Parameter inference and model comparison using theoretical predictions from noisy simulations

Niall Jeffrey[1]⋆ and Filipe B. Abdalla[1,2]

[1] *Department of Physics & Astronomy, University College London, Gower Street, London, WC1E 6BT, UK*
[2] *Department of Physics and Electronics, Rhodes University, PO Box 94, Grahamstown, 6140, South Africa*

ABSTRACT
When inferring unknown parameters or comparing different models, data must be compared to underlying theory. Even if a model has no closed-form solution to derive summary statistics, it is often still possible to simulate mock data in order to generate theoretical predictions. For realistic simulations of noisy data, this is identical to drawing realisations of the data from a likelihood distribution. Though the estimated summary statistic from simulated data vectors may be unbiased, the estimator has variance which should be accounted for. We show how to correct the likelihood in the presence of an estimated summary statistic by marginalising over the true summary statistic in the framework of a Bayesian hierarchical model. For Gaussian likelihoods where the covariance must also be estimated from simulations, we present an alteration to the Sellentin-Heavens corrected likelihood. We show that excluding the proposed correction leads to an incorrect estimate of the Bayesian evidence with JLA data. The correction is highly relevant for cosmological inference that relies on simulated data for theory (e.g. weak lensing peak statistics and simulated power spectra) and can reduce the number of simulations required.

Key words: methods: statistical – methods: data analysis – cosmology: observations

1 INTRODUCTION

It is increasingly common, especially in cosmological surveys, to attempt to make inferences from data **d** using theory summary statistics μ that can be obtained only from simulations.

One example, currently popular in cosmology, is weak lensing peak statistics (Dietrich & Hartlap 2010; Kacprzak et al. 2016; Peel et al. 2017; Shan et al. 2018; Martinet et al. 2018). Peak statistics broadly aim to use the number of density peaks in the cosmological matter distribution to constrain cosmological parameters and models. The number of peaks in the density field (or weak lensing signal) is the result of highly non-linear structure formation and large-volume dark matter simulation are often needed to generate the theoretical number. The number of peaks in a given simulation is stochastic due to cosmic variance and then further sources of noise are added to simulate realistic observed data.

The data **d** are not the raw maps or catalogues, but the observed/derived summary statistics. For example, **d** would be the observed number of peaks in a weak lensing mass map reconstruction (Kacprzak et al. 2016; Jeffrey et al. 2018).

It was recently noted by Sellentin & Heavens (2016)

that the common practice of debiasing a covariance matrix estimated from simulations of mock data (Hartlap et al. 2007) is insufficient. The sampling distribution of this estimated covariance should be incorporated into the likelihood distribution and, therefore, into the posterior distributions of the inferred parameters. Failure to do so leads to biased and overly-optimistic inferences.

In this work we note that, as with the estimated covariance described by Sellentin & Heavens (2016), an unbiased estimated summary statistic $\hat{\mu}$ is nevertheless itself a random variable, drawn from a sampling distribution with associated variance. If unaccounted for, this will lead to inaccurate parameter inference and misleading model comparison results.

In a Bayesian hierarchical framework, we present how to derive the posterior distribution of parameters by using a corrected likelihood distribution which takes into account that the theoretical predictions are based on noisy simulations.

An alternative to our presented hierarchical approach is to use likelihood-free inference (LFI) methods (e.g. Alsing et al. 2018, Leclercq 2018). In general LFI methods assume that the likelihood is unknown, and simulations are used to estimate the resulting posterior distribution conditional on data. However, if simulations are expensive and

⋆ E-mail: niall.jeffrey.15@ucl.ac.uk

we believe we know the naive likelihood[1] then LFI methods would be unnecessary and require more simulations, due to the increased number of degrees of freedom in the model. Instead, we show how to directly construct a Bayesian hierarchical model containing the free parameters of the known naive likelihood.

In Section 2 we show how to marginalise over possible summary statistics μ to derive the likelihood $P(\mathbf{d} \mid \hat{\mu})$, the probability of the data $\mathbf{d}$ conditional on the estimated summary statistic $\hat{\mu}$. In Section 3, we consider the case in which the naive likelihood is Gaussian, and derive the corrected likelihood distribution in the presence of both known and unknown (estimated) covariance matrices. In Section 4, we use a one-dimensional toy model to demonstrate the effect of estimated summary statistics; we show that the corrected likelihood distribution matches samples generated from the toy model. In Section 5, we use the public JLA supernovae data to show the effect of estimating summary statistics from simulations, using draws from the known likelihood as mock simulations.

2 LIKELIHOOD-BASED INFERENCE WITH SIMULATIONS

2.1 Posterior and likelihood

Using the example of weak lensing peak statistics, we would wish to evaluate the posterior distribution of parameters of interest θ (e.g. Ω_m, σ_8 etc.) conditional on our observed data $\mathbf{d}$ (the number of peaks in a weak lensing map), which is given by Bayes' theorem:

$$P(\theta|\mathbf{d}) = \frac{P(\mathbf{d}|\theta)\ P(\theta)}{P(\mathbf{d})} \ . \tag{1}$$

If we were able to directly model the summary statistic with a deterministic function $\mu(\theta)$ (e.g. if we could calculate the number of weak lensing peaks for given parameters θ), then we could substitute[2] $\mu(\theta)$ into the likelihood:

$$P(\mathbf{d}|\theta) = P(\ \mathbf{d} \mid \mu(\theta)\ ) \ . \tag{2}$$

Of course, this is not possible in many cases where the deterministic model is not known. We must rely on estimates $\hat{\mu}_i$ of the true, but unattainable, summary statistic, with simulations run at position i in parameter space with parameters θ_i.

The correct representation of Bayes' theorem for the posterior distribution is

$$P(\theta|\mathbf{d}, \{\hat{\mu}_i\}) = \frac{P(\mathbf{d}|\{\hat{\mu}_i\}, \theta)\ P(\theta)}{P(\mathbf{d})} \ , \tag{3}$$

or equivalently

$$P(\theta|\mathbf{d}, \{\hat{\mu}_i\}) = \frac{P(\mathbf{d}, \{\hat{\mu}_i\}|\theta)\ P(\theta)}{P(\{\hat{\mu}_i\})\ P(\mathbf{d})} \ , \tag{4}$$

---

[1] Assuming we could condition on the true summary statistics which can be calculated $P(\mathbf{d}|\mu(\theta))$.
[2] For a deterministic $\mu = \mathbf{f}(\theta)$, then $P(\mu|\theta) = \delta^{\mathrm{Dirac}}(\mathbf{f}(\theta) - \mu)$. Substitution is performed via marginalisation over a latent μ'.
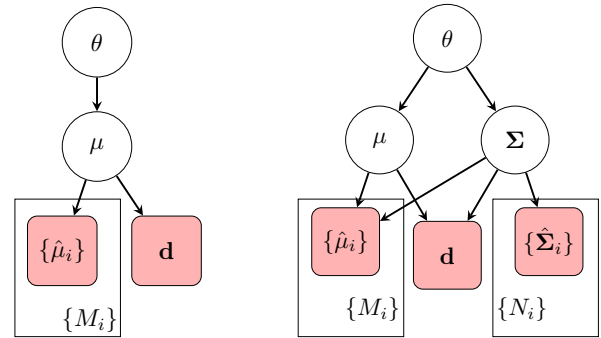

Figure 1. Probabilistic graphical representation in plate notation of Bayesian hierarchical models from Sections 2.2 & 3.1 (*left panel*) and Section 3.2. (*right panel*). Shaded nodes are "observed", either from experimental data ($\mathbf{d}$) or from simulations ($\hat{\mu}$ and $\hat{\Sigma}$). Each plate (rectangular box) includes the amount of data associated with the variable, for example each $\hat{\mu}_i$ (run at position i in parameter space) comes from M_i simulations.

keeping in mind that, as observed quantities, the observed summary statistic $\mathbf{d}$ and the statistics from the simulations $\{\hat{\mu}_i\}$ are independent and, therefore, separable. The brackets {} represent the set of simulations run over the positions i in parameter space.

The factors $P(\{\hat{\mu}_i\})$ and $P(\mathbf{d})$ are both Bayesian evidence terms for the observed data and observed simulations. As constants, they can be largely ignored.

2.2 Likelihood correction

At position i in parameter space a set of M simulated data are generated. As it is not possible to calculate[3] the summary statistic μ, we estimate it from M simulated data realisations $\mathbf{d}_{\mathrm{sim}}$. The estimate is often the mean

$$\hat{\mu}(\theta_i) = \frac{1}{M} \sum_{j=0}^{M-1} \mathbf{d}_{\mathrm{sim},j}(\theta_i) \ , \tag{5}$$

where each j_{th} data realisation is independent. In some cases, the summary statistic may not be the expectation of the likelihood $P(\mathbf{d}|\mu)$, so the estimator $\hat{\mu}$ would be another function of the simulated data (not the mean).

Even if $\hat{\mu}$ is an unbiased estimate (that is, $\langle \hat{\mu} \rangle = \mu$), it is often mistakenly assumed that

$$P(\ \mathbf{d} \mid \hat{\mu}_i\ ) = P(\ \mathbf{d} \mid \mu(\theta_i)\ ) \ , \tag{6}$$

which is not generally correct.

The correct likelihood to be used for parameter inference is with the estimated $\{\hat{\mu}_i\}$ from simulations as used

---

[3] that is, $\mu(\theta)$ cannot be calculated for a given θ (e.g. peak statistics)

in equation 3 or equation 4. This can be rewritten as a marginalisation over the unknown true summary statistic

$$
\begin{aligned}
P(\mathbf{d} \mid \{\hat{\mu}_i\},\ \theta) &= \int P(\mathbf{d},\ \mu \mid \{\hat{\mu}_i\},\ \theta)\ \mathrm{d}\mu \\
&= \int P(\mathbf{d} \mid \mu,\ \theta)\ P(\mu \mid \{\hat{\mu}_i\},\ \theta)\ \mathrm{d}\mu \\
&= \frac{1}{P(\{\hat{\mu}_i\})} \int P(\mathbf{d} \mid \mu,\ \theta)\ P(\{\hat{\mu}_i\} \mid \mu,\ \theta)\ P(\mu|\theta)\ \mathrm{d}\mu\ ,
\end{aligned}
\tag{7}
$$

or alternatively

$$
P(\mathbf{d}, \{\hat{\mu}_i\} \mid\ \theta) = \int P(\mathbf{d} \mid \mu,\ \theta)\ P(\{\hat{\mu}_i\} \mid \mu,\ \theta)\ P(\mu|\theta)\ \mathrm{d}\mu\ ,
\tag{8}
$$

which is the same up to a constant evidence factor $P(\{\hat{\mu}_i\})$, as would be expected from equations 3 and 4. Which of the previous two distributions one wishes to think of as the traditional *likelihood* is somewhat academic, as once they are included into Bayes' theorem (equation 3 or 4), the posterior is the same.

Adding the corrected likelihood into equation 3 gives the posterior distribution for the parameters of interest

$$
P(\theta|\mathbf{d}, \{\hat{\mu}_i\}) = \frac{P(\theta)}{P(\mathbf{d})\ P(\{\hat{\mu}_i\})} \int P(\mathbf{d} \mid \mu)\ P(\{\hat{\mu}_i\} \mid \mu)\ P(\mu|\theta)\ \mathrm{d}\mu\ .
\tag{9}
$$

Here and in what follows, we drop explicit dependence on θ in terms like $P(\{\hat{\mu}_i\} \mid \mu,\ \theta)$, and distributions will only condition on the previous layer in the hierarchical model (figure 1), as is a common practice (e.g. Leistedt et al. 2018). One can nevertheless keep in mind that probability distributions with μ is also conditioned on θ.

Equation 9 is the most general form, but each factor in the final integral can be evaluated for certain forms of the naive likelihood ($P(\mathbf{d} \mid \mu)$) and chosen priors:

• The first factor is the naive likelihood that would be used if the summary statistic μ could be calculated.

• The second factor is a sampling distribution of $\hat{\mu}$. If our simulated datasets $\mathbf{d}_{\mathrm{sim},j}$ for a given set of parameters θ_i are independent and realistic, then each is an independent and identically distributed (i.i.d.) draw from the naive likelihood distribution. Assuming we know the naive likelihood, it is usually possible to evaluate the sampling distribution $P(\hat{\mu} \mid \mu)$.

• The final factor of equation 7 is the prior on the summary statistic μ conditional on position in parameter space. This is not assigning prior probabilities to the values of the parameters themselves but on to possible forms of the summary statistic.

If we believe, for example, that μ should vary smoothly in parameters space, this could be enforced through a smoothness prior $P(\mu|\theta)$. This can be compared to emulation methods (e.g. Heitmann et al. 2009, Bird et al. 2019, Jennings et al. 2019), where μ is estimated from simulations using a smoothing prior, either explicitly (Gaussian processes) or implicitly (machine learning). However, the uncertainties are not generally included in the final posterior distribution in a principled hierarchical way as described here.

Conversely, if we claim to know nothing (or very little) about μ *a priori*, then we might consider a uniform prior (Section 3.1).

2.3 Bayesian Hierarchical Model

The model described in the previous section is hierarchical and has a network of parameters related by conditional probabilities. Specific probability distributions of interest, such as the posterior probability distribution of the parameters, are evaluated by appropriate use of Bayes' theorem and marginalisation.

The left panel of figure 1 shows the probabilistic graphical representation of the hierarchical model pertaining to equation 9. This graphical representation may make the logical steps of the previous section clearer, especially allowing us to see where conditions can be dropped (e.g. $P(\mathbf{d} \mid \mu,\ \theta) = P(\mathbf{d} \mid \mu)$).

3 GAUSSIAN NAIVE LIKELIHOOD

For many cosmological analyses, the data $\mathbf{d}$ are assumed to have Gaussian noise and are, therefore, drawn from a Gaussian likelihood. In this section we derive the corrected likelihood for the case with known covariance (Section 3.1) and unknown (estimated) covariance (Section 3.2).

A Gaussian likelihood is usually an approximation, as there are likely to be some sources of non-Gaussian noise. It may be a very good approximation however. By the central limit theorem it may be the correct distribution in some limit of large numbers. For example, in a survey to measure the matter power spectrum $P(k)$, if the galaxies are a Poisson process, then for modes that average many galaxies (high k) the likelihood is approximately Gaussian. Similarly, if weak lensing peaks are Poissonian, the binned counts of peaks will be approximately Gaussian for large numbers of observed peaks.

For cases where the naive likelihood is non-Gaussian and one wishes to calculate the corrected likelihood (conditional on an estimated summary statistic $\hat{\mu}$), one should evaluate equation 7 analytically or numerically.

3.1 Known Covariance

Consider the case where the naive likelihood (the first factor in the integral of equation 9) is a Gaussian/normal distribution, such that

$$
P(\mathbf{d} \mid \mu(\theta)\ ) = \mathcal{N}(\mathbf{d},\ \mu,\ \Sigma)\ ,
\tag{10}
$$

and the covariance Σ is assumed known.

In this case, the sampling distribution (the second factor in the integral of equation 9) is

$$
P(\hat{\mu}_i \mid \mu_i,\ \Sigma) = \mathcal{N}(\hat{\mu}_i,\ \mu_i,\ \frac{1}{M}\Sigma)\ ,
\tag{11}
$$

at position i in parameter space.

If we do not wish to enforce any prior knowledge about μ, it seems reasonable to use the Jeffreys prior (Jeffreys 1946,

Jeffreys 1998) as an objective prior distribution for μ, which, for our Gaussian likelihood (equation 10), is uniform

$$P(\mu|\theta) = P(\mu) \propto 1 \quad . \tag{12}$$

This flat prior on μ, if unbounded, is formally improper. However, the resulting (posterior-like) distribution, $P(\mu_i \mid \hat{\mu}_i, \Sigma)$, is Gaussian and therefore a true probability distribution.

With these distributions (equation 10-12) for a Gaussian naive likelihood, we perform the marginalisation integration (equation 7),

$$
\begin{aligned}
P(\mathbf{d} \mid \hat{\mu}_i, \Sigma) &= \frac{1}{P(\{\hat{\mu}_i\})} \int \mathcal{N}(\mathbf{d}, \mu, \Sigma) \, \mathcal{N}(\hat{\mu}_i, \mu, \tfrac{1}{M}\Sigma) \, \mathrm{d}\mu \\
&\propto \int \mathcal{N}\big(\mathbf{d}, \hat{\mu}_i, \big(\tfrac{M+1}{M}\big)\Sigma\big) \, \mathcal{N}(\mu, \mathbf{x}, \mathbf{X}) \, \mathrm{d}\mu \,,
\end{aligned}
\tag{13}
$$

where $\mathbf{x}$ is a certain function of $\{\mathbf{d}, \hat{\mu}, \Sigma, M\}$ (but not μ) and $\mathbf{X}$ is a certain function of $\{\Sigma, M\}$[4]. The first factor can be brought outside the integral. The integration over μ of the second factor, which is a normal distribution, evaluates to one, which removes the dependence on $\mathbf{x}$. Normalising the resulting distribution with respect to $\mathbf{d}$ gives the corrected likelihood:

$$P(\mathbf{d} \mid \hat{\mu}_i, \Sigma) = \mathcal{N}\big(\mathbf{d}, \hat{\mu}_i, \big(\tfrac{M+1}{M}\big)\Sigma\big) \,. \tag{14}$$

For summary statistics ($\hat{\mu}_i = \hat{\mu}(\theta_i)$) estimated from simulations (run with θ_i parameters), where the likelihood distribution for data $\mathbf{d}$ conditional on the true (but unknown) μ_i is Gaussian with known covariance Σ, and with an objective Jeffreys prior on μ, then equation 14 is the corrected form of the likelihood. It is this corrected likelihood that should be used for parameter inference.

In this case, the corrected likelihood has the same Gaussian form as the naive likelihood, but with a scaled covariance. At first glance, this scaling could be mistaken for Bessel's correction for an unbiased estimate of the sample variance; however, here we actually know the covariance Σ and the added scaling comes from uncertainty in our estimate $\hat{\mu}$.

3.2 Unknown Covariance

In the previous section, we assumed that the summary statistic μ was estimated from simulations, but the covariance Σ was known. This situation is unlikely and it is foreseeable that both the summary statistic and covariance would also be estimated from simulated data.

The estimate of the covariance from N independent data simulations is given by

$$\hat{\mathbf{S}} = \frac{1}{N-1} \sum_{i=0}^{N-1} (\mathbf{d}_{\mathrm{sim},i} - \bar{\mathbf{d}}_{\mathrm{sim}})(\mathbf{d}_{\mathrm{sim},i} - \bar{\mathbf{d}}_{\mathrm{sim}})^{\dagger} \,, \tag{15}$$

where $\dagger$ is the conjugate transpose, and $\bar{\mathbf{d}}_{\mathrm{sim}} =$

---

[4] This can be shown by completing the square, and is listed in Bromiley (2003) and Petersen & Pedersen (2012).

$\frac{1}{N} \sum_{i=0}^{N-1} \mathbf{d}_{\mathrm{sim},i}$ is the just the estimated summary statistic $\hat{\mu}$ as given by equation 5 (but with N simulations, not M).

For the case where the summary statistic μ is not estimated from simulations, Sellentin & Heavens (2016) calculate the corrected likelihood

$$P(\mathbf{d} \mid \mu, \hat{\mathbf{S}}, N) \propto \int P(\mathbf{d} \mid \mu, \Sigma) P(\hat{\mathbf{S}} \mid \Sigma) P(\Sigma) \, \mathrm{d}\Sigma \,. \tag{16}$$

For a Gaussian naive likelihood $P(\mathbf{d} \mid \mu, \Sigma)$ the distribution of the estimated covariance $P(\hat{\mathbf{S}} \mid \Sigma)$ is Wishart. With these distributions and a Jeffreys prior for Σ, the resulting Sellentin-Heavens corrected likelihood is given by

$$
\begin{aligned}
&P(\mathbf{d} \mid \mu, \hat{\mathbf{S}}) = \\
&\frac{\Gamma(\frac{N}{2})|\hat{\mathbf{S}}|^{-1/2}}{\Gamma(\frac{N-p}{2})[\pi(N-1)]^{\frac{p}{2}}} \left[1 + \frac{(\mathbf{d}-\mu)^{\dagger}\hat{\mathbf{S}}^{-1}(\mathbf{d}-\mu)}{N-1} \right]^{-\frac{N}{2}} \,.
\end{aligned}
\tag{17}
$$

where p is the number of elements in the data vector $\mathbf{d}$ (i.e. the dimensionality). This has the form of a multivariate t-distribution.

In the case considered in this work, represented by the right panel of figure 1 we are assuming that the summary statistic μ cannot be calculated, and that we must estimate $\hat{\mu}$ from simulations. The integral in equation 16 must then be replaced by

$$
\begin{aligned}
&P(\mathbf{d} \mid \{\hat{\mu}_i\}, \{\hat{\Sigma}_i\}, \theta) = \frac{1}{P(\{\hat{\mu}_i\})P(\{\hat{\Sigma}_i\})} \times \\
&\iint P(\mathbf{d}|\mu,\Sigma)P(\{\hat{\mathbf{S}}_i\}|\Sigma)P(\{\hat{\mu}_i\}|\mu,\Sigma)P(\Sigma|\theta)P(\mu|\theta) \, \mathrm{d}\mu \, \mathrm{d}\Sigma \,.
\end{aligned}
\tag{18}
$$

This integral is the (corrected) likelihood in the posterior distribution of θ for the model shown in the right-hand panel of figure 1.

We note that a given sample mean $\hat{\mu}_i$ and sample covariance $\hat{\mathbf{S}}_i$ are independent, even if the simulated data used to estimate them are the same (Anderson 2003), so that $P(\hat{\mu}, \hat{\mathbf{S}}) = P(\hat{\mu}) \, P(\hat{\mathbf{S}})$. If the simulated data used to evaluate $\hat{\mu}$ and $\hat{\mathbf{S}}$ are the same, one can set $N = M$ in what follows. If, as is often the case in cosmological analyses, the covariance is assumed not to vary with respect to the parameters of interest and is instead estimated once with more simulated data realisations, then N is fixed and the i indices on the estimated covariance are dropped.

Using the same distributions as described so far in Section 3, marginalising over the unknown true summary statistic μ and covariance Σ (equation 18) and renormalising gives the new corrected likelihood (details in Appendix A):

$$
\begin{aligned}
&P(\mathbf{d} \mid \hat{\mu}_i, \hat{\mathbf{S}}_i) = \\
&\frac{\Gamma(\frac{N}{2})\sqrt{\frac{M}{M+1}}|\hat{\mathbf{S}}_i|^{-1/2}}{\Gamma(\frac{N-p}{2})[\pi(N-1)]^{\frac{p}{2}}} \left[1 + \frac{M}{(M+1)(N-1)}(\mathbf{d}-\mu_i)^{\dagger}\hat{\mathbf{S}}_i^{-1}(\mathbf{d}-\mu_i) \right]^{-\frac{N}{2}}
\end{aligned}
\tag{19}
$$

This corrected likelihood gives the probability of observing the data $\mathbf{d}$ conditional on an estimated mean summary
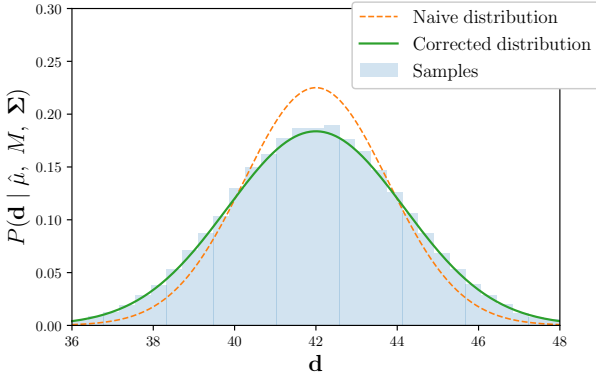
Figure 2. One-dimensional toy model for the naive Gaussian likelihood with known variance, where $\mu = 42$, $\Sigma = \pi \approx 3.14$, and $M = 2$. The samples (described in Section 4) are distributed according to the corrected likelihood distribution (equation 14), whereas the naive likelihood distribution has reduced variance.

statistic $\hat{\mu}_i = \hat{\mu}(\theta_i)$ from M simulations and an estimated covariance matrix $\hat{\Sigma}_i = \hat{\Sigma}(\theta_i)$ from N simulations, where we have assumed that the naive likelihood is Gaussian and have used Jeffreys priors on μ and Σ.

4   TOY MODEL DEMONSTRATION

As a verification and demonstration of the result given in equation 14, where $\hat{\mu}$ is estimated from simulated data and Σ is known, we construct a toy model. This toy model also relies on the assumed flat prior on μ and the fact that the sampling distribution is symmetric with respect to μ and $\hat{\mu}$.

Let us assume that in different realisations of an experiment, different experimenters randomly and independently generate M simulations, from which $\hat{\mu}$ is estimated according to equation 5. The underlying likelihood distribution for the data with known μ is Gaussian and therefore the simulated data are themselves i.i.d. draws from the Gaussian distribution (equation 10). Each experimenter then draws a realisation of the data $\mathbf{d}$ from the naive Gaussian likelihood distribution with the known variance and the mean given by their estimated $\hat{\mu}$.

Though each experimenter draws their data realisation $\mathbf{d}$ from a Gaussian likelihood with the known variance Σ, the data realisations from all the experimenters will be distributed according to the corrected likelihood (equation 14) with variance $\frac{M+1}{M}\Sigma$.

In figure 2 we take a one-dimensional case where $\mu = 42$, $\Sigma = \pi \approx 3.14$, and $M = 2$. We see that the data samples from 10^5 different experiments match the corrected likelihood distribution (equation 14), whereas the naive likelihood distribution underestimates the variance, as expected.

5   JLA SUPERNOVAE DEMONSTRATION

In this section, we use public type Ia supernova (SN Ia) data[5] from the SDSS-II/SNLS3 Joint Light-Curve Analysis

(JLA) (Betoule et al. 2014) as a demonstration of the corrected likelihoods described in the previous sections.

This is, of course, only a demonstration, as the summary statistic $\mu(\theta)$ (SN Ia apparent magnitudes) can actually be calculated for the model considered. We generate simulated data by drawing realisations from the known likelihood[6], and estimate $\hat{\mu}$. We can then constrain cosmological parameters using a likelihood distribution conditional on our estimated $\hat{\mu}$.

5.1   Data and Model

The data are observed B-band peak apparent magnitudes $\mathbf{d} = (m_{B,\text{obs},1}, m_{B,\text{obs},2}, ...)$ for 740 SN Ia over a range of redshifts up to $z = 1.3$. The supernovae also have associated light-curve stretch X_1, colour at maximum-brightness C and host stellar mass M_{stellar}, which are included in the model and covariance. The data and associated covariance are shown in figure 3.

We use the model from Betoule et al. (2014) where the SN Ia are standardizable candles with expected apparent magnitude

$$m_B = 5 \log_{10}\left(\frac{D_L}{1\text{Mpc}}\right) + 25$$
$$+ M_B + \Delta M \Theta(M_{\text{stellar}} - 10^{10} M_\odot)$$
$$- \alpha X_1 + \beta C , \tag{20}$$

where α and β are nuisance parameters for the stretch and colour respectively. M_B is the absolute magnitude of the host with a correction term ΔM depending on M_{stellar} (where Θ is the Heaviside function). We take a flat wCDM Universe, with luminosity distance

$$D_L = \frac{c(1+z)}{H_0} \int_0^z \frac{\text{d}z'}{\sqrt{\Omega_m(1+z')^3 + (1-\Omega_m)(1+z')^{3(w+1)}}} , \tag{21}$$

where c is the speed-of-light in a vacuum, H_0 is the Hubble parameter, Ω_m is the matter density parameter, and w is the equation of state for dark energy.

In our demonstration, the parameters of interest $\theta = (\Omega_m, w, M_B)$ are allowed to vary. For simplicity, we fix the other parameters: $H_0 = 70.0$ km s^{-1} Mpc^{-1}, $\Delta M = -0.05$, $\alpha = 0.1256$, $\beta = 2.6342$. The covariance is calculated according to Betoule et al. (2014) and is also fixed.

5.2   Likelihood and Priors

We assume a Gaussian likelihood, where the log-likelihood is given by:

$$\ln P(\mathbf{d}|\mu, \Sigma) = -\frac{1}{2}\ln\left[\det(2\pi\Sigma)\right] - \frac{1}{2}(\mathbf{d} - \mu)^\dagger \Sigma^{-1}(\mathbf{d} - \mu) , \tag{22}$$

where the data and covariance are those described in Section 5.1 (and shown in Fig. 3), and our summary statistic μ is given in equation 20.

---

[5] supernovae.in2p3.fr/sdss_snls_jla/ReadMe.html

[6] Alsing et al. (2018) and Leclercq (2018) take a similar approach to demonstrate likelihood-free methods.
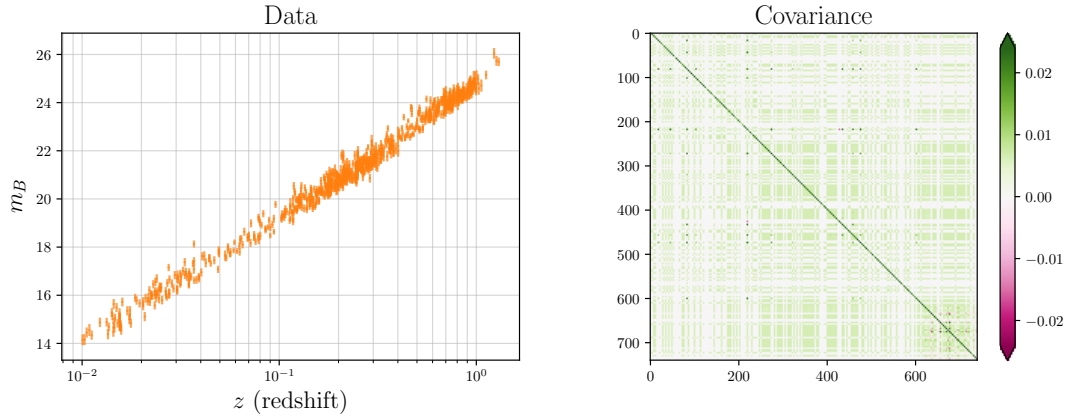
Figure 3. *Left panel*: The observed magnitude m_B data for 740 SN Ia. The error bars are taken as the square-root of the diagonal elements of the covariance. *Right panel*: The covariance matrix as described in Section 5.1.

For simplicity we take uniform priors in the ranges: $0.05 \leq \Omega_m \leq 0.6$, $-1.5 \leq w \leq -0.3$, $-19.15 \leq M_B \leq -18.95$.

Simulations for this demonstration are run on a regular grid of shape [12, 13, 11] (for Ω_m, w, M_B) spanning the prior range. The regular grid is a particularly poor choice to sample the posterior distribution when simulations are expensive. However, this is a demonstration, and for real-world analysis many better sampling schemes are available (including latin hypercubes and grid transformations to better sample the expected posterior distribution).

Once the posterior is evaluated at these grid positions, the parameter space is upsampled to a [48, 52, 44] grid. The new grid positions are evaluated by interpolating the posterior distribution from the original grid using a radial basis function 'thin plate' spline (Duchon 1976; Bookstein 1989; Jones et al. 2001). This spline interpolation worked particularly well in avoiding edge effects or artefacts around points when we compared their results with those of more poorly-performing simple polynomial splines.

5.3   Results

First, let us imagine three different experimenters, who, despite having access to the same data (described above), run their own independent simulations to estimate the summary statistics $\hat{\mu}(\theta)$ on the grid in parameter space. This results in different $\hat{\mu}_i$ for $i = 1, 2, 3$.

Evaluating the posterior distribution using the naive Gaussian likelihood set-up described in Section 5.2, and using $M = 2$ simulations per parameter grid position, results in the three posterior distributions in figure 4. The three different experimenters have three different posterior distributions due to their different $\hat{\mu}_i$.

Having different individual posterior distributions is in itself is not a problem. If different experiments have different data but the same underlying parameters, their resulting posterior distributions will look different, and will quantify their own individual uncertainty in the parameters. However, this variance of the data has been taken into account, and will be reflected in each posterior distribution. In the case of different $\hat{\mu}_i$ in figure 4, the fact that $\hat{\mu}$ was a random draw from a distribution (just like the data) has not been taken into account. As they have ignored the resulting cor-
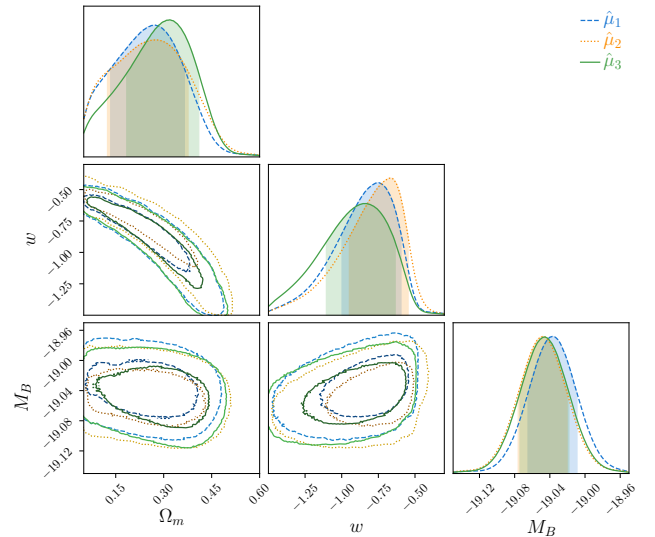

Figure 4. JLA posterior distribution for Ω_m, w, and M_B (described in Section 5.2) using three independent estimates $\hat{\mu}$ with $M = 2$ simulations per position in parameter space. This uses the naive Gaussian distribution (equation 10) without the correction (equation 14), and the contours are therefore optimistically reduced.

rection to their likelihood, each experimenter will be overly optimistic about their own inferences.

In figure 5, the posterior distribution has been calculated using the likelihood correction (equation 14), which takes into account the variance in $\hat{\mu}_i$ (using the μ summary statistic for clarity). Using ChainConsumer (Hinton 2016), we measure a 25 per cent increase in the determinant of the parameter covariance with the corrected likelihood. The resulting posterior distribution is slightly broader, reflecting the added uncertainty in the inference.
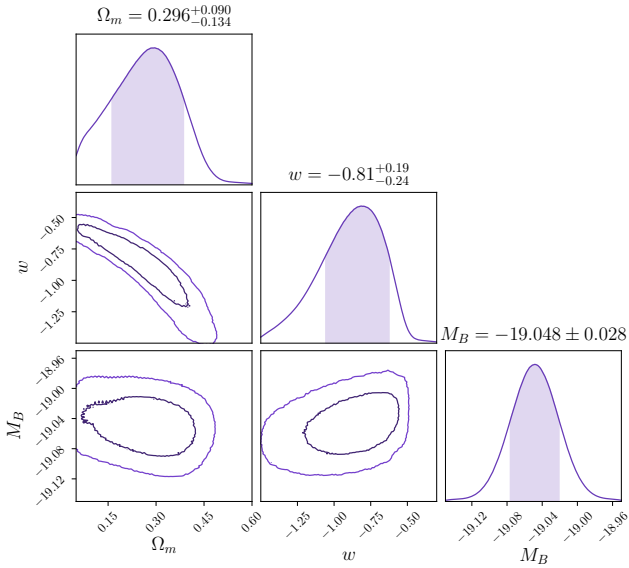
Figure 5. JLA posterior distribution for Ω_m, w, and M_B (described in Section 5.3) using the corrected likelihood (equation 14) with $M = 2$ (we set $\hat{\mu} = \mu$ in this example for clarity). The contours are broader than in figure 4 (25 per cent increase in the parameter covariance determinant) as this likelihood takes into account that the estimated summary statistic is a draw from a sampling distribution.

5.4 Model Comparison

The comparison of different theoretical models using the data in a Bayesian framework is usually done by calculating the Bayesian evidence:

$$P(\mathbf{d}|\text{Model}) = \mathcal{Z} = \int P(\mathbf{d}|\theta, \text{ Model}) \, P(\theta| \text{ Model}) \, d\theta . \quad (23)$$

Two models can be compared by evaluating the Bayes factor:

$$K = \frac{\mathcal{Z}_1}{\mathcal{Z}_2} = \frac{P(\mathbf{d}|\text{Model}_1)}{P(\mathbf{d}|\text{Model}_2)} = \frac{P(\text{Model}_1|\mathbf{d})}{P(\text{Model}_2|\mathbf{d})} \frac{P(\text{Model}_2)}{P(\text{Model}_1)} . \quad (24)$$

If one has no reason to believe a given model more than another *a priori*, then the second factor (the ratio of the prior distributions) equals one. In this case, the Bayes factor becomes a ratio of the model probabilities (conditional on the data).

We evaluate the Bayesian evidence (equation 23) for the three parameter case for the JLA analysis described in Section 5.2 with the uncorrected naive Gaussian likelihood and the corrected likelihood (equation 14) with $M = 2$. Calculating the Bayes factor gives

$$\begin{aligned} \ln K_{\hat{\mu}_1} &= 46 \\ \ln K_{\hat{\mu}_2} &= 31 \\ \ln K_{\hat{\mu}_3} &= 30 , \end{aligned} \quad (25)$$

for $\hat{\mu}_1$, $\hat{\mu}_2$, and $\hat{\mu}_3$ respectively, where

$$\ln K = \ln \mathcal{Z}_{\text{corrected}} - \ln \mathcal{Z}_{\text{uncorrected}} \quad (26)$$

using the corrected likelihood to evaluate the evidence $\mathcal{Z}_{\text{corrected}}$ and the naive, uncorrected likelihood for $\mathcal{Z}_{\text{uncorrected}}$.
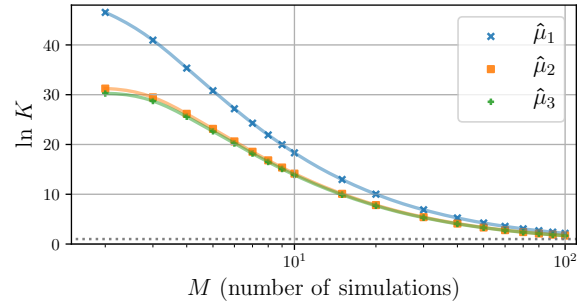

Figure 6. The log Bayes factor $\ln K$ as a function of the number of simulations M for the three estimated summary statistics $\hat{\mu}_i$ (using cubic spline interpolation between evaluated points).

As a check, after increasing the number of evaluated grid points by a factor of 4 we still calculate the same K values. Additionally, we calculate $\ln K$ using a different cosmological parametrisation, sampling scheme and data (Appendix B) and get similar results.

For all three, the corrected likelihood is more than a factor of $\exp[30]$ more probable than the uncorrected. This is further validation of the corrected likelihood; the model (i.e. the corrected likelihood) shows a better goodness-of-fit. Furthermore, if one were using an estimated summary statistic, but not using the corrected likelihood, one's belief in a model would be incorrect by this factor.

This effect would be particularly harmful if comparing two models, where it is possible to calculate μ for the first, but $\hat{\mu}$ is estimated from simulations for the second. Using the same Gaussian likelihood for both, without the correction for the second, would lead one to incorrectly favour the first model.

In figure 6, the log Bayes factors $\ln K$ (equation 25) are shown for the three estimated summary statistics $\hat{\mu}_i$ as a function of the number of simulations M. As the number of simulations increase, the error incurred by using the uncorrected likelihood decreases. The value asymptotes to, but will never reach, $\ln K = 0$. For even a large number of simulations (large M), the error is not negligible. For many analyses, rather than running 10^2 expensive simulations, it would be better to use the corrected likelihood and avoid this error.

6 DISCUSSION & CONCLUSIONS

In this work, we have shown how to take the sampling distribution of estimated summary statistics, $\hat{\mu}$, into account for parameter inference in a cosmological context. For situations where the naive likelihood is Gaussian, we have evaluated this correction (by marginalising over the unknown μ) for the case with known covariance (equation 14) and estimated covariance (equation 19).

We have validated the corrected likelihood with a toy model (Section 4). Using JLA SN Ia data, we have demonstrated the effect of the corrected likelihood on cosmological parameter inference. For model selection, in our simple three-parameter inference demonstration, we show that the log Bayesian evidence $\ln \mathcal{Z}$ will be incorrect with a difference of over 30 if the uncorrected likelihood is not used.

In the era of DES (DES Collaboration et al. 2018),

LSST ([LSST Science Collaboration et al. 2009](#)) and Euclid ([Amendola et al. 2016](#)), cosmological analyses will have access to large cosmological data sets. Sole reliance on 2-point statistics in the linear regime will be tantamount to wasting data which is rich in cosmological information. However, many summary statistics (μ) that access information beyond these 2-point statistics in the linear regime cannot be calculated analytically and need realistic simulations to be run to estimate $\hat{\mu}$.

A typical approach has been to run an excessive number of simulations at each position in parameter space, such that the variance of $\hat{\mu}$ in negligible. This approach has diminishing returns, as variance asymptotically tends to zero as $\frac{1}{M}$. This effectively aims to increase the number of simulations M until the sampling distribution $P(\hat{\mu} \mid \mu)$ can be effectively viewed as Dirac delta function (a limit it will, of course, never reach). Accepting a small increase in the resulting parameter constraints and correcting the likelihood for this sampling distribution means that fewer simulations have to be run.

If one does not wish to take the sampling distribution into account, one might use "cheap" simulations where it is possible to run enough that one effectively reaches the Dirac delta function limit. This has two potential pitfalls: firstly, the limit is never truly reached, which may affect the inferred parameters or model comparison results; secondly, cheap simulations are likely to be less realistic. It is far better to have slightly broader posterior distributions and to have used reliable simulations, than to have tighter constrains on parameters that are biased due to unreliable simulations.

The approach taken in this paper requires the acceptance that simulations are not "free". Simulations are increasingly an essential part of analyses. Like data, reliable simulations are often expensive in terms of time and resources and are, therefore, an acceptable contribution to the uncertainty of inferred parameters.

REFERENCES

Alsing J., Wandelt B., Feeney S., 2018, MNRAS, 477, 2874
Amendola L., et al., 2016, preprint, (arXiv:1606.00180)
Anderson T., 2003, An Introduction to Multivariate Statistical Analysis. Wiley Series in Probability and Statistics, Wiley
Betoule M., et al., 2014, A&A, 568, A22
Bird S., Rogers K. K., Peiris H. V., Verde L., Font-Ribera A., Pontzen A., 2019, Journal of Cosmology and Astro-Particle Physics, 2019, 050
Bookstein F. L., 1989, IEEE Transactions on pattern analysis and machine intelligence, 11, 567
Bromiley P., 2003, Tina-Vision Memo, 3, 1
DES Collaboration et al., 2018, Phys. Rev. D, 98, 043526
Dietrich J. P., Hartlap J., 2010, MNRAS, 402, 1049
Duchon J., 1976, Revue française d'automatique, informatique, recherche opérationnelle. Analyse numérique, 10, 5
Hartlap J., Simon P., Schneider P., 2007, A&A, 464, 399
Heitmann K., Higdon D., White M., Habib S., Williams B. J., Lawrence E., Wagner C., 2009, ApJ, 705, 156
Hinton S. R., 2016, The Journal of Open Source Software, 1, 00045
Jeffrey N., et al., 2018, MNRAS, 479, 2871
Jeffreys H., 1946, Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences, 186, 453
Jeffreys H., 1998, The theory of probability. OUP Oxford
Jennings W. D., Watkinson C. A., Abdalla F. B., McEwen J. D., 2019, MNRAS, 483, 2907
Jones E., Oliphant T., Peterson P., et al., 2001, SciPy: Open source scientific tools for Python, http://www.scipy.org/
Kacprzak T., et al., 2016, MNRAS, 463, 3653
LSST Science Collaboration et al., 2009, preprint, (arXiv:0912.0201)
Leclercq F., 2018, Phys. Rev. D, 98, 063511
Leistedt B., Hogg D. W., Wechsler R. H., DeRose J., 2018, arXiv e-prints, p. arXiv:1807.01391
Martinet N., et al., 2018, MNRAS, 474, 712
Peel A., Lin C.-A., Lanusse F., Leonard A., Starck J.-L., Kilbinger M., 2017, A&A, 599, A79
Petersen K. B., Pedersen M. S., 2012, The Matrix Cookbook, http://www2.imm.dtu.dk/pubdb/p.php?3274
Rollins R. P., 2015, PhD thesis, University College London, http://discovery.ucl.ac.uk/id/eprint/1469967
Scolnic D. M., et al., 2018, ApJ, 859, 101
Sellentin E., Heavens A. F., 2016, MNRAS, 456, L132
Shan H., et al., 2018, MNRAS, 474, 1116
Skilling J., 2004, in Fischer R., Preuss R., Toussaint U. V., eds, American Institute of Physics Conference Series Vol. 735, American Institute of Physics Conference Series. pp 395–405, doi:10.1063/1.1835238
Visser M., 2004, Classical and Quantum Gravity, 21, 2603

APPENDIX A:

Here the full derivation leading to equation 19 is given, starting from the integration equation 18,

$$
P(\mathbf{d}|\{\hat{\mu}_i\}, \{\hat{\mathbf{S}}_i\}) \propto
$$
$$
\iint P(\mathbf{d}|\mu,\Sigma)P(\{\hat{\mathbf{S}}_i\}|\Sigma)P(\{\hat{\mu}_i\}|\mu,\Sigma)P(\Sigma|\theta)P(\mu|\theta) \ \mathrm{d}\Sigma \ \mathrm{d}\mu \tag{A1}
$$
$$
\int \left[\int P(\mathbf{d}|\mu,\Sigma)P(\{\hat{\mu}_i\}|\mu,\Sigma)P(\mu|\theta) \ \mathrm{d}\mu \right] P(\{\hat{\mathbf{S}}_i\}|\Sigma)P(\Sigma|\theta)\mathrm{d}\Sigma \ .
$$

The first factor in the integral, $P(\mathbf{d}|\mu,\Sigma)$, is the Gaussian naive likelihood, given by equation 10. The final factor in the first integral is a uniform prior, $P(\mu) \propto 1$, as described in Section 3.1. The second line uses the fact that $\hat{\mu}$ and $\hat{\mathbf{S}}$ are independent (Anderson 2003). We can first perform the integration over μ, which, using the result from Section 3.1, gives

$$
P(\mathbf{d} \ |\hat{\mu}_i, \ \hat{\mathbf{S}}_i) \propto \int \mathcal{N}(\mathbf{d}, \hat{\mu}_i, \left(\frac{M+1}{M}\right)\Sigma)P(\hat{\mathbf{S}}_i|\Sigma)P(\Sigma|\theta)\mathrm{d}\Sigma \ . \tag{A2}
$$

The second factor is the Wishart distribution

$$
P(\hat{\mathbf{S}}_i|\Sigma, N) = \mathcal{W}(\hat{\mathbf{S}}_i|\Sigma/(N-1), N-1)
$$
$$
= \frac{|\hat{\mathbf{S}}_i|^{\frac{N-p-2}{2}} \exp\left[-\frac{N-1}{2}\mathrm{Tr}[\Sigma^{-1}\hat{\mathbf{S}}_i]\right]}{2^{\frac{p(N-1)}{2}}|\Sigma/(N-1)|^{\frac{N-1}{2}}\Gamma_P\left(\frac{N-1}{2}\right)} \ , \tag{A3}
$$

and the third factor is the Jeffreys prior for Σ

$$
P(\Sigma) \propto |\Sigma|^{-\frac{p+1}{2}} \ , \tag{A4}
$$

as described by Sellentin & Heavens (2016).

With these factors, we can rewrite equation A2, to give

$$
\begin{aligned}
P(\mathbf{d} \,|\hat{\mu}_i, \, \hat{\mathbf{S}}_i) \\
\propto \int d\Sigma \; |\Sigma|^{-\frac{N+p+1}{2}} \; \exp\Big[-\frac{1}{2}\mathrm{Tr}(\Sigma^{-1}\phi_i) \Big] \\
\propto |\phi_i|^{-N/2} \;,
\end{aligned}
\tag{A5}
$$

where

$$
\phi_i = (N-1)\,\hat{\mathbf{S}}_i + \mathbf{v}_i\mathbf{v}_i^{\dagger} \;\;,
\tag{A6}
$$

and

$$
\mathbf{v}_i = (\mathbf{d}-\hat{\mu}_i)\sqrt{\frac{M}{M+1}} \;\;.
\tag{A7}
$$

Using the identity $|\mathbf{A}+\mathbf{v}\mathbf{v}^{\dagger}| = |\mathbf{A}||1+\mathbf{v}^{\dagger}\mathbf{A}^{-1}\mathbf{v}|$ and normalising gives the new corrected likelihood in equation 19. The result is discussed in Section 3.2.

APPENDIX B:

In Section 5.4, the Bayes factors were calculated by evaluating the integral (equation 23) numerically on the parameter grid described in Section 5.2. This resulted in values of $\ln K > 30$ for the corrected vs. uncorrected likelihood using the JLA data with the parameters described. Here we briefly describe how we verified the magnitude of this effect using 1048 Pantheon SN Ia (Scolnic et al. 2018) with a different set of parameters and a different method to evaluate the Bayesian evidence.

We allow four parameters to vary: q_0 (deceleration), j_0 (jerk), M_B, and $h = H_0/[100.0 \text{ km s}^{-1} \text{ Mpc}^{-1}]$. Deceleration and jerk are the parameterisation of a third order Taylor expansion of the scale factor $a(t)$ (Visser 2004). The priors are uniform in the ranges: $-1 < q_0 < 1$, $-2 < j_0 < 2$, $-20 < M_B < -18$, and $0.4 < h < 1.2$. To evaluate the posterior distribution and Bayesian evidence we use PLINY (Rollins 2015), which is a nested sampler (Skilling 2004), and has been shown to be accurate when compared with known closed-form Bayesian evidence results.

The Bayes factor K is also differently defined here, as we set $\mu = \hat{\mu}$, inverting the result. The evaluated value $\ln K \approx -43$, validates our results from Section 5.4.