Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **521,** 1184–1199 (2023) https://doi.org/10.1093/mnras/stac2786
Advance Access publication 2022 November 25

Robust sampling for weak lensing and clustering analyses with the Dark Energy Survey

P. Lemos [1,2]★ N. Weaverdyck [3,4]★ R. P. Rollins,[5] J. Muir,[6] A. Ferté,[7] A. R. Liddle,[8] A. Campos,[9] D. Huterer,[3] M. Raveri,[10] J. Zuntz,[11] E. Di Valentino,[5] X. Fang,[12,13] W. G. Hartley,[14] M. Aguena,[15] S. Allam,[16] J. Annis,[16] E. Bertin,[17,18] S. Bocquet,[19] D. Brooks,[1] D. L. Burke,[20,21] A. Carnero Rosell,[15,22,23] M. Carrasco Kind,[24,25] J. Carretero,[26] F. J. Castander,[27,28] A. Choi,[29] M. Costanzi,[30,31,32] M. Crocce,[27,28] L. N. da Costa,[15,33] M. E. S. Pereira,[34] J. P. Dietrich,[19] S. Everett,[35] I. Ferrero,[36] J. Frieman,[16,37] J. García-Bellido,[38] M. Gatti,[10] E. Gaztanaga,[27,28] D. W. Gerdes,[3,39] D. Gruen,[40,19] R. A. Gruendl,[24,25] J. Gschwend,[15,33] G. Gutierrez,[16] S. R. Hinton,[41] D. L. Hollowood,[35] K. Honscheid,[42,43] D. J. James,[44] K. Kuehn,[45,46] N. Kuropatkin,[16] M. Lima,[15,47] M. March,[10] P. Melchior,[48] F. Menanteau,[24,25] R. Miquel,[26,49] R. Morgan,[50] A. Palmese,[12] F. Paz-Chinchón,[24,51] A. Pieres,[15,33] A. A. Plazas Malagón,[48] A. Porredon,[42,43] E. Sanchez,[52] V. Scarpine,[16] M. Schubnell,[3] S. Serrano,[27,28] I. Sevilla-Noarbe,[52] M. Smith,[53] E. Suchyta,[54] M. E. C. Swanson,[55] G. Tarle,[3] D. Thomas,[56] C. To,[42] T. N. Varga,[19,57] J. Weller[19,57] and (DES Collaboration)

Affiliations are listed at the end of the paper

Accepted 2022 September 22. Received 2022 August 17; in original form 2022 February 18

ABSTRACT

Recent cosmological analyses rely on the ability to accurately sample from high-dimensional posterior distributions. A variety of algorithms have been applied in the field, but justification of the particular sampler choice and settings is often lacking. Here, we investigate three such samplers to motivate and validate the algorithm and settings used for the Dark Energy Survey (DES) analyses of the first 3 yr (Y3) of data from combined measurements of weak lensing and galaxy clustering. We employ the full DES Year 1 likelihood alongside a much faster approximate likelihood, which enables us to assess the outcomes from each sampler choice and demonstrate the robustness of our full results. We find that the ellipsoidal nested sampling algorithm `MULTINEST` reports inconsistent estimates of the Bayesian evidence and somewhat narrower parameter credible intervals than the sliced nested sampling implemented in `POLYCHORD`. We compare the findings from `MULTINEST` and `POLYCHORD` with parameter inference from the Metropolis–Hastings algorithm, finding good agreement. We determine that `POLYCHORD` provides a good balance of speed and robustness *for posterior and evidence estimation*, and recommend different settings for testing purposes and final chains for analyses with DES Y3 data. Our methodology can readily be reproduced to obtain suitable sampler settings for future surveys.

Key words: methods: statistical – cosmological parameters – cosmology: observations – large-scale structure of the Universe.

1 INTRODUCTION

The sampling of a posterior distribution is one of the central elements in current cosmological analyses. With the increasing complexity of cosmological surveys and the large amount of data available, it is a complicated challenge to extract cosmological parameters[1] because of the high dimensionality and complex shapes of the distributions. Nuisance parameters (accounting for various calibration and systematic effects) complicate the analysis by increasing the number of parameters well beyond the six of the standard (Λ cold dark matter) ΛCDM model of cosmology.

Bayesian techniques give a principled framework for probabilistic inference, for instance characterizing information about complex, usually non-Gaussian, posterior distributions for which the mean and standard deviation alone are insufficient to fully describe the

★ E-mail: pablo.lemos.18@ucl.ac.uk (PL); nweaverd@umich.edu (NW)

[1] In this work, we use the term *parameters* to refer to the parameters characterizing a model, both nuisance and cosmological, for which we want to generate samples. We use the term *hyperparameters* to refer to the parameters specific to sampler settings, which affect their performance, such as the number of samples we want to obtain, the stopping criteria, etc. This terminology is common in the machine learning literature. Note that the term *hyperparameters* can refer to different concepts, even in the field of cosmology (Lahav et al. 2000; Hobson, Bridle & Lahav 2002; Luis Bernal & Peacock 2018).

shape of the distribution. Markov Chain Monte Carlo (MCMC) methods have traditionally been used for this purpose (Metropolis et al. 1953; Neal 1997), and have a long history of applications in cosmology (e.g. Christensen et al. 2001; Knox, Christensen & Skordis 2001; Lewis & Bridle 2002; Verde et al. 2003; Tegmark et al. 2004; Dunkley et al. 2005; Shaw, Bridges & Hobson 2007). However, for some applications (such as model comparison and the comparison of different data sets) it is necessary to calculate not only the shape of the posterior distribution but also the Bayesian evidence. Nested Sampling (Skilling 2006) is the method most commonly used for this purpose, because of its speed and its ability to obtain both the Bayesian evidence and the posterior distribution in the same calculation.

Because of their wide applicability, many tools have been developed to implement these sampling algorithms given a user-defined likelihood, and the choice to use one over another may be more driven by accessibility and ease of implementation than rigorous testing for the specific analysis at hand.

As the constraining power of cosmological data sets has grown, different analyses have begun to diverge in their inferred parameter posteriors. Perhaps most famous is the discrepancy in the measurements of H_0 by the Planck Collaboration (2020) versus that obtained via distance ladder measurements (Riess et al. 2022), but there exists also tension between measurements of $S_8 \equiv \sigma_8(\Omega_m/0.3)^{0.5}$ from large-scale structure probes and that inferred by *Planck* under ΛCDM (see Di Valentino et al. 2021a, b; Shah, Lemos & Lahav 2021, for reviews on these tensions). As these discrepancies could be indicators of new physics, it is vital that the inferences upon which they are based are robust to analysis choices such as the specific sampler and settings used.

Most samplers include hyperparameters that allow one to tune the algorithm and, in the limit of infinite computing time and resources, allow one to obtain arbitrarily precise constraints. In practice, we require a balance of speed and accuracy, where it is feasible to run a large number of chains but the error introduced by the sampler is a negligible (or at least quantifiable) contribution to the analysis' error budget. This is particularly true for the Dark Energy Survey (DES, The Dark Energy Survey Collaboration 2005) combined weak lensing and galaxy clustering cosmology analysis (henceforth, 3 × 2 pt), where the complexity of the data and analysis pipeline results in the need to run a large number of chains for validation purposes.

In this work, we perform a careful investigation of several leading sampling algorithms available within the `CosmoSIS` analysis framework. We focus on `POLYCHORD` and `MULTINEST` because of their ability to estimate the Bayesian evidence, and calculation of model comparison and data set tension statistics are of particular interest for the DES Y3 analysis. We investigate how hyperparameters impact performance and focus particularly on avoiding biases in the parameter constraints and evidence, which could lead to mistaken interpretation of the core analysis results. We make recommendations for the sampler and settings for three different use cases of the DES Y3 data, which strike different balances of speed and accuracy.

There have been previous attempts at characterizing sampler performance. For example, Allison & Dunkley (2014) compared MCMC and Nested Sampling methods, and Higson et al. (2019) developed diagnostic methods to assess errors from Nested Sampling chains, including the use of bootstrapping individual chain samples to assess uncertainty. We use some of these tools, but assess uncertainty using full independent chain realizations run over a wide range of parameter settings, and using high-resolution chains as benchmarks. We combine tests on the first year of DES data (DES Y1) and on the results of a fast, approximate version of the likelihood that allows us to generate a large number of sampling runs under the same hyperparameter settings.

The paper is structured as follows: In Section 2, we introduce the methodology and notation of Bayesian parameter estimation, as well as the summary statistics that we will use throughout this work. In Section 3, we present the methodology and data used in this work. Our results are shown in Section 4, and we present our conclusions in Section 5. All the data produced from this work are available upon request.

2 SAMPLERS

This section describes the formalism of parameter estimation in a Bayesian framework, as well as the three different sampling algorithms employed in this work. Detailed descriptions of the formalism can be found for example in MacKay (2002) and Sivia & Skilling (2006).

2.1 The Bayesian framework

In parameter estimation we have obtained some data D; we have assumed a theoretical model M, and we seek an estimate of the parameters θ of the model. This is accomplished by applying Bayes' theorem

$$P(\theta|D, M) = \frac{P(D|\theta, M) \times P(\theta|M)}{P(D|M)}. \tag{1}$$

The quantities in this equation are usually labelled as

$$\mathcal{P} = \frac{\mathcal{L} \times \Pi}{\mathcal{Z}}, \tag{2}$$

where $\mathcal{P}$ is the posterior, $\mathcal{L}$ the likelihood, Π the prior, and $\mathcal{Z}$ the marginal likelihood or Bayesian evidence. The latter can be expressed as

$$\mathcal{Z} = \int \mathcal{L} \times \Pi \, \mathrm{d}\theta. \tag{3}$$

This is typically a complicated and high-dimensional integral. Because $\mathcal{Z}$ acts as a normalizing factor that does not depend on the parameters, it often plays no role for parameter estimation. There are, however, other applications where the Bayesian evidence is fundamental; one such case is Bayesian model comparison. Here, we have two competing theoretical models M_A and M_B and we want to know which of these models is preferred given some measured data D. For this we calculate the ratio

$$\frac{P(M_A|D)}{P(M_B|D)} = \frac{P(D|M_A)}{P(D|M_B)} \times \frac{P(M_A)}{P(M_B)}, \tag{4}$$

where the equality follows from Bayes' theorem. The second factor on the right-hand side is the ratio of the prior beliefs in the two models. If there is no prior reason to prefer one model over the other, then this term is unity and hence disappears. The first factor on the right-hand side is the ratio of the Bayesian pieces of evidence for the two models. Therefore, under the assumption of equal prior beliefs in the two models, we can find which model is preferred by the data by calculating the ratio

$$R \equiv \frac{\mathcal{Z}_A}{\mathcal{Z}_B}. \tag{5}$$

This quantity is called the Bayes factor. Bayesian pieces of evidence are also used to quantify tension between different data sets (Marshall, Rajguru & Slosar 2006).

In addition to the Bayesian evidence of equation (3), we will compute two more summary statistics from our chains, which contain important information about our problem. The first one is the Kullback–Leibler divergence (Kullback & Leibler 1951), given by

$$\mathcal{D}_{KL} = \int \mathcal{P}(\theta) \log \frac{\mathcal{P}(\theta)}{\pi(\theta)} \, \mathrm{d}\theta. \tag{6}$$

The Kullback–Leibler divergence measures the information gain when going from the prior to the posterior distribution, measured in natural bits, or *nats*. The Kullback–Leibler divergence can be used amongst other things to calculate the *information* between two data sets, which in turn can be used to calculate the *Suspiciousness* (Handley & Lemos 2019b), and quantify the concordance between the data sets in a way that does not depend on prior volumes.

Our last summary statistic is the Bayesian Model Dimensionality (henceforth BMD), which provides an estimate of how many Gaussian dimensions are constrained by our data:

$$d = 2 \int \mathcal{P}(\theta) \left(\log \frac{\mathcal{P}(\theta)}{\pi(\theta)} - \mathcal{D}_{KL} \right)^2 \mathrm{d}\theta. \tag{7}$$

Handley & Lemos (2019a) discuss the advantages of characterizing dimensionality via the BMD as opposed to other commonly used measures like the Bayesian Model Complexity (Spiegelhalter et al. 2002), such as not relying on a single-point estimator. Another advantage of the BMD is that it can be computed directly from both nested sampling and MCMC chains as

$$\frac{\tilde{d}}{2} = \left\langle (\log \mathcal{L})^2 \right\rangle_{\mathcal{P}} - \langle \log \mathcal{L} \rangle_{\mathcal{P}}^2, \tag{8}$$

where $\langle \cdot \rangle_{\mathcal{P}}$ indicates an average over the posterior distribution.

2.2 Metropolis–Hastings

MCMC is one of the most widely used methods for sampling probability distributions. It consists of using chains in which each element depends only on the previous one, known as Markov Chains, to obtain samples from the target distribution. The Metropolis–Hastings algorithm (Hastings 1970) (denoted MH in the following) is a common MCMC method, widely used in various fields such as statistical mechanics or as here, Bayesian inference. Here, we use MH in order to generate samples from the posterior distribution of the cosmological and nuisance parameters. Next, we describe the fundamental aspects of MCMC algorithms in general, and MH in particular, as well as details of its implementation within this work.

Note that we include the MH sampler primarily as a benchmark against which we can compare the parameter estimation results of the nested samplers that are the main focus of this work. We have not made a significant effort to optimize the MH sampler's speed and performance, so a fair assessment of its computational cost compared to POLYCHORD and MULTINEST is beyond the scope of this paper.

2.2.1 *The Metropolis–Hastings algorithm*

The goal of MCMC algorithms is to return samples from a distribution that converges towards a unique stationary distribution $\pi(\theta)$ (where θ are the cosmological and nuisance parameters) of the target distribution, in this case the posterior $\mathcal{P}(\theta)$. Given the transition matrix p_{ij} of a Markov chain, which corresponds to the probability of moving from state i at time t to state j at time $t+1$, we thus have

$$\pi_j = \sum_j p_{ij} \pi_i. \tag{9}$$

We now need to construct such a transition matrix.

The MH algorithm proposed in Hastings (1970) does so by requiring p_{ij} and π to satisfy the so-called detailed balance

$$\pi_i \, p_{ij} = \pi_j \, p_{ji}. \tag{10}$$

In MH, the transition matrix is then defined as

$$p_{ij} = q_{ij} \alpha_{ij}, \tag{11}$$

where q_{ij} is the proposal distribution (corresponding to a proposed 'jump' in parameter space) and α_{ij} the acceptance distribution (corresponding to accepting this 'jump' or not), defined as

$$\alpha_{ij} = \min \left(1, \frac{p_j q_{ji}}{p_i q_{ij}} \right). \tag{12}$$

If a chain of samples is selected using this algorithm for a large number of steps, the density of their resulting distribution will follow the target distribution, i.e. the posterior $\mathcal{P}(\theta)$.

Depending on the initial point in parameter space and the provided proposal, some samples drawn at the beginning of this process should be discarded as they are not representative of the posterior distribution. This period is called burn-in, where the accepted points may be far from the peak of the posterior, and can be minimized if starting at a point in parameter space closer to the best-fitting value (see Hogg & Foreman-Mackey 2018 for a discussion on choices to limit the burn-in period). It can be explored by plotting the posterior or parameter values as a function of step number (or overplotting these values from chains that started at different points), where the burn-in corresponds to samples before these values converge around the typical set.

One potential way of speeding up MH algorithms often used in cosmology is to take advantage of the fact that some parameters, known as 'fast', do not affect the slowest parts of the likelihood calculation, which in the case of cosmology often involve the transfer function or line-of-sight integration. These parameters can be decorrelated and sampled separately, making sampling nearly as fast as it would for the 'slow' parameters alone (Lewis 2013). When fast and slow parameters cannot be fully decorrelated in principle, they can be sampled using 'dragging' (Neal 2005), which consists of 'dragging' the fast parameters while keeping the slow ones fixed, leading to fast likelihood evaluations. Both of these methods are implemented in the COBAYA package (Torrado & Lewis 2021).

One of the difficulties of using MCMC algorithms such as MH is the lack of definitive criteria ensuring the chain has converged towards the target distribution. Several criteria for testing the convergence have been proposed (An, Brooks & Gelman 1998; Sinharay 2003). It is also useful to study the autocorrelation of the MH chains to verify that the samples are independent on scales much smaller than the chain length. In the following text, we will mainly use the Gelman–Rubin diagnostic which is derived from the method proposed in Gelman & Rubin (1992) to monitor the convergence of MH chains. This diagnostic works by comparing parameter estimates from a number of independent chains. Specifically, adopting the standard notation, it estimates the potential scale reduction factor $\hat{R}$ for a given parameter θ, defined as

$$\hat{R} = \frac{\hat{V}}{W}, \tag{13}$$

where $\hat{V}$ is the estimator of the variance of the parameter and W is the average of the variance of θ within a chain (in the above expression, the impact of degrees of freedom defined in Gelman & Rubin 1992 is neglected). $\hat{R} \simeq 1$ implies that the distribution of the sampled parameter is close to stationary; while this does not

guarantee that the chain has converged, it is a good indication of convergence for unimodal posteriors when $\hat{R}$ is nearing 1 for all parameters. A typical convergence criterion is to stop when $\hat{R} - 1 < 0.1$. When considering M independent chains, the variance estimator $\hat{V}$ is defined as

$$\hat{V} = \frac{N-1}{N} W + \frac{M+1}{MN} B, \tag{14}$$

where N is the length of the chains and B/N is an estimate of the variance of the parameter mean $\bar{\theta}$ across chains i.e.

$$B/N = \frac{1}{M-1} \sum_{i=1}^{M} (\bar{\theta}_i - \bar{\theta})^2. \tag{15}$$

2.2.2 *Implementation*

For this study we use a simple version of the MH sampler implemented within CosmoSIS. In the configuration used here the MH sampler uses a fast–slow scheme in which each parameter subspace uses a separate multivariate Gaussian proposal, with a one-third chance of each proposed jump length being drawn instead from an exponential distribution, to better explore parameter tails. We oversampled the fast subspace by a factor of 5 and have nine fast parameters. In typical use of this sampler, the proposal is initially set to an estimate of parameter covariances, then tuned at the start of the chain. During tuning, the estimated parameter covariances are replaced with those computed from the points sampled in the chain up to that point.

For this particular study we set the initial proposal using a parameter covariance extracted from a finished high-quality POLYCHORD chain, and because that was expected to be close to the target distribution, we did not tune the proposal. This choice was motivated primarily by simplicity, in that it allowed us to use the MH sampler without adjusting its hyperparameters. Variations of this setting could have been used, namely using a more approximate initial proposal estimate – for example, using only the diagonal part of the parameter covariance – and then tuning the proposal. These choices would be expected to produce the same posterior estimate, just over a longer period of time.

Note that using the MH sampler requires sampling a scaled version of the primordial power spectrum amplitude, $10^9 A_S$. This is because the relative size of the unscaled A_S values compared to other parameters is small, which causes the proposal covariance matrix to be ill-conditioned.

2.3 MultiNest

Multinest is an example of a nested sampling algorithm (Skilling 2006) which, in contrast to MCMC samplers like MH, can be used to calculate the Bayesian evidence in addition to estimating the posterior. Instead of selecting individual samples sequentially, nested sampling starts with a large number of points (called 'live' points), and then repeatedly selects the live point with the smallest value of the posterior density, eliminates it (turning it into a 'dead point'), and then finds a new replacement live point with a posterior value larger than that of the point that was eliminated. The collection of all points (live and dead) can then be used to calculate the evidence while also serving as a (weighted) sample of the posterior. The most difficult part of Nested Sampling is finding new live points. It is extremely inefficient simply to randomly generate points until one with a higher posterior value is found (especially when most live points are close to the maximum of the posterior and when the problem has high dimensionality). This is the challenge that specific algorithms like MULTINEST are designed to address.

MULTINEST[2] is a publicly available code for Nested Sampling (Feroz, Hobson & Bridges 2009; Feroz et al. 2019). It has been extensively used for cosmology analyses, including that of the first year (Y1) of DES data (Abbott et al. 2018). MULTINEST uses a technique called ellipsoidal sampling (Mukherjee, Parkinson & Liddle 2006b), where it calculates a D-dimensional ellipsoid from current set of live points, and finds the next point within that ellipsoid, expanded by a certain factor. MULTINEST also includes a clustering algorithm to identify multiple peaks in the posterior distribution, allowing it to sample multimodal posteriors. This was its main improvement over the ellipsoidal nested sampling code COSMONEST (Mukherjee et al. 2006a, b; Pahud et al. 2006; Parkinson, Mukherjee & Liddle 2006). There are other examples of ellipsoidal Nested Sampling algorithms, such as NESTLE[3] and DYNESTY[4] (Speagle 2020), which uses dynamic sampling while still relying on ellipsoidal nested sampling.

As described in Skilling (2006), the standard Nested Sampling approach calculates the pieces of evidence using the accepted samples, and using an approximation for the distribution of sampling weights. In addition to this calculation, MULTINEST produces an alternative calculation of the Bayesian evidence using Importance Nested Sampling (henceforth INS). INS, first introduced in the context of Nested Sampling in Cameron & Pettitt (2014), uses all likelihood evaluations to estimate the evidence, instead of using only the accepted points in the MULTINEST run (which in some cases has acceptance rates as low as ~ 1 per cent). In an ideal case, both estimates of the evidence should agree. In this work, when we refer to the MULTINEST evidence, we are referencing the 'default' evidence calculation, and we will explicitly make reference to the INS evidence when that is not the case.

While ellipsoidal nested sampling leads to fast sampling, it can also lead to biases in both the posterior and the evidence estimation, as discussed later in the paper. This is illustrated in Fig. 1: If the ellipsoid is not expanded enough, the calculation of the evidence will 'miss' parts of the distribution. These issues are discussed using MULTINEST as an example, but apply to any implementation of ellipsoidal nested sampling.

2.4 PolyChord

An alternative code for Nested Sampling is POLYCHORD (Handley, Hobson & Lasenby 2015a, b).[5] The difference between this algorithm and MultiNest is in the approach to generating new live points. Instead of the ellipsoidal sampling, it uses so-called slice sampling (Aitken & Akman 2013), where new live points are generated by taking a random slice through the parameter space that includes the current live point, and randomly generating new points until one with higher likelihood is found. The process is then repeated with the new point and a slice in a new random direction, for a user-defined number of repetitions ($n_{repeats}$) until the candidate live point is sufficiently uncorrelated with the initial live point. In practice, the sample covariance of existing points is used to decorrelate and whiten the parameter space, such that slices are performed on an affine transformation of parameter space where the relevant likelihood

[2] https://github.com/farhanferoz/MultiNest
[3] http://kylebarbary.com/nestle/index.html
[4] https://dynesty.readthedocs.io/en/latest/
[5] https://github.com/PolyChord/PolyChordLite

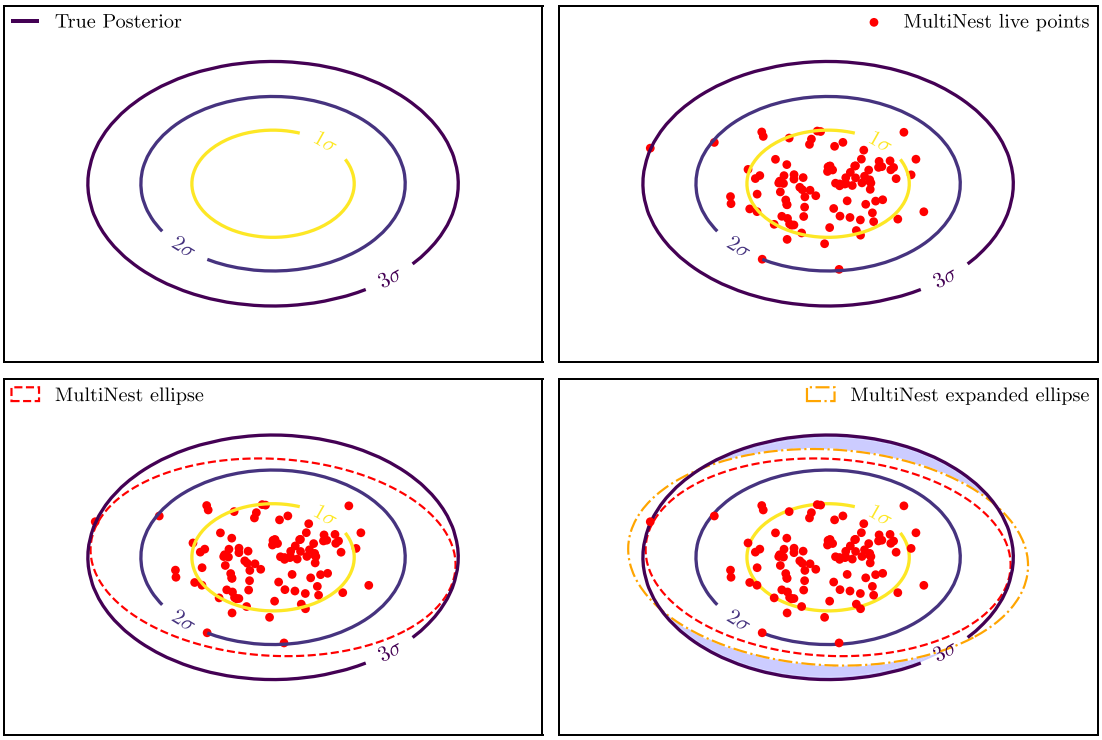

Figure 1. An example of MULTINEST's ellipsoidal sampling, and how it can lead to biases. When trying to sample a certain distribution (top left), MULTINEST randomly generates some points (top right). It then uses the covariance matrix obtained from those points to calculate an ellipsoid enclosing all existing live points (bottom left, dashed line). That ellipsoid is expanded in volume by a factor inversely proportional to the efficiency, and samples are drawn from that ellipsoid (bottom right, dot–dashed line). As the latter plot shows in the light blue regions, if the magnification factor is not big enough (i.e. the efficiency is too high), this can lead to a bias in the estimation of the evidence.

width is $\mathcal{O}(1)$ in each direction. This both simplifies and accelerates the generation of new samples.

Like MULTINEST, POLYCHORD has a clustering algorithm which allows it to sample multimodal posterior distributions. In addition, POLYCHORD is compatible with the fast–slow parameter implementation used by the code COSMOMC (Lewis & Bridle 2002; Lewis 2013), which provides a significant increase in speed for cosmological likelihoods. While it is slower than MULTINEST in obtaining posterior distributions and Bayesian pieces of evidence for the models studied here, we will show that it more reliably gives unbiased results.

2.5 Other samplers

In this work, we focus on three sampling algorithms commonly used in cosmology (Metropolis–Hastings, ellipsoidal nested sampling, and slice nested sampling). One common sampler that we do not implement is EMCEE, which is an affine-invariant MCMC sampler that uses an ensemble of walkers to traverse the posterior and update the proposal distribution before applying standard Metropolis–Hastings acceptance criteria (Foreman-Mackey et al. 2013). We found that in the large-dimensional parameter spaces tested here, the samples generated by EMCEE had high enough levels of correlation so as to require intractable runtimes. Coupled with the inability to apply convergence criteria like the Gelman–Rubin statistic to correlated walkers and the large amount of samples that need to be discarded as burn-in, we decided not to include it in this study.

There exist other algorithms that, while perhaps not yet as widely used in cosmology, could become more common in the future. `Zeus`[6] (Karamanis, Beutler & Peacock 2021) is an implementation of ensemble slice sampling (Karamanis & Beutler 2021) for MCMC, and has the advantage of not requiring tuning of any hyperparameters, thus providing a promising alternative to traditional MCMC algorithms. We provide a more detailed discussion of `Zeus`, and compare it to EMCEE, MULTINEST, and POLYCHORD in Section A. Other algorithms such as Hamiltonian Monte Carlo (Betancourt 2017) or the No-U-Turn Sampler algorithm (Hoffman & Gelman 2014) have existed for some time, but require accurate derivatives, which cannot be accessed easily in current cosmological theory codes such as CAMB.

3 METHODOLOGY

The goal of this paper is to compare the performance of the previously introduced methods for cosmological analysis. In cosmology, we usually perform inference with about six to eight cosmological parameters, and a number of nuisance parameters used to model systematic uncertainties. The nuisance parameters are usually marginalized over for cosmological constraints, though they may also be interesting in their own right (e.g. constraining galaxy bias or the amplitude of intrinsic alignment of galaxies). Here, we use the pipeline for the DES Year 1 3 × 2 pt analysis, which has 20 nuisance parameters. We assume a wCDM cosmological model, which allows for a varying equation of state for dark energy. We therefore constrain seven cosmological parameters: $\{\Omega_m, \Omega_b, h, n_s, A_s, w, \Omega_\nu h^2\}$, giving a total of 27 parameters to be sampled.

[6] https://zeus-mcmc.readthedocs.io/en/latest/

In practice, we use two different pipelines in our analysis. We use the public DES Y1 3 × 2 pt likelihood implemented in the cosmological parameter estimation code COSMOSIS (Zuntz et al. 2015), which includes all the samplers used in this work. In addition, we also use a *fast likelihood* that employs several approximations to reduce the evaluation time by a factor of ∼50. Both of these pipelines are described below.

3.1 Fast likelihood analysis

The sampling methods described in this work can be slow, and in some cases we can only understand the effects of tuning different hyperparameters by repeating the sampling a large number of times. For that purpose, we generated a *fast likelihood*, which produces posterior distributions that are similar to those of the DES Y1 pipeline, but uses multiple approximations to significantly reduce the run time. The resulting likelihood is an approximation to the true likelihood that allows for a large number of chains to be run and thus for the variance of samplers to be characterized. It can be considered a toy model that is substantially more applicable to our use case than the analytic models (e.g. Gaussian mixture models) that are often employed to characterize sampler behaviours.

The primary changes in the fast likelihood are:

(i) Using the fitting function presented in Eisenstein & Hu (1998) for the transfer function when computing the linear matter power spectrum;

(ii) Acceleration of the calculation of the Halofit non-linear scale (Equation A4 in Takahashi et al. 2012) using a non-iterative interpolation-based root-finding algorithm and trapezoidal integration;

(iii) Calculation of the lensing efficiencies and Limber angular correlation functions (Equations IV.3–IV.6 in Abbott et al. 2018) using a simplified trapezoidal integration scheme.

3.2 Application to DES Y1 data

We apply all the samplers described above to the DES 3 × 2 pt analysis, running MULTINEST and POLYCHORD with a large number of different hyperparameter settings. Because the bulk of the work presented in this paper was performed while the analysis pipeline for the recently released Y3 analysis (Abbott et al. 2022) was being developed, these tests are run using the DES Y1 data (Abbott et al. 2018) and the Y1 version of the DES modelling pipeline. These data consist of a combination of three two-point correlation function measurements: cosmic shear, galaxy–galaxy lensing, and galaxy clustering.

There are mainly two purposes to this paper: to find sampler settings that yield unbiased results for the DES analysis while minimizing the running time, and to generally understand the causes of bias in the parameter estimation or evidence calculation. The results presented in this work depend heavily on the dimensionality of the likelihood, as well as the form of the likelihood, and so cannot be generalized to all sampling problems. However, as most cosmological sampling problems have similar dimensionality and characteristics, these results should still be useful in guiding sampler choices in future cosmological analyses.

4 RESULTS

In this work, we have explored different sampling settings, to compare their performance and run time. Unless stated otherwise, all runs use the same likelihood, priors, and data, and are run using the same computing platform (the Cori system at NERSC) and with the same number of nodes.

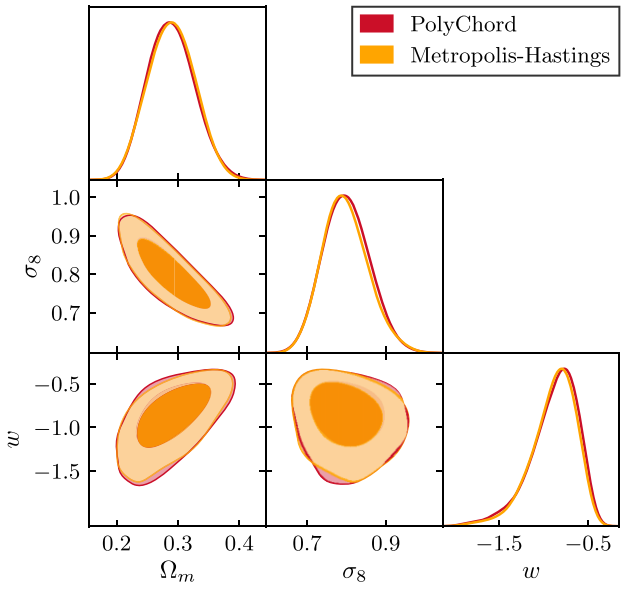

Figure 2. Posterior distribution for a high-quality Nested Sampling run (red) and a Metropolis–Hastings run (yellow) which uses a full proposal covariance.

4.1 Posterior validation with Metropolis–Hastings

MCMC methods are expected to produce more reliable posteriors than Nested Sampling, because their convergence criterion is based on the posterior, not on the Bayesian evidence (which is difficult to estimate well from standard MCMC chains). Given this, before looking in detail at the effects of hyperparameters, we compare constraint contours from MH and POLYCHORD in order to benchmark the accuracy of the nested sampling posterior estimates.

We run eight MH chains in parallel using four processors per chains, spread across two nodes. We stopped the chains once $\hat{R} - 1 < 0.02$ for all parameters (see Section 2.2 for a description of the Gelman–Rubin statistic $\hat{R}$), amounting to 762 000 samples. We burn the first 20 per cent of the chains, as described in Section 2.2. Fig. 2 shows the posterior estimated with these two sets of chains on the cosmological parameters w, Ω_m, and σ_8 along with the posterior estimated from a high-quality POLYCHORD chain.

We note that our MH run was slower than most nested sampling runs, using around 4600 CPU-hours (6 d of walltime). Nested sampling is known to scale better with dimensionality, so this is on some level expected. However, we emphasize that this is not necessarily a fair comparison because, as was discussed in Section 2.2, our MH runs did not employ a number of speed-up techniques which would likely be used in practice if MH were being used as the main sampler in an analysis. This is fine for our purposes, because as noted above we are using the MH chains to compare posterior distributions, not runtime.

We interpret the good agreement observed in Fig. 2 between MH and the high-quality POLYCHORD chain as confirmation that with good enough settings, Nested sampling can accurately sample the posterior. We then explore what these settings need to be for both MULTINEST and POLYCHORD.

4.2 MULTINEST

MULTINEST has several hyperparameters that can be tuned. These changes can increase accuracy in computing different quantities, at the expense of computing time. Table 1 shows timing and summary statistic results[7] of running MULTINEST chains with the DES likelihood using a variety of different choices for the sampler's hyperparameters. We briefly describe these MULTINEST hyperparameters (for more details see Feroz et al. 2009, 2019, henceforth F08, F13, respectively), and how they affect the sampler performance.

(i) n_{live}: the number of live points. This quantifies how many points are used to sample the posterior and is proportional to the expected final number of samples in the chain. For the full likelihood we compare two different values: $n_{live} = 250$ and $n_{live} = 675$. The latter value is based on the fiducial choice of $n_{live} = 25 * D$, where D is the number of parameters being sampled.
A higher number of live points increases the accuracy of the estimated posterior distributions and Bayesian evidence; we find that increasing n_{live} decreases the uncertainty in $\log Z$ by a factor $\sqrt{\Delta \mathrm{n_{live}}}$, and increases run time linearly.

(ii) $efficiency$: a MULTINEST-specific hyperparameter that controls the size of the ellipsoids used by MULTINEST to search for new live points. To find the next live point after every step, MULTINEST uses the covariance of the existing live points to create an N-dimensional ellipsoid, then expands the ellipsoid by a factor of $1/efficiency$ before using it to find the next live point.[8]This procedure is illustrated in Fig. 1.
As previously explained, this figure also shows a potential weakness of MULTINEST: we see that the expanded ellipsoid, shown with the dash–dotted orange line in the lower right-hand panel, is not capturing part of the tails of the true posterior distribution, shown in shaded blue. These regions will not be sampled, or considered when calculating the Bayesian evidence. This missing-posterior-tail bias will be more severe for higher values of efficiency, a finding that is reflected in the results shown in Table 1, Fig. 3, and Fig. 4. When the efficiency is too high, all the summary statistics calculated in this work are systematically wrong. Even for efficiencies as low as 10^{-3}, we find a lack of convergence in summary statistics and disagreement with the best POLYCHORD values.
This bias can be reproduced using a 27-dimensional Gaussian posterior distribution, which has a known true evidence, as shown in Fig. 5. Comparisons to this Gaussian toy model also show that POLYCHORD gets more reliable evidence estimates, motivating us to adopt the best POLYCHORD run on the DES likelihood as a benchmark 'correct value'.
We also see in Table 1 that there is an approximately power-law relation between runtime and efficiency. Therefore, it can become extremely computationally expensive to achieve a low enough value of efficiency to obtain unbiased evidence estimates with MULTINEST. The importance of the efficiency hyperparameter for MULTINEST presents another challenge in that there is no principled way of knowing what value of the efficiency should be used, or if the value used was low enough, without running the algorithm multiple times.

[7]Reported uncertainties are via bootstrap resampling as computed by the ANESTHETIC software package Handley (2019).
[8]The efficiency is thus a rough estimate of the acceptance rate, the probability that a point sampled from the expanded ellipsoid will have a higher likelihood than the point needing replacement. However, the algorithm's acceptance scales better than this due to the ellipsoid hitting the prior boundaries in some of the parameter directions, as indicated by the fact that the BMD is less than the number of sampled parameters.

Note that biases from high efficiencies have a less severe impact on marginalized posteriors than on estimates of the Bayesian evidence. We can see this in Fig. 3. While higher efficiency does cause the sampler to miss the tails of the distribution, even for very high values those missing tails are unlikely to significantly affect interpretation of the contours. Thus, if we are only interested in the posterior distributions, we do not need to use efficiencies as low as would be needed if we wanted to compute the Bayesian evidence.

(iii) **Tolerance**: the stopping criterion. Both MULTINEST and POLYCHORD can estimate how much the existing live points will contribute to the estimate of the evidence. When that contribution is smaller than the chosen value of the tolerance, the algorithm terminates. One can check whether the tolerance is low enough by plotting the progression of the weights of the chain, as shown in Fig. 6. If the tolerance is low enough, this plot will show a peak that reaches unity, and will then decay back towards zero. A spike at the end shows that the contribution to the evidence from the final set of live points is too high, and the tolerance should be decreased.
Table 1 shows that tolerance does not have a significant impact on either run-time or on summary statistics. Because of this, and because a chain initially run with higher tolerance can be resumed to reach a lower tolerance, the choice of this parameter is not considered a challenge: we simply recommend a tolerance that ensures that weights look similar to those on the right-hand panel of Fig. 6.

(iv) OMP**threads**: MULTINEST in COSMOSIS uses a double parallelization scheme: The Boltzmann solver CAMB (Lewis, Challinor & Lasenby 2000; Howlett et al. 2012) is parallelized using OPENMP, and the MULTINEST sampling algorithm uses MPI parallelization. We tested two settings, both using the same number of nodes, but changing the number of cores used on each type of parallelization. We find that not using the OPENMP parallelization greatly improves the sampling speed. We expected this, as MULTINEST will be faster when all cores are used by MPI parallelization, up to the number of live points. As expected, changing this setting does not affect the results in any way apart from the run time.

(v) **Constant Efficiency**: MULTINEST can use a different sampling method, called 'constant efficiency' mode. In this setting, we abandon the strategy of increasing the volume of the ellipse by a factor of $1/efficiency$. Instead, the increase in the size of the ellipses changes at every step to match the input 'constant efficiency' value in the sampling efficiency (i.e. the ratio of points accepted to points sampled). F13 describe how:
'Despite the increased chances of the fitted ellipsoids encroaching within the constrained likelihood volume (i.e. missing regions of parameter space for which $\mathcal{L} > \mathcal{L}_i$), past experience has shown (e.g. F08) this constant efficiency mode may nevertheless produce reasonably accurate posterior distributions for parameter estimation purposes.'
Our results agree with these statements in F13, with some caveats. Indeed, for efficiencies set to values of 0.3 and 0.1, constant efficiency mode produces significantly quicker runtimes and worse estimates of the evidence and other summary statistics than the standard mode. However, at lower efficiency values we find that this trend is inverted. For example, when the 'constant efficiency' hyperparameter is set to 10^{-3}, the constant efficiency runtime becomes longer than standard MULTINEST, and the evidence estimate also appears to converge to the correct value. However, given its longer runtime at efficiencies needed for accurate pieces of evidence, we do not recommend using 'constant efficiency' to estimate the evidence.

As previously discussed, MULTINEST produces an alternative INS evidence estimate. By examining Table 1 we can compare

Table 1. Comparison of time required and output values for MULTINEST with different settings. All runs use the DES Y1 3 × 2 pt likelihood and a wCDM cosmology, with 128 cores. The settings $OMPthreads = 1$ corresponds to 128 MPI threads, while $OMPthreads = 4$ corresponds to 32 MPI threads. CE refers to 'Constant efficiency'. L Evals is the number of Likelihood evaluations. INS refers to the Importance Nested Sampling reported by MULTINEST, D_{KL} is the Kullback–Leibler divergence, and BMD is the Bayesian Model Dimensionality. Reported uncertainties are via bootstrap resampling as computed by ANESTHETIC (Handley 2019).

n_{live}	Eff	Tol	OMP	CE	Time (h)	Acceptance	L Evals	log Z	INS log Z	D_{KL}	BMD
675	1	0.3	1	F	11.2	0.067	265 314	-277.71 ± 0.17	-285.14 ± 0.05	18.54 ± 0.16	13.0 ± 0.3
675	1	0.1	1	F	13.8	0.056	340 783	-277.68 ± 0.17	-284.93 ± 0.22	18.61 ± 0.15	13.5 ± 0.4
675	1	0.01	1	F	20.3	0.042	510 583	-277.63 ± 0.17	-284.94 ± 0.14	18.42 ± 0.15	13.0 ± 0.4
675	1	0.1	4	F	46.3	0.053	356 248	-277.53 ± 0.17	-285.09 ± 0.06	18.33 ± 0.16	12.8 ± 0.4
675	1	0.1	1	T	3.8	0.228	78 561	-275.41 ± 0.16	-286.36 ± 0.19	17.59 ± 0.15	12.3 ± 0.3
250	0.3	0.1	1	F	4.3	0.053	134 554	-278.68 ± 0.28	-285.46 ± 0.27	19.32 ± 0.28	13.3 ± 0.6
675	0.3	0.3	1	F	14.6	0.054	346 242	-278.52 ± 0.17	-285.01 ± 0.14	19.16 ± 0.16	13.5 ± 0.4
675	0.3	0.1	1	F	16.2	0.048	396 284	-278.37 ± 0.17	-285.13 ± 0.10	19.14 ± 0.17	12.6 ± 0.4
675	0.3	0.01	1	F	21.4	0.040	531 284	-278.53 ± 0.17	-284.63 ± 0.28	19.18 ± 0.16	13.7 ± 0.4
675	0.3	0.1	4	F	50.3	0.050	383 602	-278.45 ± 0.17	-285.13 ± 0.04	19.21 ± 0.16	13.0 ± 0.3
675	0.3	0.1	1	T	4.6	0.199	94 746	-276.56 ± 0.17	-285.78 ± 0.24	18.64 ± 0.16	12.3 ± 0.4
250	0.1	0.1	1	F	5.9	0.041	178 614	-279.09 ± 0.28	-285.69 ± 0.04	19.98 ± 0.28	13.4 ± 0.7
675	0.1	0.1	1	F	23.7	0.035	562 784	-278.88 ± 0.17	-285.16 ± 0.10	19.44 ± 0.17	12.9 ± 0.4
675	0.1	0.1	1	T	6.9	0.106	189 995	-278.34 ± 0.17	-285.27 ± 0.06	19.93 ± 0.16	13.2 ± 0.4
250	0.01	0.1	1	F	11.7	0.026	294 202	-280.78 ± 0.29	-285.67 ± 0.02	20.97 ± 0.30	14.4 ± 0.7
675	0.01	0.1	1	F	39.3	0.025	825 487	-280.62 ± 0.18	-285.38 ± 0.02	21.04 ± 0.17	13.6 ± 0.4
675	0.01	0.1	1	T	28.1	0.027	754 351	-280.76 ± 0.18	-285.23 ± 0.02	20.96 ± 0.16	14.0 ± 0.4
250	0.001	0.1	1	F	39.0	0.010	823 090	-281.78 ± 0.30	-285.93 ± 0.01	21.62 ± 0.28	14.6 ± 0.7
675	0.001	0.1	1	F	109.5	0.010	2116 032	-282.01 ± 0.18	-285.29 ± 0.02	21.99 ± 0.17	14.9 ± 0.4
675	0.001	0.1	1	T	126.3	0.008	2756 262	-280.92 ± 0.19	-285.52 ± 0.09	20.17 ± 0.17	15.2 ± 0.4

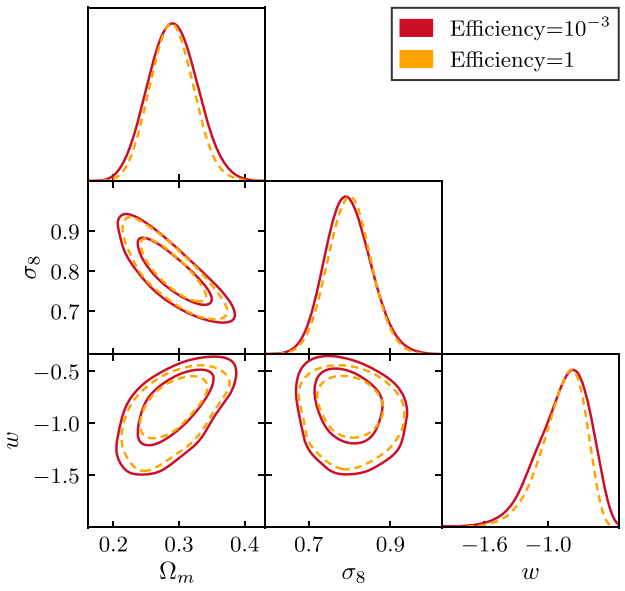

Figure 3. Marginalized one- and two-dimensional posterior distributions for two extreme values of the efficiency on MULTINEST. In red, a low value, which is therefore more likely to have fully sampled the tails of the distribution, and in the yellow dotted contours, a high value. We can see how the high efficiency does not fully sample the tails of the posterior distributions, and therefore gets narrower contours.

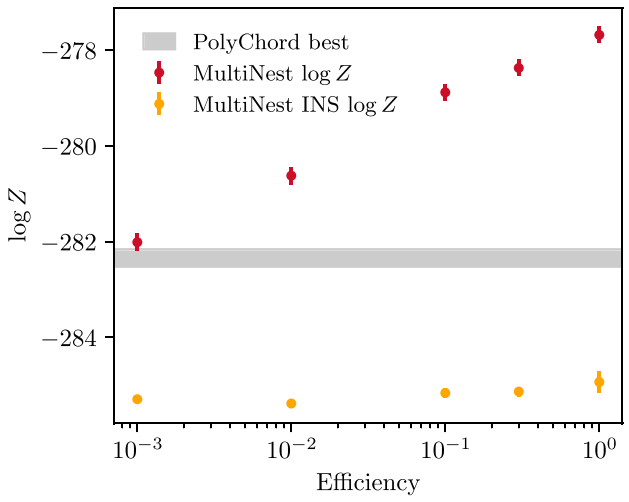

Figure 4. MULTINEST calculations of the evidence for different values of the efficiency. The MULTINEST values are plotted as red points, the MULTINEST INS evidence estimates in orange, and the grey band shows the 68 per cent confidence level of the best POLYCHORD estimate.

its dependence on hyperparameters to that of the main evidence calculation. This INS evidence estimate is very stable amongst all of our MULTINEST runs, always around the value of $\log Z \sim -285$, almost independently of the hyperparamter settings. Fig. 4 shows this graphically, with the INS estimates of the evidence in orange. They are significantly and consistently lower than the best estimate from POLYCHORD. While this might suggest convergence to the 'truth', this is belied by results from the Gaussian toy model of Fig. 5, in which the MULTINEST INS evidence estimates are also systematically biased low. Our results thus appear to contradict the findings of F13, which showed that in some toy models INS was more accurate than the baseline evidence estimate of MULTINEST. Note that in addition to being lower than the truth, the INS evidence reported also significantly underestimates its sampling error, i.e. the uncertainty caused by imperfect sampling.

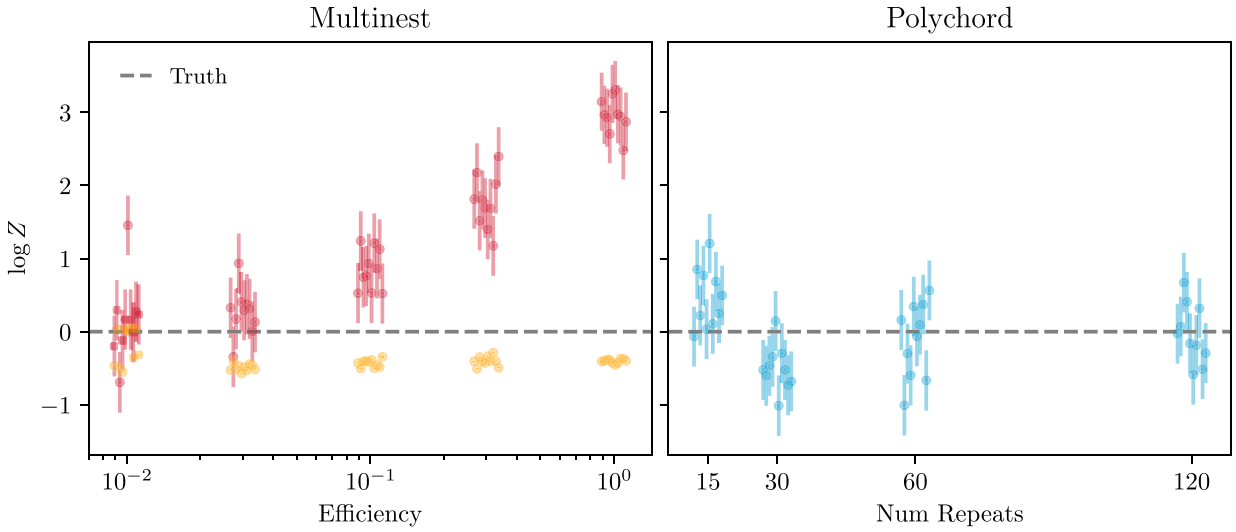

Figure 5. MULTINEST and POLYCHORD estimations of the evidence for different values of the efficiency and $n_{repeats}$ parameters, for a Gaussian posterior distribution with known truth (dashed line). In both cases, we use $nlive = 250$ and $tolerance = 0.1$. For each setting, we show 10 different sampling runs, which we displace along the x-axis for visualization purposes. The figure shows how POLYCHORD gets more reliable evidence estimates even for low values of $n_{repeats}$, where as MULTINEST gets biased estimates if the efficiency is not low enough. The yellow points include error bars, even though they are too small to be seen.

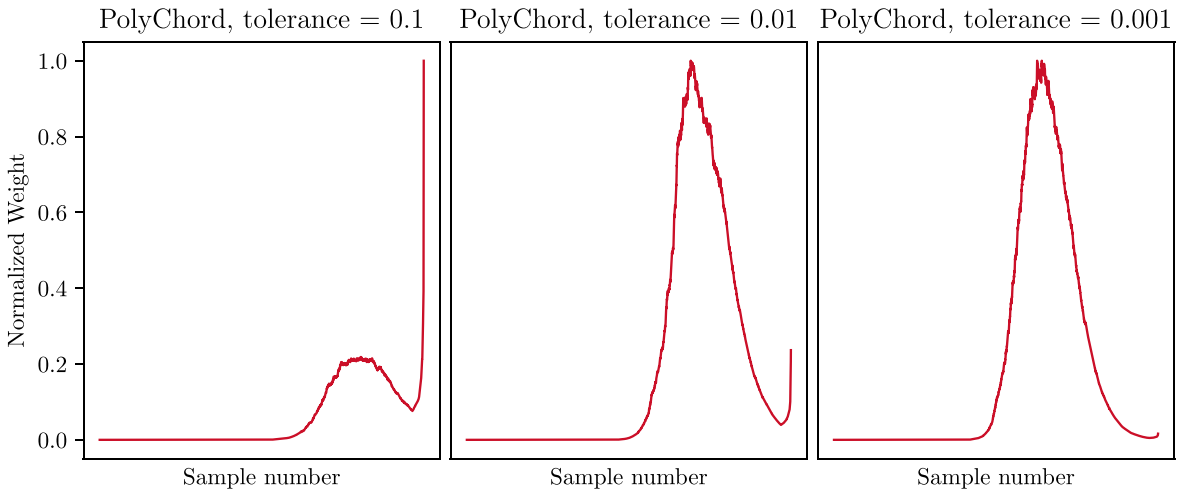

Figure 6. Normalized sample weight (weight divided by maximum weight) versus number of samples for different values of the tolerance. This plots serves as a convergence diagnostic; only the plot on the right has converged in this case. All plots use POLYCHORD with $n_{live}= 250$ and $n_{repeats}= 60$.

4.3 Polychord

Like MULTINEST, POLYCHORD has a number of hyperparameters which can be adjusted to balance running time and the accuracy of estimates for posterior distributions and summary statistics. The result of varying them is shown in Table 2.

(i) $\boldsymbol{n_{live}}$, the number of live points. As with MULTINEST, more live points lead to an increase in the accuracy of the posterior distribution, and to a decrease in the error estimate for the evidence. Also, as was the case for MULTINEST, POLYCHORD run-times scale linearly with the number of live points.

(ii) $\boldsymbol{tolerance}$, the stopping criterion. It is defined in the same way for Section 4.2, and the same conclusions apply: a lower tolerance does not have a significant impact on either runtime nor summary statistic accuracy. In fact, the tolerance seems to have even less of an impact on runtime for POLYCHORD than for MULTINEST. As before, we recommend using the weight-versus-step-number convergence diagnostic illustrated in the right-hand panel of Fig. 6 to select a reasonable tolerance for a chain, and note that a POLYCHORD chain can be resumed to reach a lower tolerance.

(iii) $n_{repeats}$, the number of repeats. This hyperparameter is specific to POLYCHORD's slice-sampling algorithm described in Section 2.4. Recall that at every step, POLYCHORD repeats the process of creating a slice through parameter space in a random direction, in which it finds a new potential live point. The value of $n_{repeats}$ dictates how many times this process is repeated for each sample selection. If this number is too low, the new live point will be correlated with the

Table 2. Comparison of time required for POLYCHORD with different settings. All runs use the DES Y1 3 × 2 pt likelihood and a wCDM cosmology, with 128 cores. The settings $OMPthreads = 1$ corresponds to 128 MPI threads, while $OMPthreads = 4$ corresponds to 32 MPI threads.

n_{live}	Tol	$n_{repeats}$	OMP	Time (h)	Acceptance	L Evals	log Z	D_{KL}	BMD
50	0.1	60	1	21.3	0.003	473 705	-282.22 ± 0.65	22.26 ± 0.64	12.1 ± 1.3
250	0.3	60	1	51.7	0.006	1171 509	-282.01 ± 0.29	21.85 ± 0.27	14.3 ± 0.7
250	0.1	15	1	11.9	0.022	342 878	-281.47 ± 0.32	21.51 ± 0.30	14.3 ± 0.7
250	0.1	30	1	23.7	0.011	675 895	-282.64 ± 0.29	22.35 ± 0.27	15.6 ± 0.7
250	0.1	60	1	46.2	0.006	1319 862	-282.51 ± 0.30	22.34 ± 0.26	15.2 ± 0.7
250	0.1	60	4	75.5	0.007	1016 058	-282.48 ± 0.29	22.26 ± 0.29	14.3 ± 0.7
250	0.1	120	1	87.4	0.003	2597 251	-282.87 ± 0.30	22.30 ± 0.32	14.9 ± 0.8
250	0.03	60	1	60.9	0.006	1379 582	-282.46 ± 0.30	22.21 ± 0.30	14.4 ± 0.7
250	0.01	60	1	62.4	0.006	1504 244	-282.72 ± 0.30	22.72 ± 0.34	13.5 ± 0.6
250	0.001	60	1	68.3	0.005	1670 676	-282.09 ± 0.30	22.12 ± 0.29	14.2 ± 0.6
675	0.1	15	1	22.5	0.026	733 506	-281.54 ± 0.18	21.38 ± 0.18	14.1 ± 0.4
675	0.1	30	1	47.5	0.014	1458 873	-282.52 ± 0.19	22.32 ± 0.19	14.6 ± 0.4
675	0.1	60	1	117.9	0.007	2863 702	-282.09 ± 0.18	21.95 ± 0.17	14.4 ± 0.5
675	0.01	60	1	131.2	0.007	3289 309	-282.14 ± 0.18	21.79 ± 0.17	14.5 ± 0.4
675	0.1	120	1	191.4	0.003	5795 180	-282.34 ± 0.18	22.32 ± 0.17	14.3 ± 0.4

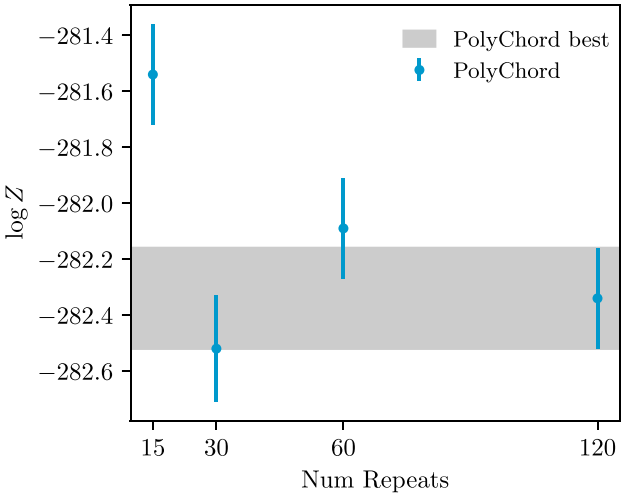

Figure 7. POLYCHORD estimations of the evidence for different values of $n_{repeats}$. The different values are plotted as red points, and the blue band shows the 68 per cent confidence level of the best POLYCHORD estimate.

previous sample in the chain. Such correlations between live points could bias the posterior distribution and summary statistics. In this sense, a higher value of $n_{repeats}$ is somewhat akin to increasing the degree of 'thinning' of a standard MCMC chain, with the result that more likelihood evaluations are performed but then discarded, in the interest of reducing systematic uncertainty.

The official POLYCHORD paper (Handley et al. 2015b) recommends using at least $n_{repeats} \sim 2D$, where D is the number of dimensions being sampled. In this work, we tested values that are approximately $\{0.5, 1, 2, 4\}$ times the number of dimensions. Fig. 7 shows the corresponding values of the summary statistics. We see that a value $n_{repeats} = 15$ obtains biased estimates of the evidence for the full likelihood, but values as low as $n_{repeats} = 30$ (for $D = 27$) already obtain valid results. The main conclusion is that, as expected, while a poor choice for $n_{repeats}$ can lead to biased results, the accuracy of estimates for evidence and other summary statistics are not nearly as sensitive to $n_{repeats}$ as they are to MULTINEST's efficiency hyperparameter.

(iv) OMP**threads**. As with MULTINEST, we obtain the best results when we use all our cores for MPI parallelization, up to the number of cores matching the number of live points.

4.4 Fast-likelihood tests of sampler variance

While the studies above give us an indication of how changing the values of POLYCHORD and MULTINEST hyperparameters affect runtime and the accuracy of summary statistic estimates, they do not tell us much about noise in those relations due to the particular realization of random points sampled in a given chain run. In order to assess this sampler variance, we run a large number of independent chains, which long runtimes make infeasible with the full DES likelihood studied above. Thus, to more robustly assess convergence properties and characterize how sampler variance changes across settings, we use the approximate fast likelihood to generate multiple chain realizations at each set of sampler hyperparameters.

We use a set of three 'high-quality' POLYCHORD chains with $n_{live} = 1000$, $n_{repeats} = 120$, and tol $= 0.001$ to approximate the truth and compare with the performance of chains run with lower quality settings. Unless otherwise stated, we ran 20 independent chains for each combination of settings. While the previous sections illustrated the different impacts of the $efficiency$ and $n_{repeats}$ parameters (as the unique hyperparameters for each sampler), we found that varying n_{live} had the most pronounced impact on posterior and evidence estimates for both MULTINEST and POLYCHORD. This aligns with the recommendations of Higson et al. (2019), who note that increasing n_{live} is the most computationally efficient way to increase accuracy, as it decreases both stochastic and systematic contributions to the uncertainty. We therefore mostly show results in this section with respect to varying n_{live}. Unless otherwise stated, MULTINEST chains were run with $efficiency = 0.1$ and $tolerance = 0.1$, and POLYCHORD chains with $n_{repeats} = 30$ and $tolerance = 0.01$.

Fig. 8 shows the marginalized 1D constraints on Ω_m and log$Z$ for multiple independent MULTINEST (red) and POLYCHORD (blue) chains with different numbers of live points. The average of the high-quality POLYCHORD chains are shown in grey. Constraints on Ω_m are consistent across the different of n_{live} values for both samplers. As expected, based on results in previous sections, we find the

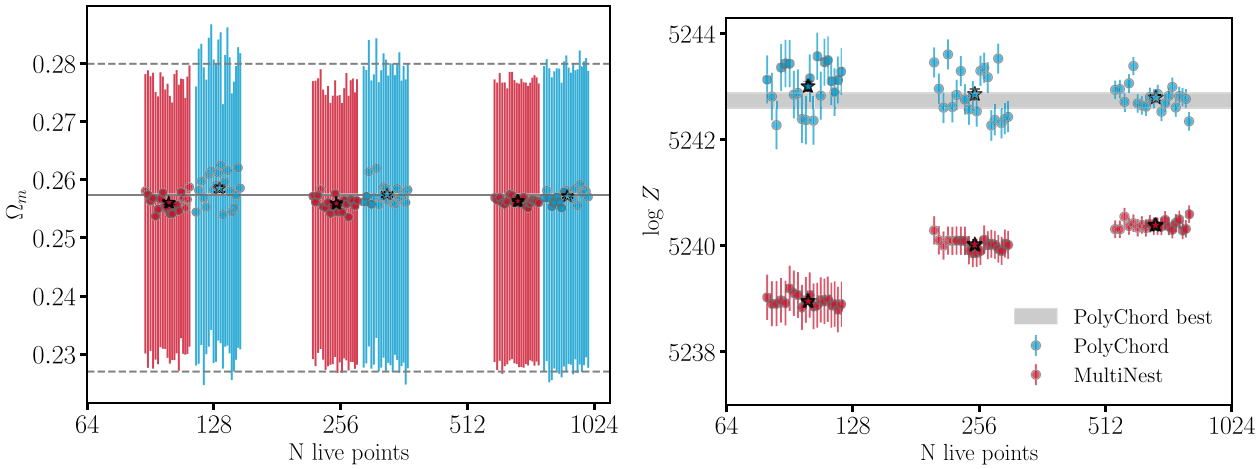

Figure 8. *Left*: 1D marginalized mean and 68 per cent credible intervals of Ω_m for both MULTINEST (red) and POLYCHORD (blue) run using the fast likelihood at different values of nlive ({100, 250, 675}), while other hyperparameters are held constant at the default values listed in Section 4.4. 20 chains are run at each setting of n_{live}, corresponding to the clusters of jittered points. The mean and standard error of chains at each setting are indicated by stars at the centre of each cluster. The horizontal lines correspond to the average mean and 1σ credible intervals of three high-quality POLYCHORD runs ($n_{live} = 1000$, $n_{repeats} = 120$, $tolerance = 10^{-3}$). There is good agreement between POLYCHORD and MULTINEST on the mean and small discrepancies between the credible intervals. MULTINEST credible intervals are consistently smaller than those reported by POLYCHORD. Fig. 9 shows this in greater detail for S_8. *Right*: Estimates of the Bayesian evidence (and its sampler-reported uncertainty) for the same chains as the left-hand plot. The shaded band shows $\pm 1\sigma$ uncertainty on the mean evidence of the three high-quality POLYCHORD chains. The POLYCHORD values are consistent at all settings, including at the lowest settings, with each individual run consistent with the high-resolution 'truth' within its reported uncertainty. In contrast, MULTINEST evidence estimates display a systematic bias that is greater for small n_{live}, and the reported uncertainty for individual chains is insufficient to make runs consistent across different values of n_{live}. The reported uncertainties in $\log Z$ for each individual chain (given by the error bars) is consistent with the sampling variance across chain means for POLYCHORD, but is greatly overestimated for MULTINEST where the means across chains are much more tightly clustered. This is shown more directly in Fig. 10.

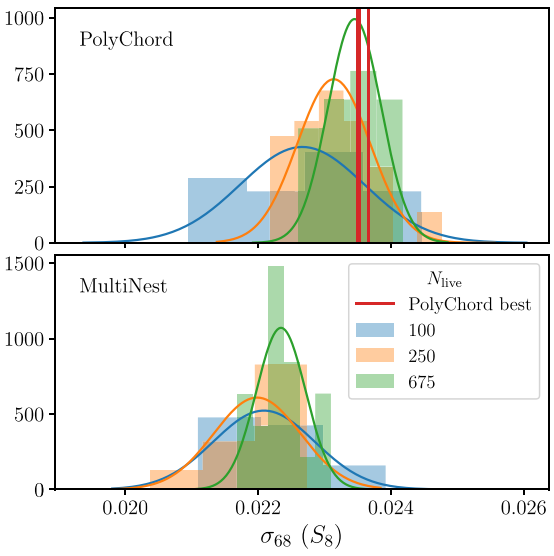

Figure 9. Histogram of the half-width of 68 per cent credible intervals for S_8 for many chains using the fast likelihood for both POLYCHORD (top) and MULTINEST (bottom). MULTINEST systematically reports smaller credible intervals than POLYCHORD for the range of settings tested, here shown with colours indicating different numbers of live points. The red lines indicate the credible intervals for the three high-quality POLYCHORD chains.

Ω_m credible intervals reported by MULTINEST to be consistently ~ 10 per cent smaller than those reported by POLYCHORD. The same is true to a lesser extent for S_8. This is shown in Fig. 9, which depicts the estimated uncertainty in S_8 inferred from POLYCHORD (top) and MULTINEST (bottom) chains run with different n_{live} settings. The uncertainty is represented by the half-width of the 68 per cent credible interval ($\sigma_{68}(S_8)$), a quantity comparable to the standard deviation but more directly related to the marginalized quantities we are interested in for Bayesian cosmological inference and less sensitive to the tails of the posterior. As was noted in Section 4.2, underestimation of credible intervals is expected when the MULTINEST *efficiency* parameter is too large. We see here that having low n_{live} can also potentially cause parameter constraint error bars to be slightly underestimated.

Estimates of both the mean and dispersion of the Bayesian evidence differ significantly between the samplers. The values of $\log Z$ reported by POLYCHORD are consistent across the range of settings we tested, indicating minimal systematic bias in the estimates due to hyperparameter settings. In contrast, the reported MULTINEST evidence changes significantly as a function of n_{live}, and to a degree much larger than the algorithm's reported uncertainty. This can also be seen in Figs 4 and 5, which show results for different likelihoods. As we increase n_{live} in Fig. 8, the MULTINEST evidence estimate shifts towards that reported by PolyChord. This behaviour suggests that the MultiNest settings tested here are insufficient to obtain accurate pieces of evidence.

In addition to being robust to systematic shifts in the reported evidence, the sample variance in evidence estimates between different POLYCHORD chains with the same settings shows good agreement with the sampler variance uncertainty reported from each individual chain. Fig. 10 shows the distribution of reported evidence values across the 20 chains at each value of n_{live}. The dashed Gaussian curves are drawn according to

$$\mathcal{N}\left(\langle \log Z\rangle_c, \langle \sigma(\log Z)\rangle_c^2\right), \tag{16}$$

where the averages are computed over each ensemble of 20 chains. For POLYCHORD, these closely match the empirical distribution of

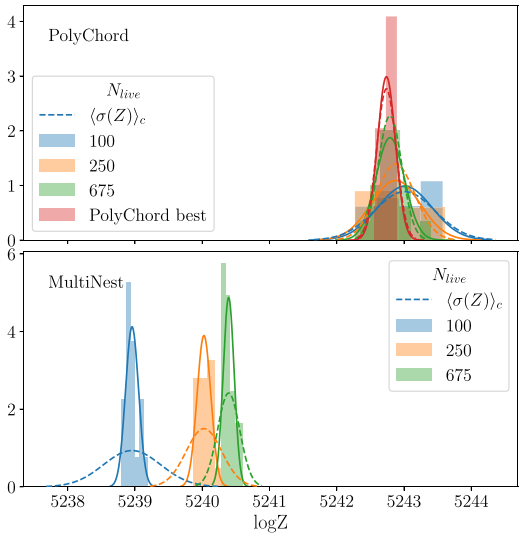

Figure 10. Histogram of reported $\log Z$ for many POLYCHORD (top) and MULTINEST (bottom) chains run with different numbers of live points. The solid curves are Gaussian fits to the distribution in reported $\log Z$ across different chain realizations. The dashed Gaussian curves show the expected distribution based on the mean $\log Z$ and mean claimed uncertainty across chains. The relatively close agreement between solid and dashed curves indicates that the reported uncertainty in POLYCHORD chains is fairly representative of the true sampling error. MULTINEST reports uncertainties considerably larger than the observed sampling variance.

reported pieces of evidence, indicated by the solid Gaussian curves which are fits to the histogram. In contrast, the reported statistical uncertainty in MULTINEST pieces of evidence is much greater than the observed scatter across chain realizations.

The uncertainty in $\log Z$ given by POLYCHORD is more solidly grounded than that of MULTINEST, having been derived analytically by Keeton (2011), whereas the MULTINEST evidence estimates use the relative entropy and are based on information theoretic arguments. Our numerical results affirm the greater reliability of evidence errors from POLYCHORD.

Despite MULTINEST overestimating the *statistical* uncertainty in $\log Z$, the reported pieces of evidence are still inconsistent across hyperparameter settings, due to even greater *systematic* shifts as n_{live} is increased. However, none of the MULTINEST settings tested resulted in evidence values approaching the stable POLYCHORD estimates.

We can also use this ensemble of chains to estimate and limit the contribution of sampler biases to the systematic error budget of the DES Y3 analysis. The DES Y3 analysis requires that unmodelled systematics shift the maximum posterior point of key cosmological parameters by not more than 0.3σ in the 2D plane of $\Omega_{\rm m}$ and S_8 (c.f. Krause et al. 2021). Here, we adopt a somewhat simplified but none the less strict criterion that the typical variation of parameter means across chain realizations is far below their statistical uncertainty. We thus require

$$\sigma_c\left[\bar{\theta}\right] < 0.1\langle\sigma_s\left[\theta\right]\rangle_c, \tag{17}$$

where the left-hand side is the standard deviation of the parameter mean across chain realizations and the right-hand side is a threshold, set equal to a fraction of the average of the standard deviations of the parameter computed from the individual chains. Note that this requirement is closely related to requiring the Gelman–Rubin statistic defined in Section 2.2 to be below a given threshold.

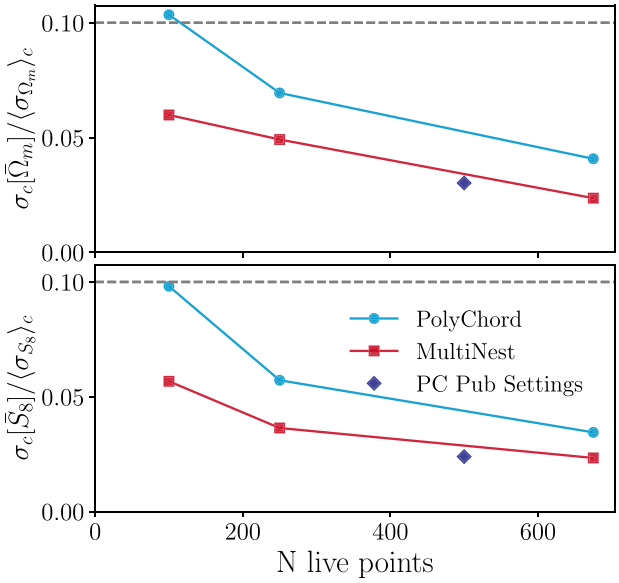

Figure 11. Standard deviation of parameter means across chains relative to the average within-chain parameter standard deviation. We require that the contribution to mean parameter shifts from sampler variance is small for settings used to run chains testing the impact of unmodelled systematics. The POLYCHORD publication settings can be found in Table 3.

We are primarily interested in shifts on $\Omega_{\rm m}$ and S_8, and so require that any recommended sampler settings satisfy equation (17) for those parameters. Fig. 11 confirms this requirement is fulfilled for $n_{live}>200$ with the fiducial values of the other hyperparameters.

5 CONCLUSIONS

In this paper, we have studied the performance of two commonly used tools used to sample posteriors for cosmological analysis, MULTINEST and POLYCHORD, as a function of their hyperparameter settings. Our analysis had two parts: testing multiple sampler settings on the DES Y1 3 × 2 pt analysis to calibrate the time needed to get unbiased posterior distributions in the Y3 analysis, and also using a faster approximate version of the likelihood to characterize the amount of sampler variance and further validate those findings.

We found that these Nested Sampling algorithms require careful tuning of their hyperparameters, especially n_{live} and MULTINEST's ellipsoidal sampling efficiency. Particularly for MULTINEST, the wrong settings can lead to a poor sampling of the tails of the posterior distribution, and to a biased evidence estimation. Furthermore, the superior speed of the MULTINEST algorithm compared with POLYCHORD's slice sampling method is not present when sufficiently accurate sampling hyperparameters are used. POLYCHORD produces unbiased evidence estimates with reasonable settings, as well as contours that are in good agreement with those we find using Metropolis–Hastings. Therefore, our findings lead us to prefer POLYCHORD over MULTINEST.

The studies described in this paper were used to guide recommendations for sampler settings used for the DES Y3 cosmology analysis (Krause et al. 2021) as well as some Y1 follow-up papers (Chen et al. 2021; Muir et al. 2021). These recommendations are summarized in Table 3 for three use-cases. We recommend that MULTINEST only be used for preliminary testing and pipeline debugging. While it is relatively fast, we found the (fairly standard) settings described in Case III of Table 3 to produce marginalized pos-

Table 3. Recommended sampler settings based on this work, used for the DES Y3 cosmology analysis. Approximate wall-time estimates are given for a wCDM DES Y1 3 × 2 chain run on 128 cores.

Case I: Publication quality (one-time runs)	
Sampler	POLYCHORD
n_{live}	500
Tol	0.01
$n_{repeats}$	60
fast_fraction	0.01
Time to run:	∼4 d
Case II: Testing (noisy contours)	
Sampler	POLYCHORD
n_{live}	250
Tol	0.1
$n_{repeats}$	30
fast_fraction	0.0
Time to run:	∼1 d
Case III: Very preliminary results only (unreliable pieces of evidence, Ω_m and σ_8 posterior widths underestimated by ∼10 per cent)	
Sampler	MULTINEST
n_{live}	250
Efficiency	0.3
Tol	0.1
constant_efficiency	F
Time to run:	∼6 h

terior widths for Ω_m and σ_8 that are systematically underestimated by about 10 per cent and unreliable pieces of evidence. For most pre-publication testing, we recommend using the fast POLYCHORD settings described in Case II. Those settings will produce unbiased posteriors and evidence values, though the resulting contours for marginalized posteriors will be somewhat noisy. Case I in Table 3 presents our recommendation for publication-quality results, with an increased number of live points extending the run-time but resulting in reduced noise in both posterior contours and evidence estimates.

While this study was performed specifically for DES Y3, our findings should be useful as a guide for cosmological analyses of similar dimensionality. Though our results were broadly consistent across three different likelihoods: the full DES Y1 likelihood, a fast approximate DES Y1 likelihood, and a 27D Gaussian toy model, the exact settings of Table 3 will likely need to be adjusted for problems where the number of dimensions or the shape of the posterior distribution change significantly. We found that good posterior and evidence estimates can be obtained with relatively low settings of $n_{repeats} \sim D$ with POLYCHORD, but that MULTINEST pieces of evidence were systematically biased except at extreme values of $efficiency$, and MULTINEST missed tail regions of the posterior such that reported credible intervals were consistently ∼ 10 per cent smaller than those reported by POLYCHORD. We found that increasing n_{live} had the greatest impact in improving accuracy of the summary statistics of interest.

Sampling algorithms are a key component of modern cosmological analyses and it is important to characterize their impacts on inference. As demonstrated in this work, poor choice of sampler and/or hyperparameter settings can lead to biased estimates of parameter constraints and other key summary statistics, but it is possible to achieve sufficiently unbiased estimates in realistic use cases. Through this work, we motivate the sampling methods used for the DES Y3 analyses, and note that the fine margins demanded by precision cosmology will increasingly require heightened scrutiny of sampling tools.

ACKNOWLEDGEMENTS

PL acknowledges STFC Consolidated Grants ST/R000476/1 and ST/T000473/1. NW is supported by the Chamberlain fellowship at Lawrence Berkeley National Laboratory.

The analysis used the software tools SCIPY (Jones et al. 2001), NUMPY (Oliphant 2006), MATPLOTLIB (Hunter 2007), CAMB (Lewis et al. 2000; Howlett et al. 2012), GETDIST (Lewis 2019), MULTINEST (Feroz & Hobson 2008; Feroz et al. 2009, 2019), POLYCHORD (Handley et al. 2015a, b), ANESTHETIC (Handley 2019), and COSMOSIS (Zuntz et al. 2015). Elements of the DES modelling pipeline additionally use COSMOLIKE (Krause & Eifler 2017), HALOFIT (Bird, Viel & Haehnelt 2012; Takahashi et al. 2012), FASTPT (McEwen et al. 2016), and NICAEA (Kilbinger et al. 2009).

This work was supported through computational resources and services provided by the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility operated under Contract No. DE-AC02-05CH11231; and by the Sherlock cluster, supported by Stanford University and the Stanford Research Computing Center.

Funding for the DES Projects has been provided by the U.S. Department of Energy, the U.S. National Science Foundation, the Ministry of Science and Education of Spain, the Science and Technology Facilities Council of the United Kingdom, the Higher Education Funding Council for England, the National Center for Supercomputing Applications at the University of Illinois at Urbana-Champaign, the Kavli Institute of Cosmological Physics at the University of Chicago, the Center for Cosmology and Astro-Particle Physics at the Ohio State University, the Mitchell Institute for Fundamental Physics and Astronomy at Texas A&M University, Financiadora de Estudos e Projetos, Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro, Conselho Nacional de Desenvolvimento Científico e Tecnológico and the Ministério da Ciência, Tecnologia e Inovação, the Deutsche Forschungsgemeinschaft and the Collaborating Institutions in the Dark Energy Survey.

The Collaborating Institutions are Argonne National Laboratory, the University of California at Santa Cruz, the University of Cambridge, Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas-Madrid, the University of Chicago, University College London, the DES-Brazil Consortium, the University of Edinburgh, the Eidgenössische Technische Hochschule (ETH) Zürich, Fermi National Accelerator Laboratory, the University of Illinois at Urbana-Champaign, the Institut de Ciències de l'Espai (IEEC/CSIC), the Institut de Física d'Altes Energies, Lawrence Berkeley National Laboratory, the Ludwig-Maximilians Universität München and the associated Excellence Cluster Universe, the University of Michigan, NSF's NOIRLab, the University of Nottingham, The Ohio State University, the University of Pennsylvania, the University of Portsmouth, SLAC National Accelerator Laboratory, Stanford University, the University of Sussex, Texas A&M University, and the OzDES Membership Consortium.

Based in part on observations at Cerro Tololo Inter-American Observatory at NSF's NOIRLab (NOIRLab Prop. ID 2012B-0001; PI: J. Frieman), which is managed by the Association of Universities for Research in Astronomy (AURA) under a cooperative agreement with the National Science Foundation.

The DES data management system is supported by the National Science Foundation under Grant Numbers AST-1138766 and AST-1536171. The DES participants from Spanish institutions are partially supported by MICINN under grants ESP2017-89838, PGC2018-094773, PGC2018-102021, SEV-2016-0588, SEV-2016-0597, and MDM-2015-0509, some of which include ERDF funds from the European Union. IFAE is partially funded by the CERCA program of the Generalitat de Catalunya. Research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Program (FP7/2007-2013) including ERC grant agreements 240672, 291329, and 306478. We acknowledge support from the Brazilian Instituto Nacional de Ciência e Tecnologia (INCT) do e-Universo (CNPq grant 465376/2014-2).

This manuscript has been authored by Fermi Research Alliance, LLC under Contract No. DE-AC02-07CH11359 with the U.S. Department of Energy, Office of Science, Office of High Energy Physics.

DATA AVAILABILITY STATEMENT

The data underlying this article are available in the Dark Energy Survey Data Management platform, at https://des.ncsa.illinois.edu

REFERENCES

Abbott T. M. C. et al., 2018, Phys. Rev. D, 98, 043526
Abbott T. M. C. et al., 2022, Phys. Rev. D, 105, 023520
Aitken S., Akman O. E., 2013, BMC Syst. Biol., 7, 72
Allison R., Dunkley J., 2014, MNRAS, 437, 3918
An L., Brooks S., Gelman A., 1998, J. Comput. Graph. Stat., 7, 434
Betancourt M., 2017, preprint (arXiv:1701.02434)
Bird S., Viel M., Haehnelt M. G., 2012, MNRAS, 420, 2551
Cameron E., Pettitt A., 2014, Stat. Sci., 29, 397
Chen A. et al., 2021, Phys. Rev. D, 103, 123528
Christensen N., Meyer R., Knox L., Luey B., 2001, Class. Quantum Gravity, 18, 2677
Di Valentino E. et al., 2021a, Astropart. Phys., 131, 102604
Di Valentino E. et al., 2021b, Astropart. Phys., 131, 102605
Dunkley J., Bucher M., Ferreira P. G., Moodley K., Skordis C., 2005, MNRAS, 356, 925
Eisenstein D. J., Hu W., 1998, ApJ, 496, 605
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., 2019, Open J. Astrophys., 2 10
Foreman-Mackey D., Hogg D. W., Lang D., Goodman J., 2013, PASP, 125, 306
Gelman A., Rubin D. B., 1992, Stat. Sci., 7, 457
Handley W., 2019, J. Open Source Softw., 4, 1414
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
Hastings W. K., 1970, Biometrika, 57, 97
Higson E., Handley W., Hobson M., Lasenby A., 2019, MNRAS, 483, 2044
Hobson M. P., Bridle S. L., Lahav O., 2002, MNRAS, 335, 377
Hoffman M. D., Gelman A., 2014, J. Mach. Learn. Res., 15, 1593
Hogg D. W., Foreman-Mackey D., 2018, ApJS, 236, 11
Howlett C., Lewis A., Hall A., Challinor A., 2012, J. Cosmol. Astropart. Phys., 1204, 027
Hunter J. D., 2007, Comput. Sci. Eng., 9, 90
Jones E. et al., 2001, SciPy: Open source scientific tools for Python, http://www.scipy.org/
Karamanis M., Beutler F., 2021, Stat. Comput., 31, 61
Karamanis M., Beutler F., Peacock J. A., 2021, MNRAS, 508, 3589
Keeton C. R., 2011, MNRAS, 414, 1418
Kilbinger M. et al., 2009, A&A, 497, 677
Knox L., Christensen N., Skordis C., 2001, ApJ, 563, L95
Krause E. et al., 2021, preprint (arXiv:2105.13548)
Krause E., Eifler T., 2017, MNRAS, 470, 2100
Kullback S., Leibler R. A., 1951, Ann. Math. Stat., 22, 79
Lahav O., Bridle S. L., Hobson M. P., Lasenby A. N., Sodre L., Jr., 2000, MNRAS, 315, L45
Lewis A., 2013, Phys. Rev. D, 87, 103529
Lewis A., 2019, GetDist: MCMC sample analysis, plotting and GUI, https://github.com/cmbant/getdist
Lewis A., Bridle S., 2002, Phys. Rev. D, 66, 103511
Lewis A., Challinor A., Lasenby A., 2000, ApJ, 538, 473
Luis Bernal J., Peacock J. A., 2018, J. Cosmol. Astropart. Phys., 07, 002
MacKay D. J. C., 2002, Information Theory, Inference & Learning Algorithms. Cambridge University Press, New York, NY, USA
Marshall P., Rajguru N., Slosar A., 2006, Phys. Rev. D, 73, 067302
McEwen J. E., Fang X., Hirata C. M., Blazek J. A., 2016, J. Cosmol. Astropart. Phys., 09, 015
Metropolis N., Rosenbluth A. W., Rosenbluth M. N., Teller A. H., Teller E., 1953, J. Chem. Phys., 21, 1087
Muir J. et al., 2021, Phys. Rev. D, 103, 023528
Mukherjee P., Parkinson D., Corasaniti P. S., Liddle A. R., Kunz M., 2006a, MNRAS, 369, 1725
Mukherjee P., Parkinson D., Liddle A. R., 2006b, ApJ, 638, L51
Neal R. M., 1997, Technical Report CRG-TR-93-1
Neal R. M., 2005, preprint(arXiv Mathematics e-prints, p. math/0502099)
Oliphant T. E., 2006, Guide to NumPy. Trelgol Publishing, USA
Pahud C., Liddle A. R., Mukherjee P., Parkinson D., 2006, Phys. Rev. D, 73, 123524
Parkinson D., Mukherjee P., Liddle A. R., 2006, Phys. Rev. D, 73, 123523
Planck Collaboration, 2020, A&A, 641, A6
Riess A. G. et al., 2022, ApJ, 934, L7
Shah P., Lemos P., Lahav O., 2021, A&AR, 29, 9
Shaw J. R., Bridges M., Hobson M. P., 2007, MNRAS, 378, 1365
Sinharay S., 2003, ETS Research Report Series, 2003, 52
Sivia D. S., Skilling J., 2006, Data Analysis: A Bayesian Tutorial. Oxford science publications, Oxford University Press, Oxford, New York,
Skilling J., 2006, Bayesian Anal., 1, 833
Speagle J. S., 2020, MNRAS, 493, 3132
Spiegelhalter D. J., Best N. G., Carlin B. P., Van Der Linde A., 2002, J. R. Stat. Soc. B, 64, 583
Takahashi R., Sato M., Nishimichi T., Taruya A., Oguri M., 2012, ApJ, 761, 152
Tegmark M. et al., 2004, Phys. Rev. D, 69, 103501
The Dark Energy Survey Collaboration, 2005, preprint(pp astro–ph/0510346)
Torrado J., Lewis A., 2021, J. Cosmol. Astropart. Phys., 05, 057
Verde L. et al., 2003, ApJS, 148, 195
Zuntz J. et al., 2015, Astron. Comput., 12, 45

APPENDIX A: OTHER SAMPLERS

In this section, we briefly compare two nested samplers used in this paper, `MULTINEST` and `POLYCHORD`, with two other available sampling codes, `EMCEE` and `ZEUS`, in terms of the stability of their posterior estimates. As already noted, `EMCEE` showed very poor convergence with the full DES likelihood and `ZEUS` was not integrated into the `COSMOSIS` framework at the time of testing, a requirement for its use in the fiducial DES analysis. Furthermore, neither of these samplers provide an estimate of the Bayesian Evidence, a key summary statistic of interest for the DES Y3 analysis. We therefore use the toy model described in Section 4.2, a 27-dimensional Gaussian likelihood distribution, and compare the stability of posterior estimates.

For each sampler, we run 10 chains with the same settings, and then plot the standard deviation of each parameter mean across chains divided by the average parameter standard deviation within

Table A1. Comparison between POLYCHORD, EMCEE, and ZEUS. All three samplers are run with settings that generate around 1.5×10^6 likelihood evaluations, and therefore are expected to take similar amounts of time in realistic cases, where likelihood evaluations are the slowest part of the calculation. Likelihood evaluations and acceptance are averaged over each of the 10 chains. The last column is averaged over each of the 27 parameters.

Sampler	⟨L Evals⟩	⟨Acceptance⟩	$\sigma_c[\bar{p}]/\langle\sigma_p\rangle_c$
POLYCHORD	1749 508	0.0074	0.06
EMCEE	1500 000	0.26	0.06
ZEUS	1666 525	0.18	0.01

each chain. Defining similar settings across different samplers can be challenging; we use the Case II settings showed in Table 3 for POLYCHORD, and tune the settings in the other samplers to obtain a similar number of likelihood evaluations, of order 5×10^5. Our results are shown in Table A1.

We find that for similar numbers of likelihood evaluations, we get many more samples for EMCEE and ZEUS as shown by their higher acceptance rates. This is expected for MCMC samplers when compared to nested samplers. We also see how, for a similar number of likelihood evaluations, ZEUS seems to obtain more stable posterior means than the other two samplers, which shows the great potential of ZEUS for future cosmological analyses. It is important, however, to highlight once more the two main advantages of Nested Sampling, despite its low acceptance rate: It provides us with an estimate of the Bayesian Evidence, needed for Bayesian model comparison and tension quantification; and it can sample multimodal spaces, something with which MCMC samplers tend to struggle. Therefore, it is key to select the sampling method that better suits the problem at hand, and if possible to use multiple methods to ensure robustness.

APPENDIX B: EFFICIENCY RULE OF THUMB

From the results of this paper, as well as our understanding of ellipsoidal nested sampling, we can set up an approximately guideline to correctly choose an efficiency that will not lead to biased sampling.

We can estimate the enlargement of the ellipse in each direction from the efficiency and the number of dimensions N_d, as

$$\text{enlargement} = \text{eff}^{-1/N_d} \tag{B1}$$

From Mukherjee et al. (2006a) and our work, we estimate that we need an enlargement of approximately $\gtrsim$1.5. Therefore, the required efficiency is approximately

$$\text{eff} \sim 1.5^{-N_d}. \tag{B2}$$

This shows how poorly the required efficiency scales with an increase in dimensionality. In the case of DES, the required efficiency is $e \sim 10^{-5}$, which is consistent with the findings of Fig. 4, showing that an efficiency of 10^{-3} is not enough to reach the correct evidence values.

As a side note, we can also estimate the acceptance of MULTINEST as

$$\text{acc} = \text{eff}^{\text{BMD}/\text{N}_\text{d}}, \tag{B3}$$

where BMD is the Bayesian Model Dimensionality.

[1]*Department of Physics & Astronomy, University College London, Gower Street, London WC1E 6BT, UK*
[2]*Department of Physics and Astronomy, Pevensey Building, University of Sussex, Brighton BN1 9QH, UK*
[3]*Department of Physics, University of Michigan, Ann Arbor, MI 48109, USA*
[4]*Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720, USA*
[5]*Jodrell Bank Center for Astrophysics, School of Physics and Astronomy, University of Manchester, Oxford Road, Manchester M13 9PL, UK*
[6]*Perimeter Institute for Theoretical Physics, 31 Caroline St North, Waterloo, ON N2L 2Y5, Canada*
[7]*Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Dr., Pasadena, CA 91109, USA*
[8]*Instituto de Astrofísica e Ciências do Espaço, Faculdade de Ciências, Universidade de Lisboa, P-1769-016 Lisboa, Portugal*
[9]*Department of Physics, Carnegie Mellon University, Pittsburgh, PA 15312, USA*
[10]*Department of Physics and Astronomy, University of Pennsylvania, Philadelphia, PA 19104, USA*
[11]*Institute for Astronomy, University of Edinburgh, Edinburgh EH9 3HJ, UK*
[12]*Department of Astronomy, University of California, Berkeley, 501 Campbell Hall, Berkeley, CA 94720, USA*
[13]*Department of Astronomy/Steward Observatory, University of Arizona, 933 North Cherry Avenue, Tucson, AZ 85721-0065, USA*
[14]*Department of Astronomy, University of Geneva, ch. d'Ecogia 16, CH-1290 Versoix, Switzerland*
[15]*Laboratório Interinstitucional de e-Astronomia - LIneA, Rua Gal. José Cristino 77, Rio de Janeiro, RJ 20921-400, Brazil*
[16]*Fermi National Accelerator Laboratory, P. O. Box 500, Batavia, IL 60510, USA*
[17]*Institut d'Astrophysique de Paris, CNRS, UMR 7095, F-75014 Paris, France*
[18]*Institut d'Astrophysique de Paris, Sorbonne Universités, UPMC Univ Paris 06, UMR 7095, F-75014 Paris, France*
[19]*Faculty of Physics, Ludwig-Maximilians-Universität, Scheinerstr 1, D-81679 Munich, Germany*
[20]*Kavli Institute for Particle Astrophysics & Cosmology, P. O. Box 2450, Stanford University, Stanford, CA 94305, USA*
[21]*SLAC National Accelerator Laboratory, Menlo Park, CA 94025, USA*
[22]*Instituto de Astrofisica de Canarias, E-38205 La Laguna, Tenerife, Spain*
[23]*Dpto. Astrofísica, Universidad de La Laguna, E-38206 La Laguna, Tenerife, Spain*
[24]*Center for Astrophysical Surveys, National Center for Supercomputing Applications, 1205 West Clark St, Urbana, IL 61801, USA*
[25]*Department of Astronomy, University of Illinois at Urbana-Champaign, 1002 W. Green Street, Urbana, IL 61801, USA*
[26]*Institut de Física d'Altes Energies (IFAE), The Barcelona Institute of Science and Technology, Campus UAB, E-08193 Bellaterra (Barcelona) Spain*
[27]*Institut d'Estudis Espacials de Catalunya (IEEC), E-08034 Barcelona, Spain*
[28]*Institute of Space Sciences (ICE, CSIC), Campus UAB, Carrer de Can Magrans, s/n, E-08193 Barcelona, Spain*
[29]*Department of Physics, California Institute of Technology, 1200 East California Blvd, MC 249-17, Pasadena, CA 91125, USA*
[30]*Astronomy Unit, Department of Physics, University of Trieste, via Tiepolo 11, I-34131 Trieste, Italy*
[31]*INAF-Osservatorio Astronomico di Trieste, via G. B. Tiepolo 11, I-34143 Trieste, Italy*
[32]*Institute for Fundamental Physics of the Universe, Via Beirut 2, I-34014 Trieste, Italy*
[33]*Observatório Nacional, Rua Gal. José Cristino 77, Rio de Janeiro, RJ 20921-400, Brazil*
[34]*Hamburger Sternwarte, Universität Hamburg, Gojenbergsweg 112, D-21029 Hamburg, Germany*
[35]*Santa Cruz Institute for Particle Physics, Santa Cruz, CA 95064, USA*
[36]*Institute of Theoretical Astrophysics, University of Oslo, P.O. Box 1029 Blindern, NO-0315 Oslo, Norway*
[37]*Kavli Institute for Cosmological Physics, University of Chicago, Chicago, IL 60637, USA*
[38]*Instituto de Fisica Teorica UAM/CSIC, Universidad Autonoma de Madrid, E-28049 Madrid, Spain*
[39]*Department of Astronomy, University of Michigan, Ann Arbor, MI 48109, USA*
[40]*Excellence Cluster Origins, Boltzmannstr 2, D-85748 Garching, Germany*

[41]*School of Mathematics and Physics, University of Queensland, Brisbane, QLD 4072, Australia*
[42]*Center for Cosmology and Astro-Particle Physics, The Ohio State University, Columbus, OH 43210, USA*
[43]*Department of Physics, The Ohio State University, Columbus, OH 43210, USA*
[44]*Center for Astrophysics |Harvard & Smithsonian, 60 Garden Street, Cambridge, MA 02138, USA*
[45]*Australian Astronomical Optics, Macquarie University, North Ryde, NSW 2113, Australia*
[46]*Lowell Observatory, 1400 Mars Hill Rd, Flagstaff, AZ 86001, USA*
[47]*Departamento de Física Matemática, Instituto de Física, Universidade de São Paulo, CP 66318, São Paulo, SP 05314-970, Brazil*
[48]*Department of Astrophysical Sciences, Princeton University, Peyton Hall, Princeton, NJ 08544, USA*
[49]*Institució Catalana de Recerca i Estudis Avançats, E-08010 Barcelona, Spain*
[50]*Physics Department, University of Wisconsin-Madison, 2320 Chamberlin Hall, 1150 University Avenue Madison, WI 53706-1390, USA*
[51]*Institute of Astronomy, University of Cambridge, Madingley Road, Cambridge CB3 0HA, UK*
[52]*Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas (CIEMAT), Madrid, Spain*
[53]*School of Physics and Astronomy, University of Southampton, Southampton SO17 1BJ, UK*
[54]*Computer Science and Mathematics Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA*
[55]*National Center for Supercomputing Applications, 1205 West Clark St, Urbana, IL 61801, USA*
[56]*Institute of Cosmology and Gravitation, University of Portsmouth, Portsmouth PO1 3FX, UK*
[57]*Max Planck Institute for Extraterrestrial Physics, Giessenbachstrasse, D-85748 Garching, Germany*

This paper has been typeset from a TEX/LATEX file prepared by the author.