We proposed a method to approximate the probability distribution of the evidence via repeated draws of mock data from the likelihood and then obtaining the maximum likelihood for each mock data set, which will add negligible computation time to a full exploration of the posterior distribution. Conclusions drawn from noisy model comparison measures inevitably become more conservative, e.g. the tension significance according to the suspiciousness for an internal consistency analysis of KiDS weak lensing data reduces from 1.6$\sigma$ in the traditional approach to 1.1$\sigma$ when scatter is accounted for. While in this application the two models under comparison were nested, our formalism and conclusions also hold for the more general case in which parameter spaces differ.

Finally, we demonstrated that data compression suppresses the impact of noisy data on the evidence, in the case of Planck CMB constraints to negligible levels. In light of this, the following pre-processing steps are beneficial before any form of model...
comparison: (i) compress the data vector as much as possible as long as the compression is essentially lossless; and (ii) choose a parametrisation such that the model is close to linear in the parameters (see e.g. Schuhmann et al. 2016) which increases the chances of achieving a near-perfect fit for any noise realisation of the data.

Acknowledgements. We thank C. Jenkins, J. Peacock, and R. Trotta for insightful discussions. We are also grateful to C. Jenkins for alerting us to the work of Jenkins and Peacock. BJ acknowledges the kind hospitality of IPMU where part of this work was carried out. FK acknowledges support from the World Premier International Research Center Initiative (WPI), MEXT, Japan. PL acknowledges STFC Consolidated Grant ST/R000476/1.

The KiDS analysis is based on observations made with ESO Telescopes at the La Silla Paranal Observatory under programme IDs177.A-3016, 177.A-3017, 177.A-3018, 179.A-2004, 298.A-5015, and on data products produced by the KiDS Consortium.

References

Abbott, T. M. C., Abdalla, F. B., Alarcon, A., et al. 2018, Phys. Rev. D, 98, 043526
Adhikari, S. & Huterer, D. 2019, J. Cosmology Astropart. Phys., 2019, 036
Asgari, M., Tröster, T., Heymans, C., et al. 2020, A&A, 634, A127
Audren, B., Lesgourgues, J., Benabed, K., & Prunet, S. 2015, J. Cosmol. Astropart. Phys., 2015, 001
Brinckmann, T. & Lesgourgues, J. 2018, arXiv e-prints [arXiv:1804.07261]
Charnock, T., Battye, R. A., & Moss, A. 2017, Phys. Rev. D, 95, 123535
Feroz, F. & Hobson, M. P. 2008, MNRAS, 384, 449
Feroz, F., Hobson, M. P., & Bridges, M. 2009, MNRAS, 398, 1601
Feroz, F., Hobson, M. P., Cameron, E., & Pettitt, A. N. 2013, arXiv e-prints [arXiv:1306.2144]
Gelman, A., Meng, X. L., & Stern, H. 1996, Statistica Sinica, 6, 733
Good, I. J. 1992, J Am Stat Assoc, 87, 597
Handley, W. & Lemos, P. 2019a, Phys. Rev. D, 100, 023512
Handley, W. & Lemos, P. 2019b, Phys. Rev. D, 100, 043504
Handley, W. J., Hobson, M. P., & Lasenby, A. N. 2015a, MNRAS, 450, L61
Handley, W. J., Hobson, M. P., & Lasenby, A. N. 2015b, MNRAS, 453, 4384
Heavens, A. F., Jimenez, R., & Lahav, O. 2000, MNRAS, 317, 965
Heavens, A. F., Kitching, T. D., & Verde, L. 2007, MNRAS, 380, 1029
Heymans, C., Tröster, T., Asgari, M., et al. 2020, arXiv e-prints, arXiv:2007.15632
Hildebrandt, H., Viola, M., Heymans, C., et al. 2017, MNRAS, 465, 1454
Jaffe, A. 1996, ApJ, 471, 24
Jenkins, C. 2014, in American Institute of Physics Conference Series, Vol. 1636.
Bayesian Inference and Maximum Entropy Methods in Science and Engineering, 106–111
Jenkins, C. R. & Peacock, J. A. 2011, MNRAS, 413, 2895
Joudaki, S., Blake, C., Heymans, C., et al. 2017, MNRAS, 465, 2033
Kass, R. E. & Raftery, A. E. 1995, J Am Stat Assoc, 90, 773
Köhlinger, F., Joachimi, B., Asgari, M., et al. 2019, MNRAS, 484, 3126
Kozaczuk, E., Mőlósavójecse, M., & Jimenez, R. 2002, Phys. Rev. D, 66, 063007
Kuijken, K., Heymans, C., Hildebrandt, H., et al. 2015, MNRAS, 454, 3500
Kullback, S. & Leibler, R. A. 1951, Ann. Math. Statist., 22, 79
Kunz, M., Trotta, R., & Parkinson, D. R. 2006, Phys. Rev. D, 74, 023503
Lazaresis, G., Ruiz de Austri, R., & Trotta, R. 2004, Phys. Rev. D, 70, 123527
Lemos, P., Köhlinger, F., Handley, W., et al. 2020, MNRAS, 496, 4647
Lin, W. & Ishak, M. 2017, Phys. Rev. D, 96, 023502
MacKay, D. J. C. 2003, Information Theory, Inference, and Learning Algorithms (Cambridge University Press)
Marshall, P., Raguram, N., & Slosar, A. 2006, Phys. Rev. D, 73, 067302
Nicola, A., Amara, A., & Refregier, A. 2019, J. Cosmology Astropart. Phys., 2019, 011
Petersen, K. B. & Pedersen, M. S. 2012, The Matrix Cookbook, version 20121115
Planck Collaboration, Aghanim, N., Akrami, Y., et al. 2020a, A&A, 641, A6
Planck Collaboration, Aghanim, N., Akrami, Y., et al. 2020b, A&A, 641, A5
Prince, H. & Dunkley, J. 2019, Phys. Rev. D, 100, 083502
Ravert, M. & Hu, W. 2019, Phys. Rev. D, 99, 043508
Riess, A. G., Casertano, S., Yuan, W., et al. 2018, ApJ, 861, 126
Riess, A. G., Casertano, S., Yuan, W., Macri, L. M., & Scoclin, D. 2019, ApJ, 876, 88
Riess, A. G., Macri, L. M., Hoffmann, S. L., et al. 2016, ApJ, 826, 56
Schuhmann, R. L., Joachimi, B., & Peiris, H. V. 2016, MNRAS, 459, 1916
Seehars, S., Amara, A., Refregier, A., Paranjape, A., & Akerron, J. 2014, Phys. Rev. D, 90, 023533
Skilling, J. 2006, Bayesian Anal., 1, 833
Tegmark, M., Taylor, A. N., & Heavens, A. F. 1997, ApJ, 480, 22
Trotta, R. 2007, MNRAS, 378, 819
Trotta, R. 2008, Contemporary Physics, 49, 71
Verde, L., Protopapas, P., & Jimenez, R. 2013, Physics of the Dark Universe, 2, 166
Verde, L., Treu, T., & Riess, A. G. 2019, Nature Astronomy, 3, 891
Wong, K. C., Suyu, S. H., Chen, G. C. F., et al. 2020, MNRAS, 498, 1420
Appendix A: The complete Gaussian case

Assume that an experiment produces a single observation of \( n \) data points, drawn from a Gaussian distribution about some true mean \( \hat{d} \) with covariance \( \mathbf{C} \):

\[
\ln \Pr(d) = -\frac{1}{2} \ln |2\pi \mathbf{C}| - \frac{1}{2} (d - \hat{d})^T \mathbf{C}^{-1} (d - \hat{d}).
\]

(A.1)

In general we do not know \( \hat{d} \), but design a model \( M \) which parameterises the data by some function \( f(p) \), with \( m \ll n \) parameters, in the hope that the true data are well approximated by our model. Assuming a given model, the likelihood becomes

\[
\ln \Pr(d | p, M) = -\frac{1}{2} \ln |2\pi \mathbf{C}| - \frac{1}{2} (d - f(p))^T \mathbf{C}^{-1} (d - f(p)).
\]

(A.2)

Often a likelihood may be approximated as a Gaussian in the parameter space

\[
\ln \Pr(d | M) = \ln \mathcal{L}_{\text{max}} - \frac{1}{2} \hat{\delta}^T \mathbf{C}^{-1} \hat{\delta}.
\]

(A.3)

with mean \( \mu \) and covariance \( \Sigma \), and the corresponding log-evidence reads

\[
\ln \mathcal{Z} \equiv \ln \Pr(d | M) = \ln \mathcal{L}_{\text{max}} + \ln \sqrt{\frac{2\pi \Sigma}{\mathcal{V}_{\text{prior}}}},
\]

where \( \mathcal{V}_{\text{prior}} \) is the volume of a uniform prior fully encompassing the posterior. One can make the link between Eqs. (A.2) and (A.3) explicit by assuming that we can model our function \( f \) as linear in the region of parameter space around \( p_* \) where the likelihood is significantly non-zero,

\[
f(p) \approx f(p_*) + \nabla f(p_*)(p - p_*) = : \hat{d} + J(p - p_*),
\]

(A.4)

from which one can identify

\[
\ln \mathcal{L}_{\text{max}} = -\frac{1}{2} \ln |2\pi \mathbf{C}| - \frac{1}{2} (d - \hat{d})^T \mathbf{C}^{-1} (d - \hat{d});
\]

\[
\mathbf{C}^{-1} = J^T \mathbf{C}^{-1} J; \quad \mu = p_* + \Sigma J \mathbf{C}^{-1} (d - \hat{d}),
\]

(A.6)

where we defined

\[
\mathbf{C}^{-1} = \mathbf{C}^{-1} - \mathbf{C}^{-1} \mathbf{J} J^T \mathbf{C}^{-1}.
\]

(A.7)

As an aside, Eq. (A.7) shows that noisy data realisations affect the posterior mean but not its covariance. In other words, while the posterior shape is unaffected, the distribution moves as a whole in parameter space with different realisations of the data.

From the above expressions we can immediately see that the evidence is quite a noisy statistic, driven by the second term in Eq. (A.6). Taking the variance of Eq. (A.4) after inserting Eq. (A.6) and assuming that \( d \) follows the distribution of Eq. (A.1) yields

\[
\text{Var}(\ln \mathcal{Z}) = \frac{1}{2} \text{Tr} \left( (\mathbf{C}^{-1} \mathbf{C})^2 + (d - \hat{d})^T \mathbf{C}^{-1} \mathbf{C} \mathbf{C}^{-1} (d - \hat{d}) \right).
\]

(A.9)

The first term here is equal to \( \frac{1}{2}(n - m) \), and hence the variance in the raw evidence is large for \( n \gg m \), even in the event of a good fit to the data (i.e. \( d \approx \hat{d} \)). We also see that in the case of heavily compressed data, \( n \sim m \), the evidence scatter reduces considerably.

To derive the expression (A.9) above, and some of the following ones, it is helpful to note that for a Gaussian-distributed variable \( x \) with covariance \( \Sigma \) centered on zero [Petersen & Pedersen 2012]

\[
\langle (x - a)^T A(x - a) \rangle = \text{Tr} [\mathbf{AC}] + a^T a \Sigma ;
\]

\[
\text{Cov} [(x - a)^T A(x - a)], \quad (x - b)^T B (x - b)] = 2 \text{Tr} [\mathbf{AC}^2] + 4b^T B \mathbf{CA} a.
\]

(A.10)

where \( A \) and \( B \) are symmetric matrices, and \( a \) and \( b \) are arbitrary, non-stochastic vectors.

We are of course really interested in how model comparison (i.e. a difference in evidence) scatters with noisy data, so we introduce two models with \( \tilde{d}_1, \tilde{d}_2, J_1\) and \( J_2 \), respectively\(^8\), and ask what is the variance in their evidence difference. Under the true distribution of Eq. (A.1), we find that the log Bayes factor (under the same assumptions as in Eq. (1)), \( \ln R_{12} = \ln \mathcal{Z}_{1} - \ln \mathcal{Z}_2 \), has mean

\[
\langle \ln R_{12} \rangle = \frac{1}{2} \langle (\tilde{d}_1 - \tilde{d}_2)^T \tilde{C}_2^{-1} (\tilde{d}_1 - \tilde{d}_2) \rangle - \frac{1}{2} \langle (\tilde{d} - \hat{d}_1)^T \tilde{C}_1^{-1} (\tilde{d} - \hat{d}_1) \rangle + \frac{1}{2} \text{Tr} [\Delta] + \ln \sqrt{\frac{2\pi \tilde{\Sigma}}{\mathcal{V}_{\text{prior},2}}} - \ln \sqrt{\frac{2\pi \tilde{\Sigma}}{\mathcal{V}_{\text{prior},1}}},
\]

(A.11)

(see also Lazarides et al. 2004, Heavens et al. 2007 for similar, less general expressions) and variance

\[
\text{Var}(\ln R_{12}) = \frac{1}{2} \text{Tr} \left( \Delta^2 + (\tilde{\Delta} - \hat{\Delta}) \tilde{C}_1^{-1} (\Delta \hat{\Delta} - \hat{\Delta}) \right),
\]

(A.12)

where we defined

\[
\Delta := C(\tilde{C}_2^{-1} - \tilde{C}_1^{-1}); \quad \hat{\Delta} := C(\tilde{C}_1^{-1} \tilde{d}_2 - \tilde{C}_1^{-1} \tilde{d}_1) .
\]

(A.13)

The mean in Eq. (A.11) has three portions, a set of misfit terms on the first line, a constant trace term equal to \( \frac{1}{2}(m_1 - m_2) \) and an Occam factor. The trace contribution can be understood as a typically small modification of the Occam factor.

In the variance (Eq. (A.12)) there is a trace term which is roughly the dimensionality of the parameter space(s) \( \lesssim \frac{1}{2}(m_1 + m_2) \), as well as a data misfit term. The trace term is always present, and represents the ‘order unity’ term for the general Gaussian case, but can reduce to zero (via a cross-term dependent on both models) the more similar the two model parametrisations (as quantified by \( J \)) are to each other. As opposed to the variance of the evidence (cf. Eq. (A.9)), the trace term in the variance of the Bayes factor does not depend on \( n \), so if \( n \gg m \), the scatter in \( R_{12} \) is significantly smaller than the scatter in either \( \mathcal{Z}_1 \) and \( \mathcal{Z}_2 \) if the data is well fitted. This is the situation we encountered in Fig. 2.

The second term may be small if the models are good, but can become arbitrarily large, which corresponds to the scatter seen in Fig. 3. It should be noted that in the event of large misfits, the mean and variance are both of the same order, which gives a Poisson-type evidence error associated with measurement noise. This is reassuring as it means the evidence in theory becomes relatively less noisy the larger it becomes.

We note that, if Eq. (A.3) is a reasonable approximation, provided one can compute (by numerical derivatives or otherwise) the Jacobian \( J \), one may use Eq. (A.12) to evaluate the expected
The data are publicly available at [http://kids.strw.leidenuniv.nl/sciencedata.php](http://kids.strw.leidenuniv.nl/sciencedata.php) for Version 3.8 from [polychordlite](https://github.com/polychord/leidenuniv.nl/sciencedata.php) and at [Multinest](https://github.com/multinest/multinest) and [MontePython](https://github.com/brinckmann/montepython_public) for Version 1.16 from [MontePython](https://www.montepython-project.org/). The implementation of the inference pipeline is that presented in [Köhlinger et al. (2019)](https://arxiv.org/abs/1909.02763), which is independent of, but in excellent agreement with, the analysis of [Hildebrandt et al. (2017)](https://arxiv.org/abs/1709.09010).

We opt for nested sampling ([Skilling 2006](https://arxiv.org/abs/0804.3631)) to explore the posterior distribution as the most efficient way to evaluate high-dimensional likelihoods and calculate Bayesian evidence simultaneously. To avoid significant algorithm-induced scatter in the evidence values, we check three variants of nested sampling algorithms for their consistency. We use [MULTINEST](https://github.com/fcoehlin/montepython_2cosmos_public) ([Feroz & Hobson 2008](https://arxiv.org/abs/0804.3631), [Feroz et al. 2009](https://arxiv.org/abs/0901.4830), [2013](https://arxiv.org/abs/1303.6955)), [POLYCHORD](https://github.com/polychord/polychordlite) ([Handley et al. 2015a,b](https://arxiv.org/abs/1504.01999), [2015b](https://arxiv.org/abs/1510.03198), [2016](https://arxiv.org/abs/1608.00381)), which primarily differ in the key step of how new ‘live’ sampling points are drawn at each likelihood contour. Moreover, we consider an importance-sampled determination of the evidence in [MULTINEST](https://github.com/fcoehlin/montepython_2cosmos_public) that utilises the full set of generated sample points and can achieve higher accuracy ([Feroz et al. 2013](https://arxiv.org/abs/1303.6955)). For [MULTINEST](https://github.com/fcoehlin/montepython_2cosmos_public) 1000 live points are used with a sampling efficiency of 0.3 and a final error tolerance on the log-evidence of 0.1. Live points for [POLYCHORD](https://github.com/polychord/polychordlite) runs are 25 times the number of parameters (7 for Model 0; 14 for Model 1) with a final error tolerance on the log-evidence of 0.001.

Figure B.1 shows evidence values for a KiDS-like noise-free simulated data vector, as well as for ten realisations with noise included. It is evident that in all cases, and for both the joint and split cosmological models, the three nested sampling variants agree very well, with the residual scatter at a small fraction of the statistical errors. Our [MULTINEST](https://github.com/fcoehlin/montepython_2cosmos_public) and [POLYCHORD](https://github.com/polychord/polychordlite) settings were optimised to yield accurate evidences. However, we note that evidence values are faithfully recovered as soon as the bulk of the posterior is explored, while credible regions of the parameters as well as the effective dimension (see Eq. 2) are sensitive to the tails of the distribution. Therefore, when these latter quantities are required in high-stakes real-data applications, we recommend increasing the accuracy settings of the nested sampling runs.

The $\chi^2$ minimisation for the approximate method is performed with the built-in [MONTE PYTHON](https://www.montepython-project.org/) maximum likelihood determination, with a precision tolerance of $10^{-9}$ on the log-likelihood. With this setup a minimisation run consumes about 500 times less wall-clock time than full, parallelised sampling on high-performance computing infrastructure.

---

10 The data are publicly available at [http://kids.strw.leidenuniv.nl/sciencedata.php](http://kids.strw.leidenuniv.nl/sciencedata.php)
11 Likelihood pipelines available in [MONTE PYTHON](https://www.montepython-project.org/). [Audren et al. (2013)](https://arxiv.org/abs/1303.6955), [Brinckmann & Lesgourgues (2018)](https://arxiv.org/abs/1805.06592), and from [https://github.com/fcoehlin/montepython_2cosmos_public](https://github.com/fcoehlin/montepython_2cosmos_public)
12 Version 3.8 from [http://ccpforge.cse.rl.ac.uk/gf/project/multinest/](http://ccpforge.cse.rl.ac.uk/gf/project/multinest/)
13 Version 1.16 from [https://github.com/polychord/polychordlite](https://github.com/polychord/polychordlite)