$$R = \frac{P(M_1|\mathbf{D})}{P(M_0|\mathbf{D})} = \frac{\mathcal{Z}_1}{\mathcal{Z}_0}\frac{P(M_1)}{P(M_0)} = \frac{\mathcal{Z}_1}{\mathcal{Z}_0} \; . \tag{5}$$

Here, $P(M_1)/P(M_0)$ is the probability ratio for the two models *a priori*, which is conventionally set to unity; the *evidence* $\mathcal{Z}$ of a model M is the marginalized likelihood of the data, i.e. the probability of having obtained the data $\mathbf{D}$ integrated over all possible values of the model parameters $\boldsymbol{\theta}$:

$$\mathcal{Z} = \int \mathcal{L}(\mathbf{D}|M(\boldsymbol{\theta}))\, \pi(\boldsymbol{\theta})\, \mathrm{d}^D\boldsymbol{\theta} \; , \tag{6}$$

where $\mathcal{L}(\mathbf{D}|M(\boldsymbol{\theta}))$, $\pi(\boldsymbol{\theta})$ and D are, respectively, the likelihood of the data, the prior of the parameters in the model and the dimensionality of the parameter space. In this work, we will use M_1 and M_0 to denote the feature and featureless ΛCDM models;[3] the cosmological parameter ranges we studied are listed in Tab.I. And the multidimensional integration in Eq. (6) was sampled via the multi-modal implementation of the nested sampling algorithm MultiNest [59–61].

Parameter	Range (min, max)
$\Omega_b h^2$	$(0.005, 0.100)$
$\Omega_c h^2$	$(0.01, 0.99)$
$100\vartheta_*$	$(0.5, 10.0)$
$\tau_{\rm reio}$	$(0.01, 0.80)$
n_s	$(0.9, 1.1)$
$\ln(10^{10}A_s^2)$	$(2.7, 4.0)$
B	$(-0.2, 0)$
$\ln\beta$	$(0, 7.5)$
$\ln(-\tau_0)$	$(4.3, 6.0)$

TABLE I. List of the parameters used in the multimodal nested sampling. Besides these parameters, we also sample and marginalise over the fourteen nuisance parameters of the Planck likelihood and one bias parameter of the WiggleZ likelihood. We have sampled B up to -0.5, but nothing interesting was found beyond the upper value cited in this table.

The Bayesian evidence, Eq. (6), measures the predictivity of a model. The integral is bigger the more amount of likelihood mass falls inside regions with substantial prior probability. The evidence is penalised by the volume $\mathcal{V}$ of the parameter space allowed by the theory, since the prior density goes roughly like $\pi \sim \mathcal{V}^{-1}$. In turn, the Bayesian ratio quantifies the relative predictivity of two models given a data set: if its value is much smaller than one, the model M_0 is a more likely explanation of the data than the model M_1, and vice versa. In the frequentist approach, this is comparable to the increase of p-values[4] due to the *look-elsewhere effect*. For example, in particle physics, if one allows the predicted mass of a particle to vary within a broad range, the p-value of an apparent peak in particle production with a corresponding mass within this range will increase, just because a wider range of energies makes a random, non-physical peak-like feature more likely. Correspondingly, this indicates that the evidence of this model with a new parameter, like the new particle's mass, gets reduced.

In the particular case of localized primordial features in the CMB and LSS spectra, the Bayesian approach is motivated by the similarity that said features share with shot noise in the corresponding bands. This similarity, when the features are small, will result in the multi-modality of the likelihood of the corresponding parameters, and likelihood enhancements similar to those obtained by fitting the model to feature-less, noisy data. For example, for a specific linear oscillation template, using 5000 Planck-like, signal-less simulated CMB maps, the authors of [53] found that the noise could account for up to $\Delta\chi^2 \equiv 2\Delta\ln\mathcal{L} \sim 30$ at 3σ confidence level, with a typical enhancement of $\Delta\chi^2 \sim 10$ for the best fit of this kind of model. Considering this, it is not easy to assess whether we are fitting noise based on the likelihood enhancement only. Therefore, we focus on the predictivity of the models, given by their

[3] ΛCDM denotes the 6-parameter base model considered by the Planck collaboration [62].

[4] From Wikipedia.org, "a p-value is the probability of obtaining a test statistic result at least as extreme as the one that was actually observed, assuming that the null hypothesis is true. A researcher will often "reject the null hypothesis" when the p-value turns out to be less than a predetermined significance level, often 0.05 or 0.01. Such a result indicates that the observed result would be highly unlikely under the null hypothesis".