Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **455,** 2461–2473 (2016) doi:10.1093/mnras/stv2217

Bayesian model selection without evidences: application to the dark energy equation-of-state

S. Hee,[1,2]★ W. J. Handley,[1,2] M. P. Hobson[1] and A. N. Lasenby[1,2]

[1]*Astrophysics Group, Battcock Centre, Cavendish Laboratory, JJ Thomson Avenue, Cambridge CB3 0HE, UK*
[2]*Kavli Institute for Cosmology Cambridge, Madingley Road, Cambridge CB3 0HA, UK*

Accepted 2015 September 22. Received 2015 September 16; in original form 2015 June 30

ABSTRACT

A method is presented for Bayesian model selection without explicitly computing evidences, by using a combined likelihood and introducing an integer model selection parameter n so that Bayes factors, or more generally posterior odds ratios, may be read off directly from the posterior of n. If the total number of models under consideration is specified *a priori*, the full joint parameter space (θ, n) of the models is of fixed dimensionality and can be explored using standard Markov chain Monte Carlo (MCMC) or nested sampling methods, without the need for reversible jump MCMC techniques. The posterior on n is then obtained by straightforward marginalization. We demonstrate the efficacy of our approach by application to several toy models. We then apply it to constraining the dark energy equation of state using a free-form reconstruction technique. We show that Λ cold dark matter is significantly favoured over all extensions, including the simple $w(z) =$ constant model.

Key words: equation of state – methods: data analysis – methods: statistical – cosmological parameters – dark energy.

1 INTRODUCTION

Comparing two or more models given some data is central to the scientific method. The field of model selection within statistical inference attempts to address this problem, and numerous techniques for choosing between models exist, including: Akaike's Information Criterion (Akaike 1974), Schwarz's Bayesian Information Criterion (Schwarz 1978) and the Bayesian evidence (Jeffreys 1961; MacKay 2003). Here, we focus on Bayesian model selection using the evidence $\mathcal{Z}$ (also known as the prior predictive or marginal likelihood) and posterior odds ratios (PORs) $\mathcal{P}_{ij}$ (a generalization of the more commonly used Bayes Factors $\mathcal{B}_{ij}$), as this technique is inherent to Bayes theorem and both are widely used throughout cosmology and astrophysics (Liddle, Mukherjee & Parkinson 2006).

PORs provide a quantitative means for selecting between models and are usually calculated directly from the evidence of each model. In higher dimensions, techniques to calculate evidences include thermodynamic integration (Gelman & Meng 1998, also known as simulated annealing), approximations to the evidence when certain favourable conditions are met (such as unimodality and Gaussianity; Tierney & B. 1986; Liddle et al. 2006) and nested sampling (Skilling 2004, 2006; Sivia & Skilling 2006). Calculating Bayes factors directly, without calculating $\mathcal{Z}$ for each model, is also possible using the Savage–Dickey density ratio for nested models (Verdinelli & Wasserman 1995, where a more complex model reduces to the simpler by setting its additional parameters appropriately). A good review from before nested sampling's rise in popularity can be found in Clyde et al. (2007); for a thorough review of these methods in cosmology see Trotta (2008).

In this paper, we propose a method to calculate PORs without the problems associated with evidence calculations or simplifying assumptions. PORs are calculated directly from a set of models explored simultaneously without constraints on the forms these models might take. The new method circumvents the challenges associated with accurate evidence calculations by computing PORs using Bayesian parameter estimation, which is typically a more reliable and computationally less expensive task. Additionally, parameter estimation algorithms are more commonly used and therefore the method provides an easy means for extending existing codes to the domain of model selection. This is achieved by introducing a parameter that selects between models, and allows the calculation of PORs from the posterior probability of this parameter. We note that similar approaches have been proposed previously (Hobson & McLachlan 2003; Goyder & Lasenby 2004; Brewer & Donovan 2015), but these typically rely on the use of sampling techniques capable of jumping between parameter spaces of different sizes, such as reversible jump Markov chain Monte Carlo (MCMC; Green 1995), which requires special sampling methods that are often very computationally demanding. Our approach is much simpler, requiring no special sampling methods, provided the number of models under consideration is specified a priori, and is related to the class

★E-mail: sh767@cam.ac.uk

of product–space MCMC methods originally proposed by (Carlin & Chib 1995, see also Sisson 2005; Lodewyckx et al. 2011).

We apply our method to toy models and the cosmological problem of constraining the dark energy equation of state, with particular emphasis on determining the complexity supported by data for deviations from Λ cold dark matter (ΛCDM). In both cases, we are solving the problem of how many nodes are required in a piecewise linear model to reconstruct a one-dimensional function. With the number of nodes defining the models, we show explicitly that this new method agrees with the evidences-based approach for calculating PORs.

The rest of the paper is organized as follows. Section 2 provides a brief statistical overview of PORs and evidence calculation. Section 3 discusses the statistical framework for calculating posteriors odds ratios using parameter estimation instead of calculating evidences. Thereafter, results are presented in Section 4 for a toy model data fitting problem and in Section 5 for the cosmological problem of characterizing the dark energy (DE) equation of state parameter as a function of redshift using recent cosmological data sets. We summarize our findings and conclude in Section 6.

2 BACKGROUND

Bayes Theorem (Bayes & Price 1763; MacKay 2003; Sivia & Skilling 2006) states that

$$\Pr(X|Y,I) = \frac{\Pr(Y|X,I)\Pr(X|I)}{\Pr(Y|I)}, \tag{1}$$

where X and Y are propositions, $\Pr(X)$ specifies our belief that the proposition is true, and I is the background information. Using this, we can calculate the probability that a set of parameters θ of a model $\mathcal{M}$ takes specific values given some data $\mathcal{D}$ to constrain them (note we drop the dependence on I as it is implicit throughout):

$$\Pr(\theta|\mathcal{D},\mathcal{M}) = \frac{\Pr(\mathcal{D}|\theta,\mathcal{M})\Pr(\theta|\mathcal{M})}{\Pr(\mathcal{D}|\mathcal{M})} \equiv \frac{\mathcal{L}\pi}{\mathcal{Z}}, \tag{2}$$

where $\mathcal{L}$, π and $\mathcal{Z}$ are shorthands for the likelihood, prior, and evidence, respectively. This is Bayesian parameter estimation, where $\Pr(\theta|\mathcal{D},\mathcal{M})$ is the posterior probability distribution. Similarly, we can calculate the probability of a model given some data:

$$\Pr(\mathcal{M}|\mathcal{D}) = \frac{\Pr(\mathcal{D}|\mathcal{M})\Pr(\mathcal{M})}{\Pr(\mathcal{D})} = \frac{\mathcal{Z}\pi_{\mathcal{M}}}{\Pr(\mathcal{D})}. \tag{3}$$

Taking the ratio of the probabilities of two models signifies our degree of belief in one model over another. Taking the logarithm of this ratio and using equation (3) above gives us PORs:

$$\mathcal{P}_{ij} = \ln\left[\frac{\Pr(\mathcal{M}_j|\mathcal{D})}{\Pr(\mathcal{M}_i|\mathcal{D})}\right] = \ln\left(\frac{\mathcal{Z}_j}{\mathcal{Z}_i}\right) + \ln\left(\frac{\pi_{\mathcal{M}_j}}{\pi_{\mathcal{M}_i}}\right). \tag{4}$$

If $\pi_{\mathcal{M}_i} = \pi_{\mathcal{M}_j}$, then $\mathcal{P}_{ij} = \mathcal{B}_{ij}$, the Bayes factor, which is more commonly used in the literature despite being a less general treatment than the fully Bayesian PORs that also considers the prior probability of each model. For both, criteria to give meaning to this quantification are given by the Jeffreys guideline (Jeffreys 1961), shown in Table 1. Model selection using Bayesian statistics thus requires the calculation of ratios of evidences. Typically, the evidences are first calculated separately and their ratios evaluated. Calculating the evidence for each model is inherently difficult. From equation (2), we see that $\mathcal{Z}$ is a normalization constant for $\Pr(\theta|\mathcal{D},\mathcal{M})$, allowing us to calculate it as

$$\mathcal{Z} = \int_{\text{all }\theta} \mathcal{L}(\theta)\pi(\theta)\,\mathrm{d}\theta. \tag{5}$$

Table 1. Jeffreys guideline for interpreting PORs. As $\mathcal{P}_{ji} = -\mathcal{P}_{ij}$, negative PORS imply reversed model favouring.

POR	Favouring of $\mathcal{M}_j$ over $\mathcal{M}_i$
$0.0 \leq \mathcal{P}_{ij} \leq 1.0$	None
$1.0 \leq \mathcal{P}_{ij} \leq 2.5$	Slight
$2.5 \leq \mathcal{P}_{ij} \leq 5.0$	Significant
$5.0 \leq \mathcal{P}_{ij}$	Decisive

Equation (5) is a multidimensional integral over the whole parameter space of a model. Computationally, it is not possible to calculate these by brute force even for modest dimensionalities, and the techniques mentioned in the introduction have been developed as an alternative means to do so. The most promising of these techniques is nested sampling, and with steady advances made in both computing power and algorithms to implement nested sampling, many cosmological and astrophysical model selection problems can now be solved by computing evidences, which is the current standard practice.

3 METHOD

We propose a method here for calculating PORs, using parameter estimation techniques, that avoids calculating evidences directly. The method places no constraints on the models that can be considered and has the advantage of being simple to implement and undisruptive for members of the community familiar with Bayesian parameter estimation techniques.

Consider a number of different models $\mathcal{M}_n$ ($n = 1, 2, \ldots, N$). We combine these into a single hypermodel $\mathcal{M}$. The parameters of $\mathcal{M}$ are the integer variable n that 'switches' between the models $\mathcal{M}_n$, and the union θ of the parameter vectors θ_n of each individual model. Note that, if there is some overlap between the parameter vectors θ_n and $\theta_{n'}$ of two different models, then the coincident parameters are notionally included only once in the union θ. In practice, the parameter n can be implemented as a continuous parameter and a suitable binning used to convert it to an effective integer parameter, thereby simplifying the implementation (provided the technique used to explore the parameter space does not rely on gradient information). Indeed, the implementation of our approach is, in general, straightforward, since one needs only to write a simple 'wrapper' hyperlikelihood function for $\mathcal{M}$, which calls the existing likelihood function for the appropriate individual model $\mathcal{M}_n$ depending on the (integer) value of n.

In general, the parameter vectors θ_n and $\theta_{n'}$ for different models will be of different dimensionalities. In the case of nested models, where $\theta_n \subset \theta_{n+1}$, such problems are usually accommodated using reversible-jump Markov chain Monte Carlo (RJMCMC) methods, which are capable of making transitions between spaces of different dimensionality. In principle, such methods might also be used in the case of non-nested models, even in the extreme case where θ_n and $\theta_{n'}$ have no parameters in common, although such applications have not been widely explored.

Here, we adopt a different approach that accommodates nested and non-nested models equally well, including the extreme case mentioned above, and avoids the algorithmic complication and computational expense of RJMCMC methods. The only assumption required is that N (the number of models under consideration) is known a priori. Although this seems an innocuous requirement, it does constitute a mild limitation. Consider, for example, the classic

nested problem of fitting a polynomial of unknown degree to a set of (x, y) data points. In our approach, one is required to fix the maximum allowed degree N of the polynomial in advance, whereas this is not necessary in the traditional RJMCMC approach. None the less, in realistic applications such a limitation is not too severe.

By fixing N, the full parameter space (θ, n) is determined a priori, and is of fixed dimensionality, so it may be explored using standard sampling methods, such as MCMC or nested sampling (MacKay 2003; Skilling 2006; Brewer, Pártay & Csányi 2011). Explicitly, suppose at some MCMC step or nested sampling iteration one considers the point (θ, n), possibly after suitable binning of the continuous parameter n to obtain an integer value. For any given value of n so obtained, the union parameter space may be partitioned into those parameters θ_n on which the model $\mathcal{M}_n$ depends and the remaining parameters ϕ_n that are not used by $\mathcal{M}_n$. The 'wrapper' hyperlikelihood function thus may pass only the parameters θ_n to the likelihood function for the appropriate model $\mathcal{M}_n$. The remaining parameters ϕ_n are thus 'ignored', which is equivalent to assigning a constant likelihood value over this subspace. By considering the full space (θ, n), however, the sampling method will typically need to accommodate moderate to large dimensionality, most likely possessing multiple modes and/or pronounced degeneracies. In practice, nested sampling is well suited to such problems, and therefore we adopt it here.

Once one has obtained a set of posterior samples from the space (θ, n), one may calculate $\Pr(n|\mathcal{D}, \mathcal{M})$ by simply marginalizing out all other parameters to produce a marginalized posterior probability:

$$\Pr(n|\mathcal{D}, \mathcal{M}) = \int \Pr(\theta, n|\mathcal{D}, \mathcal{M})\, \mathrm{d}\theta \tag{6}$$

$$= \frac{1}{\mathcal{Z}_{\mathcal{M}}} \int \mathcal{L}(\theta, n)\, \pi(\theta, n)\, \mathrm{d}\theta, \tag{7}$$

where $\mathcal{Z}_{\mathcal{M}}$ is the evidence for this hypermodel $\mathcal{M}$. Since for any given value of n the union parameter space may be partitioned into those parameters θ_n on which the model $\mathcal{M}_n$ depends and the remaining parameters ϕ_n that are not used by $\mathcal{M}_n$, one may write the likelihood in equation (7) as $\mathcal{L}(\theta_n)$ and the priors as $\pi(\theta|n) = \pi(\theta_n|n)\pi(\phi_n)\pi(n)$, where $\pi(n) \equiv \Pr(n|\mathcal{M})$. Hence equation (7) becomes

$$\Pr(n|\mathcal{D}, \mathcal{M}) = \frac{\pi(n)}{\mathcal{Z}_{\mathcal{M}}} \int \mathcal{L}(\theta_n)\, \pi(\theta_n|n)\, \mathrm{d}\theta_n, \tag{8}$$

where we have used the fact that the integral over the priors for unused parameters is unity, namely $\int \mathrm{d}\phi_n\, \pi(\phi_n) = 1$. We recognize the integral in equation (8) as the evidence $\mathcal{Z}_n$ of the model $\mathcal{M}_n$, so that we have

$$\pi(n)\mathcal{Z}_n = \mathcal{Z}_{\mathcal{M}} \Pr(n|\mathcal{D}, \mathcal{M}). \tag{9}$$

We are interested in the PORs between two models, $\mathcal{M}_i$ and $\mathcal{M}_j$:

$$\mathcal{P}_{ij} = \ln\left[\frac{\Pr(n{=}j|\mathcal{D}, \mathcal{M})}{\Pr(n{=}i|\mathcal{D}, \mathcal{M})}\right], \tag{10}$$

where the $\mathcal{Z}_{\mathcal{M}}$ cancels. Thus, the POR is given simply by the ratio of values of the posterior $\Pr(n|\mathcal{D}, \mathcal{M})$ for the two models, which is obtained using the parameter estimation formulation of Bayes theorem and the process of marginalization, without the need to calculate evidences directly. The key feature is that the unused parameters ϕ_n marginalize out to unity. Moreover, the posteriors on ϕ_n should simply equal the priors on ϕ_n. Visual inspection of these posteriors thus provides a useful check that the method is performing correctly.

A potential downside to this method is the requirement that the prior probabilities of the models are specified in advance. For signal detection problems with an unknown number of sources, for example Hobson & McLachlan (2003) and Feroz & Skilling (2013), this is in principle undesirable but in practice a suitable prior choice can always be found. Additionally, if calculating PORs for another model $\mathcal{M}_{N+1}$ was desired, after having completed the analysis for the first N models, then a repetition of the method with only this new model and the most favourable model is possible, at a computational cost of exploring the most favourable model[1] a second time.

It is also important to note, however, that our new method does not produce an estimate of the error on the PORs in a single computation, whereas this is possible when calculating evidences directly using nested sampling. Throughout we therefore use multiple repeat runs to obtain an error on the PORs.

4 APPLICATION TO TOY-MODELS

In this section, we demonstrate our approach by applying it to some toy-models and in the next section we apply our method to constraining the DE equation of state as a function of redshift using recent cosmological data sets.

In both applications, we seek to model a one-dimensional function $y(x)$ using a piecewise linear interpolation scheme between a set of nodes and ask the model selection question 'how many nodes are needed to fit the data?'. Thus we place a set of nodes $y_i(x_i)$ in the plane, where the amplitude y_i and the position x_i are model parameters to be varied. At $x_{\min}$ and $x_{\max}$, fixed-position nodes are placed with varying amplitude only, such that for the model defined by n internal nodes there are $2 + 2n$ parameters. As shown in Fig. 1, linear interpolation is used to construct y at all points (with $y(x)$ set constant outside the range $[x_{\min}, x_{\max}]$). Of course, other interpolation schemes between nodes may be used, such as splines, although we do not consider these here. The application of these approaches to constraining $w(z)$ is described by Vázquez et al. (2012b).

A specific model is defined by how many nodes are used in reconstructing $y(x)$. Comparing multiple models with increasing numbers of nodes identifies how many nodes are needed to fit the data, in other words the preferred complexity inherent in the data. As the final result, one can plot either $\Pr(y|x, n_\star)$, where $n_\star$ denoted the number of nodes in the most favoured model, or $\Pr(y|x)$ averaged over all models weighted by their PORs (Parkinson & Liddle 2013; Planck Collaboration XX 2015). Either approach identifies clearly the nature of the data constraints on $y(x)$.

The key strength of the reconstruction is its free-form nature, which can capture any shape of function in the $y(x)$ plane by adding arbitrarily large numbers of nodes. Providing the model selection criterion penalizes overcomplex models appropriately by weighing 'goodness-of-fit' against the numbers of parameters in the model (Occam's Razor), identifying how much complexity the data support is performed in a clear and unambiguous manner by the favoured number of nodes. Model selection techniques can thus be used to solve questions on the constraining power of the data, as successfully shown in various cosmological applications (Vázquez et al. 2012a,b; Planck Collaboration XX 2015).

[1] The most favourable is best used, in light of discussions on the size of error bars in Section 4.2.

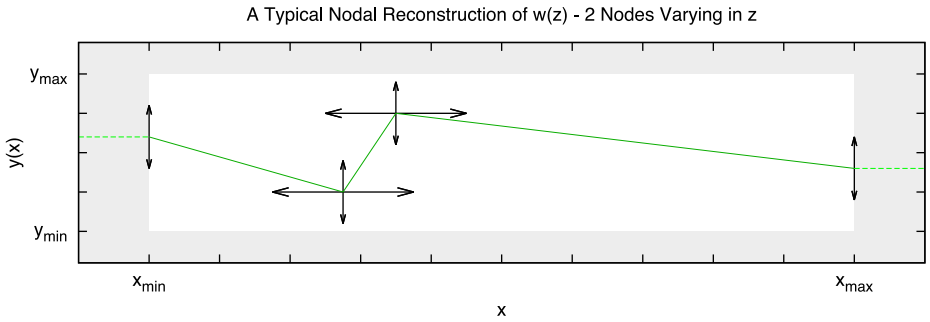
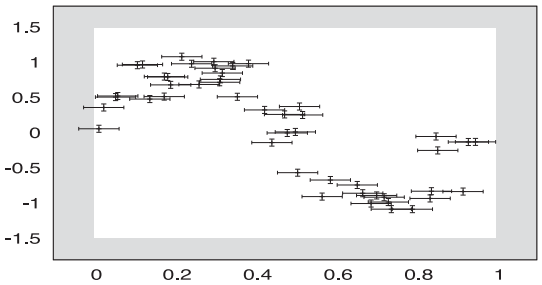
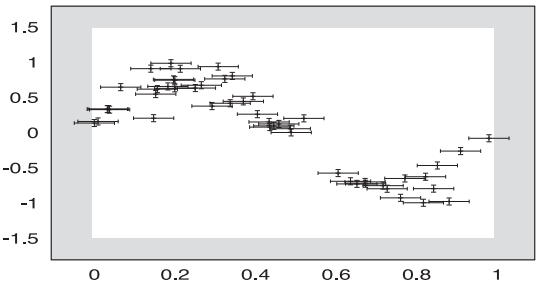


Figure 1. Illustration of the nodal reconstruction, which flexibly allows the parameter estimation process to define the preferred shape of $y(x)$ from the data by linearly interpolating nodes whose amplitudes, positions (for internal nodes) and number can vary as required. The figure shows the interpolation process, and highlights how nodes can be positioned inside the unshaded prior space (with sorting of node positions such that $x_i < x_{i+1}$).



(a) $\sin(2\pi x)$: 47 points sampled from the $\sin(2\pi x)$ function.



(b) $\mathrm{line}(2\pi x)$: 49 points sampled from the $\mathrm{line}(2\pi x)$ function.

Figure 2. Data points plotted in the (x, y) plane for each data set (a) and (b). The unshaded region represents the prior space for the y_i amplitudes and x_i positions of the nodes, over which a uniform prior is assumed (with sorting of the node position parameters such that $x_i < x_{i+1}$).

The nodal reconstructions are clearly nested models. Since our general approach does not require this, for completeness we also consider a non-nested model selection problem by comparing a 2-internal node reconstruction with a sinusoidal model. The rest of this section presents the results obtained and highlights further strengths and weaknesses of our approach.

4.1 Fitting a function to data

Consider a set of $j_{\max}$ data points $\{(x_j, y_j), j = 1, \ldots, j_{\max}\}$ with experimental errors $\{(\sigma_{x_j}, \sigma_{y_j})\}$ on each of the points. Assuming there is a functional relationship between the independent variable x and dependent variable y, captured by $y = f(x)$, then the likelihood of observing these data is given by

$$\Pr(\{x_j, y_j\}|\{\sigma_{x_j}, \sigma_{y_j}\}, f, X_-, X_+) = \prod_{j=1}^{j_{\max}} \int_{X_-}^{X_+} \mathrm{d}X_j \, \frac{\exp\left[-\frac{(x_j - X_j)^2}{2\sigma_{x_j}^2} - \frac{(y_j - f(X_j))^2}{2\sigma_{y_j}^2}\right]}{2\pi\sigma_{x_j}\sigma_{y_j}(X_+ - X_-)}, \quad (11)$$

where X_-, X_+ are the end points of the uniform region in which the data points may be found a priori. A Bayesian derivation of this likelihood can be found in Appendix A; for more detail see Sivia & Skilling (2006). The integral is calculated numerically using standard quadrature techniques.

Given the data, the Bayesian approach is to use this likelihood to infer the probability distribution of the parameters in some parametric form of the function f. We will do this for the family of functions described above, and use PORs to determine how many nodes optimally reconstruct the function.

We test two different data sets, shown in Fig. 2. The traditional evidence-based approach and our new method for calculating PORs are compared for each data set. The constraints on $y(x)$ given the data are also discussed.

Data set (a) has 47 data points drawn uniformly in x from the function $y = \sin(2\pi x)$ in the range $x \in [0, 1]$, with each point adjusted in x and y by random Gaussian noise with mean $= 0$ and $\sigma = 0.05$ (error bars on data points are σ).[2] Data set (b) has 49 data points drawn as in (a) but from a piecewise-linear function coinciding with the function $y = \sin(2\pi x)$ at $x = 0, 0.25, 0.75, 1$, so that it is very difficult by eye to distinguish the two data sets as being drawn from different functions. We call the function used in (b) $\mathrm{line}(2\pi x)$ for brevity. Clearly, a linearly interpolated nodal model with $n = 2$ internal nodes can represent this function exactly.

For each of the data sets, we test models with one internal node up to seven internal nodes (i.e. three total nodes up to nine total nodes or two line segments up to eight line segments), using POLYCHORD (Handley, Hobson & Lasenby 2015) to calculate evidences (the vanilla method henceforth) and again using POLYCHORD to implement the new method (Post(n) method henceforth).[3] POLYCHORD is a

[2] 50 points were drawn initially for each data set, but some fellow outside the prior range due to the Gaussian noise, and were not included.

[3] Note the marginalized posterior probability on n is calculated from the *chain_unnormalized.txt* file using the standard nested sampling technique (Skilling 2006). It is important to use this file over the usual *chain.txt* file and set up POLYCHORD to output all interchain points of the algorithm. This ensures good reconstruction of $\Pr(n|\mathcal{D}, \mathcal{M})$ over the lower probability regions in light of the computing 'log-sum-exp' problem.

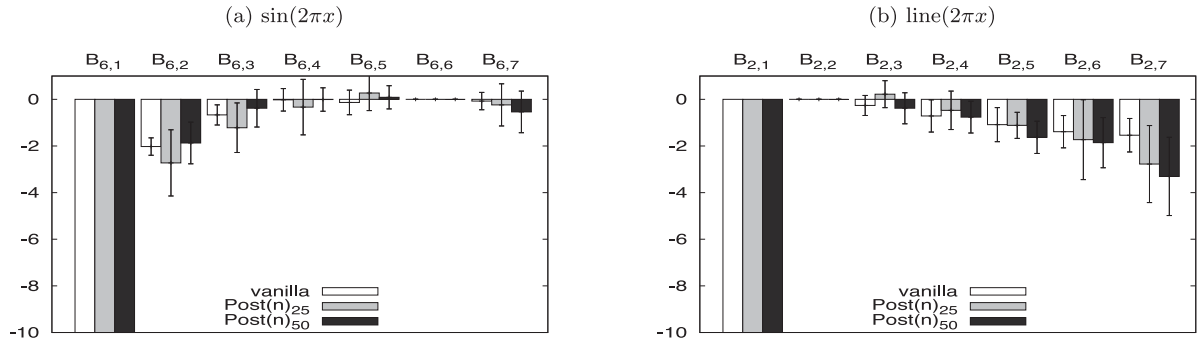


Figure 3. PORs (or Bayes factors) for data sets (a) and (b) defined by Fig. 2. $\mathcal{B}_{n,n'}$ denotes the Bayes factor for the models with n and n' internal nodes. Histograms represent PORs with respect to the most probable model. White, light grey and dark grey bars are for the vanilla, Post$(n)_{25}$ and Post$(n)_{50}$ results, respectively. Error bars shown are sample standard deviations obtained from 10 repeat trials. The PORs agree well between methods.

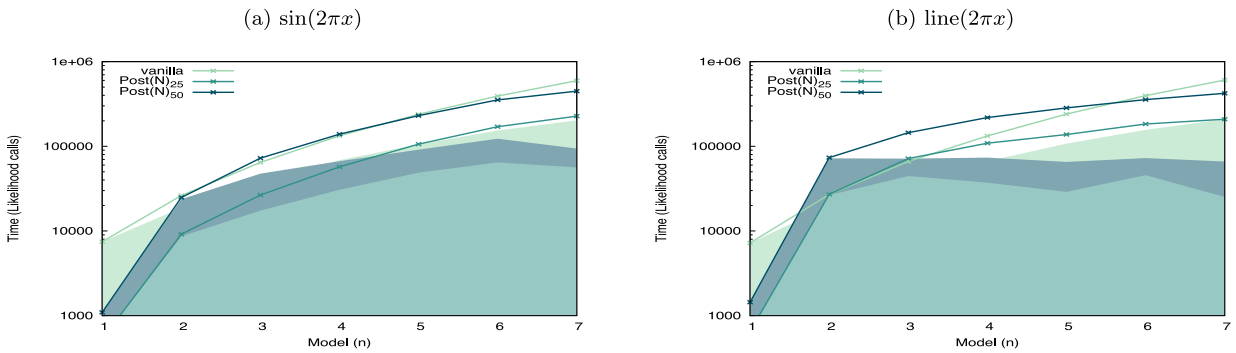


Figure 4. Average timing data for data sets defined in Fig. 2 and the vanilla, Post$(n)_{25}$ and Post$(n)_{50}$ results defined in the text. The shaded regions show the approximate number of likelihood calculations made for each model n and the solid lines show the cumulative numbers. More detail and an analysis of the timing benefits of using our new method are given in Appendix B. Considering error bars on the PORs for the different methods, it is clear that the Post$(n)_{50}$ method (darkest plots) can produce comparable accuracy in less likelihood calls than the vanilla method (lightest plot).

relatively new nested sampler and was found to be very suitable for this problem. We use uniform priors on the y amplitudes of nodes, and sorted uniform priors on the x position parameters of nodes, where the x priors are uniform but forced to adhere to $x_i < x_{i+1}$ to avoid the scenario where the n internal nodes are interchangeable with each other. We assign equal prior probabilities for each model, so PORs are equal to Bayes factors.

Each data set is analysed 10 times for each method to determine the statistical uncertainty on the derived PORs. In each case, the PORs are normalized to the model with the highest evidence in the vanilla method. Errors on the PORs are given as the sample standard deviation from the 10 repeats. POLYCHORD was run with $N_{\rm live} = 25N_{\rm dim}$ live points initially to obtain the results labelled Post$(n)_{25}$, where $N_{\rm dim} = 2n + 2$ is the number of parameters to be explored (the dimension of the space) and the number of live points, $N_{\rm live}$, is the only tuning parameter associated with the POLYCHORD sampling algorithm. To highlight accuracy and timing considerations when using the method, we also repeat the analysis with $N_{\rm live} = 50N_{\rm dim}$ to obtain the results labelled Post$(n)_{50}$.

4.2 Results for nested nodal models

The PORs (or Bayes factors) for the vanilla method with $N_{\rm live} = 25N_{\rm dim}$ and Post(n) method with $N_{\rm live} = 25N_{\rm dim}$ and $50N_{\rm dim}$, per data set, are shown in Fig. 3 and show good agreement between the two methods regardless of $N_{\rm live}$. From this we conclude that the methods produce consistent PORs. As one might expect, for the line$(2\pi x)$ data set, the preferred model has $n = 2$ internal nodes, whereas a larger number of nodes is preferred for the sin$(2\pi x)$ data set. The timing data in Fig. 4 suggests that Post$(n)_{25}$ results were faster to obtain by about a factor of 2.5 when using the same $N_{\rm live}$ per parameter, however this comes at a cost in accuracy as the errors on the vanilla PORs are clearly tighter than the Post$(n)_{25}$ results. Post$(n)_{50}$, however, takes less time to produce similar accuracy for the significant PORs. In general, we observe that our method can produce Bayes factors faster than the vanilla method in a systematic manner, and discuss this in Appendix B. Furthermore, the reconstructions of the favoured models for each method are shown in Figs 5 and 6, respectively. The reconstructions are identical in all key features between methods. The Post$(n)_{50}$ graph is not plotted as it was very similar.

The important discrepancies between the vanilla and Post(n) methods are in the errors on the PORs, where we have identified two issues: first for large negative PORs the errors from the Post(n) method are quite large and, secondly, the errors on the vanilla method are tighter for equivalent $N_{\rm live}$. The first discrepancy might be expected given that POLYCHORD, and nested samplers in general, rapidly converge to the central peak(s) in a distribution, thus spending less time in lower likelihood regions and sampling those regions proportionately less thoroughly. Given that each model investigated is a separate mode in the computation, a model with low likelihood will be less thoroughly explored than the models with larger likelihoods – making the calculation of $\Pr(n|\mathcal{D}, \mathcal{M})$ less

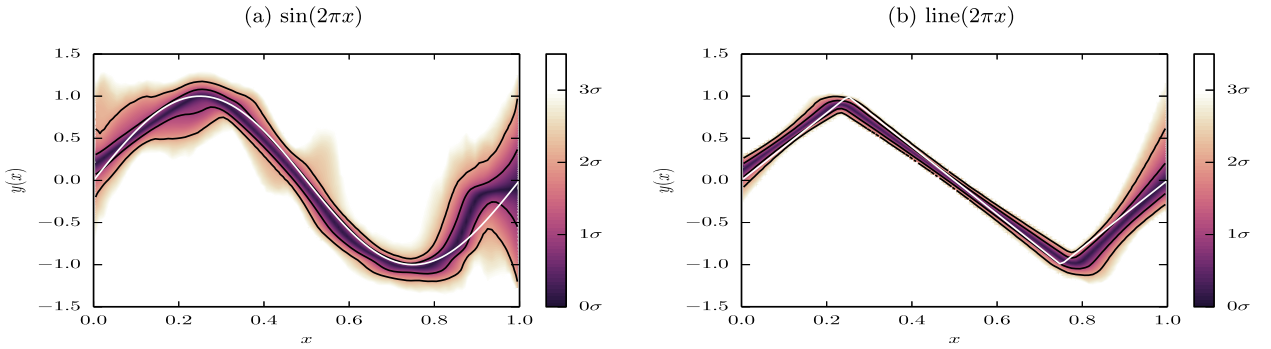


Figure 5. Reconstructions of $y(x)$ using the vanilla method of explicitly calculating evidences to obtain PORs. Plots are from one of the 10 trials, arbitrarily chosen, and are of the model with the largest POR, i.e. (a) six internal node model, (b) two internal node model. Each figure shows the posterior probability $\Pr(y|x, \mathcal{D}, \mathcal{M})$, in normalized slices of constant x to show the deviation from the peak y at each x, binned in 100 bins in both x and y. The color bars to the right show the iso-probability confidence intervals at a given slice in x Sivia & Skilling (2006), see Planck Collaboration XX (2015) Section 8.2 equation (68) for details. The 1σ and 2σ intervals are plotted as black lines for clarity and the cube-helix colour scheme by Green (2011) is used for linearity in grey-scale. In white is plotted the underlying function from which the data was sampled, and even with less than 50 data points a good reconstruction is obtained.

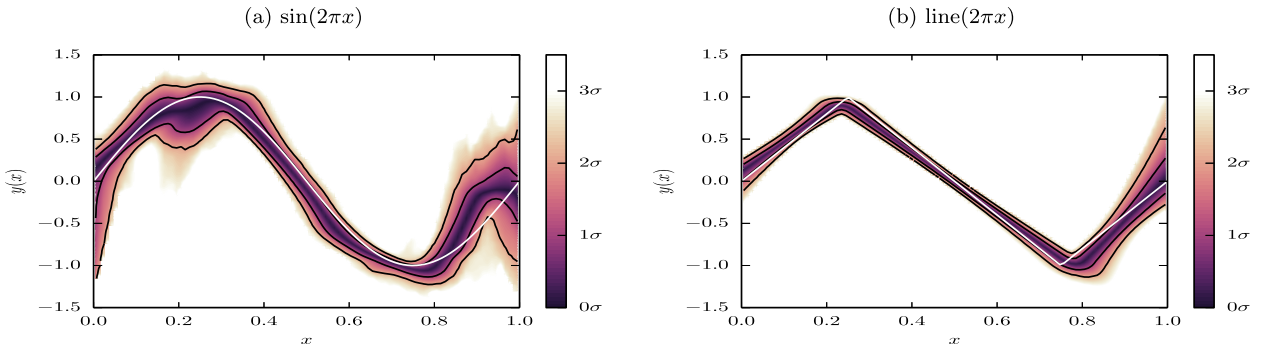


Figure 6. Reconstructions of $y(x)$ for the Post$(n)_{25}$ results to obtain PORs. Plots are for comparison to the vanilla results of Fig. 5, and are plotted in the same way. The Post$(n)_{25}$ results agree well with the vanilla method results in all key features.

reliable for these models. This is, however, desirable behaviour. Spending compute time only on probable models reduces the overall time taken to find the most probable model(s), whilst the less probable models are still sampled sufficiently well to identify them as less probable.

The second discrepancy is more significant but equally predictable. The number of live points in POLYCHORD defines how fully the space is explored. For the vanilla method, the $N_{\rm live} = 25N_{\rm dim}$ calculation provides adequate sampling per model, whilst for the Post(n) method a similar number of live points needs to explore several models simultaneously, effectively reducing the live points available to explore each model and producing larger errors. This suggests that users need to ensure that algorithm tuning parameters such as $N_{\rm live}$ are chosen appropriately and check that the results on repetitions of the algorithm are consistent. The Post$(n)_{50}$ results demonstrate clearly that results are confidently extracted in comparable compute-times when best practice is adhered to. Being aware of the increased modality of the space that is inherent to the method and ensuring that the sampling algorithm adequately handles such complex parameter spaces helps ensure accurate results.

Finally, it is worth making some brief comments on the 'physical' results of the model selection process for each of the data sets. In data set (a), a more complex underlying shape in $y(x)$ is identified needing more nodes than data set (b), consistent with the distinction between $\sin(2\pi x)$ and $\mathrm{line}(2\pi x)$. It should be noted too that overfitting (adding more parameters than needed) is not heavily penalized for data set (b), as observed in the slow decrease in Bayes factors after the favoured model is found – this is standard behaviour (Sivia & Skilling 2006, p. 93) and can be understood by considering the Occam factor associated with a parameter which is constrained without increasing the fit of the model (MacKay 2003, p. 349). In general, the model selection and nodal reconstruction technique produces strong conclusions on the shapes of the $y(x)$ plane, given the data in each case, and clearly identifies the inherent complexity of the various data sets, as we desired it to.

4.3 Results for non-nested models

Our new method does not require that the models be nested. A model is nested inside another 'larger' model if setting some parameters to specific values in the larger model allows one to obtain the smaller nested model. The nodal reconstructions are clearly nested in this sense. Here, we quickly demonstrate that our method also works for non-nested models.

We test data sets (a) and (b) against two models. The first model is the sinusoid function $y(x) = A\sin(2\pi Bx + C) + D$ and the second model is the two internal node reconstruction, so that we expect data set (a) to favour the sinusoidal model and (b) to favour the linear model. Parameters A and B are scale parameters for the amplitude and frequency, respectively; we assign to these logarithmic priors in the range $[0.1, 5]$. Parameters C and D are shift parameters and we assign uniform priors in the ranges $[-\pi, \pi]$ and $[-1.5, 1.5]$, respectively. These priors reflect sufficient coverage of the prior

space defined in Fig. 2 and are adequate for comparing the vanilla and new methods. It is important to note that in this test, both the vanilla method and Post(n) method used $N_{\rm live} = 25N_{\rm dim}$. For the vanilla method, this resulted in $N_{\rm live} = 100$ for the sinusoidal model and $N_{\rm live} = 150$ for the four node model, whilst for the Post(n) method the parameters were searched simultaneously (along with n) to give 11 parameters and $N_{\rm live} = 275$.

The PORs for data set (a) favour the sinusoid by 1.94 ± 0.93 and 2.01 ± 1.08 units, for vanilla and Post(n) methods, respectively. The PORs for data set (b) favour the linear model by 13.82 ± 1.02 and 14.87 ± 2.58 units, respectively, for vanilla and Post(n) methods. Taking into account the previous discussion, it is clear that the new method produces PORs consistent with the vanilla method. The Post(n) method here was about 5 per cent slower for data set (a) and 30 per cent slower for data set (b). However, with the significantly larger number of live points that the Post(n) method used, the fact that the methods are of comparable time is a desirable result and suggests that the unconstrained parameters for a given n are not significantly increasing the compute time of those isolated nodes in the parameter space.

In general, we conclude that the discussions in Section 3 regarding unconstrained parameters is correct. When parameters were reviewed for the chains files produced in a given model, the parameters that were not used by that model were distributed according to their priors. This is one of the core strengths and novelties of the method and allows PORs to be calculated without constraints on the models to be compared. This verifies that the method works for non-nested models, and we proceed now to apply it to a cosmological application using the nodal reconstruction.

5 APPLICATIONS TO THE DARK ENERGY EQUATION OF STATE

Having validated our approach on a toy problem, we now apply our method to a cosmological application, for which the vanilla method is not computationally suited. The aim is to demonstrate the method in a typical model selection application to obtain PORs efficiently and with estimates of the error that do not require excessive repetition of long computations. We probe the DE equation of state parameter $w(z)$ as a function of redshift to update the work of Vázquez et al. (2012b), using more modern data sets. We further showcase the usefulness of the nodal reconstruction approach, briefly described in Section 4 and more fully in Vázquez et al. (2012b), in defining the complexity supported by the data and identifying features in $w(z)$, adding to the list of papers using the reconstruction (Vázquez et al. 2012a,b; Aslanyan et al. 2014; Planck Collaboration XX 2015).

5.1 Method

We combine CMB data from the Planck 2013 data release (Planck Collaboration XV 2014a; Planck Collaboration XVI 2014b; Planck Collaboration XVII 2014c) with the *Wilkinson Microwave Anisotropy Probe WMAP* 9-yr polarization data (Bennett et al. 2013), baryonic acoustic oscillation (BAO) from the BOSS data release 11 (Anderson et al. 2014) and supernovae Type Ia (SNIa) data from the Union 2.1 catalogue (Suzuki et al. 2012) to provide constraints on DE behaviour. We focus on the redshift range $z \in [0, 2]$ in the reconstruction, where we set to constant values $w(z) = w(2)$ when $z > 2$. We use the COSMOMC code package (Lewis & Bridle 2002), which contains the CAMB code (Lewis, Challinor & Lasenby 2000; Howlett et al. 2012), and substitute the MCMC sampler for the MULTINEST nested sampling plugin running in constant efficiency mode (Feroz & Hobson 2008; Feroz, Hobson & Bridges 2009; Feroz et al. 2013), which is a well-established nested sampling implementation for evidence calculations and parameter estimation, and was the sampler used by Vázquez et al. (2012a,b) thereby enabling a direct comparison. To facilitate deviations away from the standard ΛCDM equation of state parameter $w = -1$, we implement the 'Parameterized Post-Friedmann' framework (PPF) modification to CAMB (Fang, Hu & Lewis 2008). For further details on the method and data sets see Vázquez et al. (2012b) and Planck Collaboration XVI (2014b), respectively.

Using PORs to identify the optimal number of nodes tells us the complexity of $w(z)$ features supported by the data. Further, the nodal reconstruction, as shown in the toy model, is highly adept at identifying constraints in the (w, z) plane. Of particular interest is whether deviations in $w(z)$ away from the successful ΛCDM cosmological model are supported by modern data and to identify which DE extensions are favoured. Theories incorporating deviations from $w = -1$ include quintessence scalar fields for $w > -1$ (Ratra & Peebles 1988; Caldwell, Dave & Steinhardt 1998; Tsujikawa 2013) and phantom DE models with supernegative $w < -1$ (Caldwell 2002; Sahni 2005). The possibility of crossing of the phantom divide line at $w = -1$ in dynamical models has also been considered (Zhang 2009). Modified gravity or brane-world models also make predictions about $w(z)$ (Sahni 2005). Thus, paramount to understanding DE is determining $w(z)$.

To do this, we compare six models, in order of increasing complexity: ΛCDM with $w = -1$, wCDM with w constant in z but allowed us to vary in amplitude, *tilt*CDM with $w(z = 0)$ and $w(z = 2)$ allowed us to vary and linear interpolation for $w(z)$ between them (0 internal node model), and then nodal models with 1, 2 and 3 internal nodes, respectively. Models are abbreviated to Λ, w, t, 1, 2 and 3, respectively, where appropriate. Priors on each w parameter are uniform on the range $[-2, 0]$ and were chosen to be conservative, we did not check the robustness of results with respect to prior choice and leave this for future work, see Vázquez et al. (2012b) for such an analysis. Priors on each z parameter are uniform on [0, 2] such that for more than one internal node $z_i < z_{i+1}$ (i.e. sorted uniform priors as in the toy model). The previous work by Vázquez et al. (2012b) found that ΛCDM was favoured, whilst the two internal node model had the second largest evidence, pointing to structure in $w(z)$ that could not be captured by a constant equation of state parameter wCDM, or even the one internal node model. Here, we show clearly that Planck 2013 era data sets do not have this feature and only ΛCDM can be considered favoured.

An important point is that the Planck data require the addition of 14 so called nuisance parameters. These must be sampled and, together with the six parameters of CDM models, produce an at least 20-dimensional parameter space. As MULTINEST is a rejection nested sampling algorithm, it is expected that computation times increase significantly in higher dimensions as the volume on the shell increases.[4] MULTINEST has the algorithm search parameters $N_{\rm live}$ and eff, where decreasing eff (in constant efficiency mode) typically achieves more accurate results more effectively than increasing $N_{\rm live}$.

[4] Specifically, it constructs multidimensional ellipsoids to estimate sampling within an iso-likelihood region, as required by nested sampling. The ellipsoids expand by a fraction to ensure no viable regions of the true iso-likelihood contour are outside this estimate. Points are sampled inside these ellipsoids and rejected until meeting the nested sampling criterion.

Table 2. The 30 priors that define the parameter space. The top set of parameters are the CDM parameters, the middle ones show the nuisance parameters associated with the Planck 2013 data release, and the bottom set are the parameters introduced by dark energy model extensions, including n for selecting between models and $\theta_{\rm uniform}$ for testing a MULTINEST edge-effect problem. Planck Collaboration XVI (2014b) has more details about the CDM and nuisance parameters, whilst the dark energy extension parameters are defined in the text.

Parameter	Prior range	Prior type
$\Omega_{\rm b}h^2$	[0.019, 0.025]	Uniform
$\Omega_{\rm c}h^2$	[0.095, 0.145]	Uniform
$100\theta_{\rm MC}$	[1.03, 1.05]	Uniform
τ	[0.01, 0.4]	Uniform
$n_{\rm s}$	[0.885, 1.04]	Uniform
$\ln(10^{10}A_{\rm s})$	[2.5, 3.7]	Uniform
$A_{100}^{\rm PS}$	[0, 360]	Uniform
$A_{143}^{\rm PS}$	[0, 270]	Uniform
$A_{217}^{\rm PS}$	[0, 450]	Uniform
$A_{143}^{\rm CIB}$	[0, 20]	Uniform
$A_{217}^{\rm CIB}$	[0, 80]	Uniform
$A_{143}^{\rm tSZ}$	[0, 10]	Uniform
$r_{143\times217}^{\rm PS}$	[0, 1]	Uniform
$r_{143\times217}^{\rm CIB}$	[0, 1]	Uniform
$\gamma^{\rm CIB}$	[−2, 2]	Uniform
c_{100}	[0.98, 1.02]	Uniform
c_{217}	[0.95, 1.05]	Uniform
$\xi^{\rm tSZ-CIB}$	[0, 1]	Uniform
$A^{\rm kSZ}$	[0, 10]	Uniform
β_1^1	[−20, 20]	Uniform
$w(z_i)\vert_{i=1,\ldots,5}$	[−2, − 0.01]	Uniform
$z_i\vert_{i=2,\ldots,4}$	[0.01, 2.0]	Sorted-uniform
n	[Λ, w, t, 1, 2, 3]	Uniform
$\theta_{\rm uniform}$	[−2, − 0.01]	Uniform

With the new method there seems to be no way to estimate the errors on the PORs from a single run, and attaining these is best done via repeat simulation and the calculation of sample standard deviations from these. We therefore performed three repetitions each using $N_{\rm live} = 500$ with eff $= 0.01$ (the repeat runs) and the default 2014 July COSMOMC priors for the 20 CDM and nuisance parameters and the priors mentioned above for additional model parameters; an overview is shown in Table 2. Constant efficiency mode had to be used to attain feasible computing times, similarly the search parameters could not just be increased arbitrarily. With these MULTINEST search parameters and constant efficiency mode, it was found that the edges of the priors were not sampled effectively. The error is reproducible with a 20-dimensional Gaussian test likelihood with a covariance matrix given by Planck chains. To ensure this problem had no impact on our results, first we added a prior for an unconstrained parameter, the $\theta_{\rm uniform}$ parameter in Table 2, which should produce a flat posterior. Observing the edge effects problem on this parameter gives a clear indication of the severity of the problem, and allows us to reconsider parameter estimation conclusions if needed. Secondly, we tested for convergence of the marginalized posterior on n with respect to search parameter changes to ensure that our parameter estimation results were robust. We thus performed a single further run using MULTINEST with the search parameters $N_{\rm live} = 1000$, eff $= 0.005$ (full run) for which the edges of the prior were sampled effectively. Given the concerns about the accuracy of the MULTINEST evidence calculation for Planck data (due to nuisance parameters, high dimensionality, and the need for constant efficiency mode), the new method combined with the two robustness checks thus provides a valuable alternative way to obtain PORs.

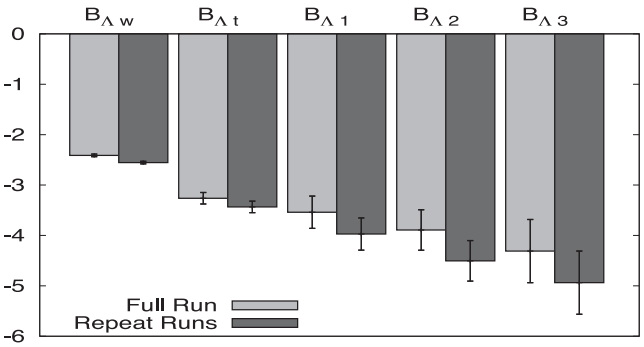


Figure 7. The PORs obtained from the new method comparing the five DE extension models to ΛCDM. The error bars on each histogram are the sample standard deviations of the three repeat runs. It is clear that the two sets of results agree very well, with discrepancies between them small compared both to the error bars and the absolute values used to draw conclusions based on Jeffreys guideline. This shows that the results are robust with respect to changes in MULTINEST search parameters, as required. Numerical results are given in Table 3.

Table 3. Summary of the Bayes factors from the four computations. The full run and repeat averages columns show results using the MULTINEST search parameters discussed in the text. For both columns, the errors are sample standard deviations of the three repeat trials. The results agree well within 1σ confidence intervals for all but the $\mathcal{B}_{\Lambda\, w}$, where a larger discrepancy occurs due to small error bars despite a small difference in log-units. The results show clearly that the new method implementation is robust to changes in MULTINEST parameters.

Bayes factor	Full run	Repeat averages
$\mathcal{B}_{\Lambda\, w}$	-2.41 ± 0.03	-2.55 ± 0.03
$\mathcal{B}_{\Lambda\, t}$	-3.26 ± 0.11	-3.43 ± 0.11
$\mathcal{B}_{\Lambda\, 1}$	-3.54 ± 0.32	-3.97 ± 0.32
$\mathcal{B}_{\Lambda\, 2}$	-3.89 ± 0.40	-4.50 ± 0.40
$\mathcal{B}_{\Lambda\, 3}$	-4.31 ± 0.63	-4.94 ± 0.63

5.2 Results

The POR results for the full run and the 3 repeat runs are shown in Fig. 7 and Table 3. The key points are first that the PORs are consistent with each other, demonstrating convergence of $\Pr(n|\mathcal{D}, \mathcal{M})$ with respect to MULTINEST search parameters, and secondly that the $w(z)$ investigation clearly favours ΛCDM.

The toy model showed that error bars on PORs will depend on how thoroughly the sampling explores the space. Note that the error bars used are the sample standard deviations from the PORs of the three repeat runs. The repeat run PORs are consistent with the full run and sufficiently tight to resolve differences to make conclusions based on Jeffreys guideline, suggesting that the space is well explored. This convergence on reruns, together with the convergence between different MULTINEST search parameters, suggests that the POR results are robust. Additionally, the edge effect problem previously mentioned was thoroughly checked for using an

unconstrained parameter $\theta_{\rm uniform}$. The posterior of $\theta_{\rm uniform}$ was close to flat for all runs. The edge effect problem presumably affects all parameters a small amount, as the strength of this effect is different between the different MULTINEST search parameter settings whilst the PORs are consistent, it suggests that the PORs are not significantly biased. From these four runs, we therefore conclude that we have accurate PORs and proceed to quote those of the full run combined with the errors from the three repeat runs as upper estimates for those of the full run (as repeats of a more well sampled run will produce tighter estimates, shown in the toy model when doubling $N_{\rm live}$).

From these PORs, it is clear that ΛCDM is the only favourable model. The decrease in PORs with an increase in the number of parameters to model DE suggests that further additions of parameters to model deviations from ΛCDM are penalized more strongly by the Occam's Razor principle than the gain in constraining power that they provide. One can estimate the Occam factor associated with adding an additional nodal amplitude parameter, using the analysis in (MacKay 2003, page 349), as $\sigma_{w|\mathcal{D}}/\sigma_w$, where $\sigma_{w|\mathcal{D}}$ is the width around the peak of a Laplace approximation inside the evidence integral and σ_w is the prior width. We estimating $\sigma_{w|\mathcal{D}}/\sigma_w$ for non-Gaussian parameters with a full width half max (FWHM) calculation of the 1D marginalized w-amplitude posterior. Doing this for the wCDM model's additional parameter yields a drop in the Bayes factor due to the approximated Occam factor of -2.63. The observed -2.41 ± 0.03 therefore suggests that the parameter is not improving the likelihood fit to the data significantly. Doing something similar for the three internal node model gives an Occam factor of -0.45 (using the average of the five amplitudes; assuming that an additional z-position parameter is unconstrained as there are no additional $w(z)$ features it would constrain). This is the anticipated decay in the POR when adding unnecessary nodes, and the Bayes factor drop from 2CDM to 3CDM at -0.42 suggests that three nodes already saturate the $w(z)$ space.

A clear and strong conclusion from this analysis is that there is considerably less evidence for deviations from ΛCDM in the Planck era data sets used here than in the *WMAP* era data sets used by Vázquez et al. (2012b), which is consistent with other results (Planck Collaboration XVI 2014b; Shafer & Huterer 2014). The next most favoured model is the next simplest one, wCDM, and at a PORs of -2.41 ± 0.03 it is almost significantly disfavoured according to the Jeffreys guideline. All other models are significantly disfavoured at between 3.3 and 4.3 log units. The constraints in the (w, z) plane for each of the model extensions beyond ΛCDM, shown in Fig. 8, do however indicate some deviations from $w = -1$. Typically the data seem to favour the phantom region, potentially more so at the ends of the considered redshift range and less so at redshift 0.4–0.7, where the data gives the tightest constraints. However, the $1\,\sigma$ and $2\,\sigma$ contours clearly indicate that these effects are not significant. At all z and for all models, $w = -1$ is comfortably within the peak of the $\Pr(w|z)$ distribution and more so in the regions where we have strong data constraints, suggesting that any deviations or apparent systematic patterns are dominated by a lack of data. The plane reconstructions also support the model selection conclusions that ΛCDM is significantly favoured over other models, as the constraints in the data do not deviate from $w = -1$ beyond even 1σ.

The correct Bayesian way to view the $w(z)$ plane reconstructions for all models considered is to sum over all the models whilst weighting by the Bayesian evidence, or equivalently PORs. This is exceptionally easy to implement with our new method, as a program like GETDIST (included with COSMOMC) can use the chains file produced by the new method to correctly weight all the models automatically whilst marginalizing out the parameter n. Fig. 9 shows this for the five DE extension models beyond ΛCDM. When plotting with ΛCDM, the plot is centred on $w = -1$, with 85 per cent of the peak confidence interval region contained in the $w = -1$ line, and thus a plot showing only the model extensions is more insightful. The plane reconstruction shows clearly the constraining power of the data at different redshifts as our knowledge of $w(z)$ moves from the prior on the left to the posterior on the right. The result is a tightly constrained function of $w(z)$ slightly below -1 for all redshifts, suggesting a small favouring of the phantom region at an insignificant level. Most importantly, ΛCDM is fully compatible, well within 1σ of the model extension results, as is expected given the Bayesian model selection analysis. This insignificant deviation away from $w = -1$ explains clearly why ΛCDM is so heavily favoured.

Of practical importance is the strength with which the nodal reconstruction identifies features, and especially that the reconstruction is data driven. Most of our data sets that can constrain $w(z)$ are in the redshift range $z \in [0.5, 0.8]$ and this is shown by where the reconstructions most tightly constrain the plane. This reconstruction technique is clearly of merit and in the future, with more powerful data sets, can hopefully act as a tool to identify features (if any) in $w(z)$. At present, the work here can only suggest that dark energy models with $w(z)$ close to -1 are needed. Finally, the posteriors of the CDM parameters are plotted in Fig. 10 for each of the six models tested. The posteriors of the DE extensions agree well with the ΛCDM values, as can be expected given that there is no significant deviation from $w = -1$.

6 CONCLUSIONS

We demonstrated a novel method for calculating PORs through a toy model application and then applied it to a cosmological model selection problem.

Our new method uses Bayesian parameter estimation on a parameter that switches between models, via a hyperlikelihood that wraps around the individual model likelihoods, to infer PORs (or Bayes factors if desired) without calculating evidences. It uses novel partitioning of the parameter space via the parameter n, and marginalization of posterior probabilities, to allow sampling of a variable length parameter space when moving between models, thus facilitating any models to be tested without restriction and without reversible jump Monte Carlo techniques. To use the method, one needs to have a parameter estimation algorithm capable of sampling from multimodal spaces and to decide which models one wants to test a priori.

The toy model demonstrated clearly that the method is valid and consistent with the existing method of calculating PORs by evaluating evidences. We conclude that the new method is not necessarily faster, despite avoiding evidence integrals, for two reasons. First, to get errors on the PORs it requires rerunning several times, whereas nested sampling algorithms such as MULTINEST and POLYCHORD can attain error estimates of evidences from a single run. Secondly, the parameter space needs to be explored comparably thoroughly in both methods, as shown by the increase in error bars on the PORs in the toy model when spending less computational time on the new method.

A peculiar feature of the new method in combination with nested sampling (which likely applies to other samplers too) is that computation time dedicated to a model is dependent on how strongly the model is favoured over others. Less favoured models become depopulated with live points as the nested sampling algorithm

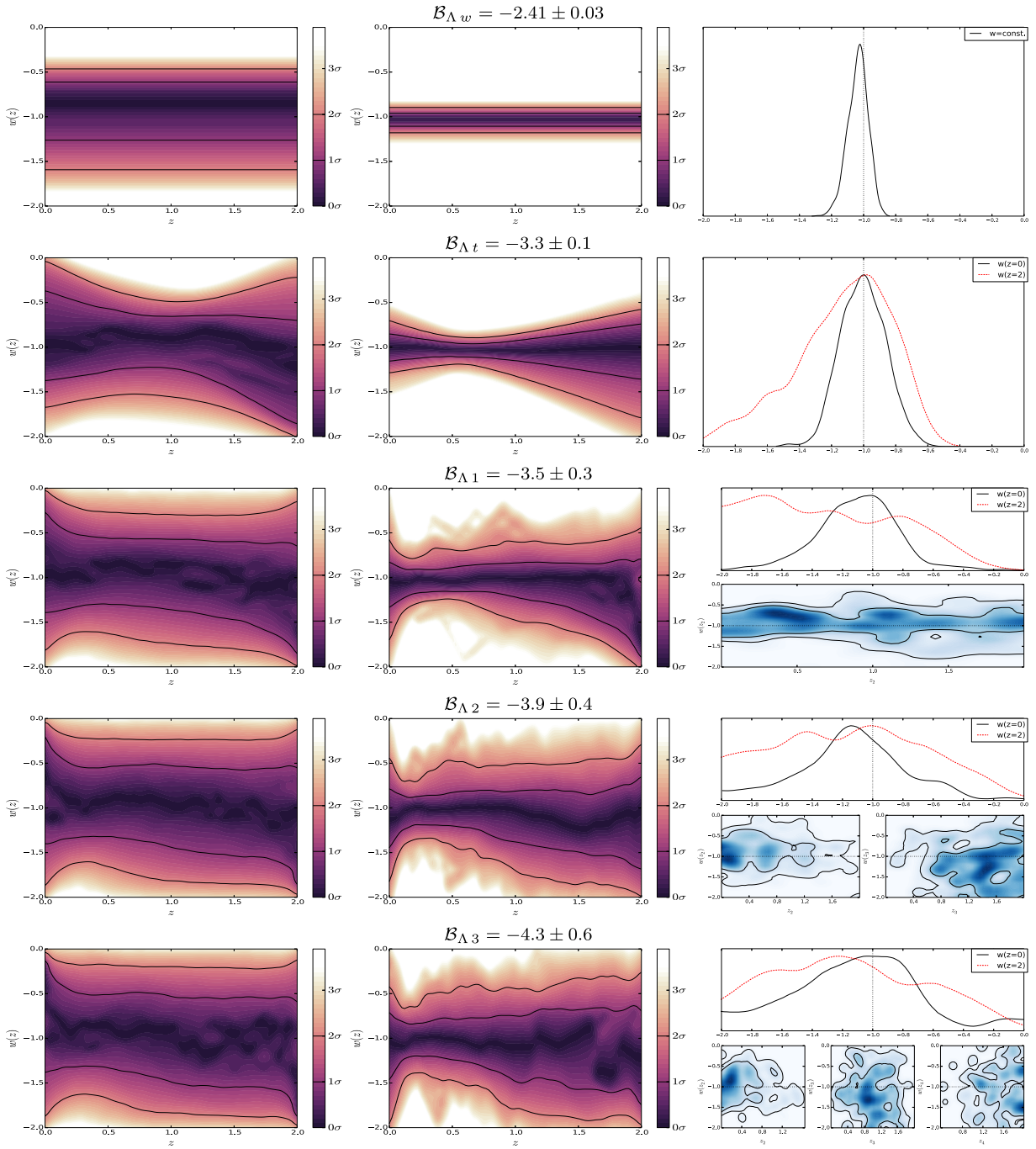


Figure 8. The $w(z)$ priors, $w(z)$ reconstructions and parameter constraints for each of the five model extensions beyond ΛCDM. The leftmost plot is the prior space on the function $w(z)$ as a result of our flat priors on amplitude and position parameters and the central plots show the posterior on $w(z)$ defining the data and model constraints on the $w(z)$-plane. These plots show the posterior probability $\Pr(w|z)$ similar to Fig. 5. Here, it is the probability of w as normalized in each slice of constant z, with colour scale in confidence interval values shown. The 1σ and 2σ confidence intervals are plotted as black lines. Note that the priors on $w(z)$ include implicit prior information from COSMOMC, and therefore are not flat. The posteriors show that the data constrains $w(z)$ strongly compared to our priors. Rightmost are the 1D and 2D marginalized posteriors of the additional model parameters. Plots were produced using GETDIST and with the cubehelix colour scheme by Green (2011) for linearity in grey-scale.

removes lowest likelihood points. As a result, we observed that less favoured models typically had less accurate POR calculations, which helps to reduce computing time, but still in such a way that they were always identifiable as less favoured. The reduction in computing time can be substantial, especially in applications where there are a number of computationally expensive models with low PORs.

The toy models illuminated precautionary measures that best be adhered to by users. As with all Bayesian parameter estimation, robustness of posterior probabilities to changes in algorithm-specific tuning parameters needs to be tested for and in the case of the new method, where a posterior is used to infer evidence ratios, it is especially important to check this. It is best to test that the PORs obtained from the posterior on n are consistent on repetitions of the

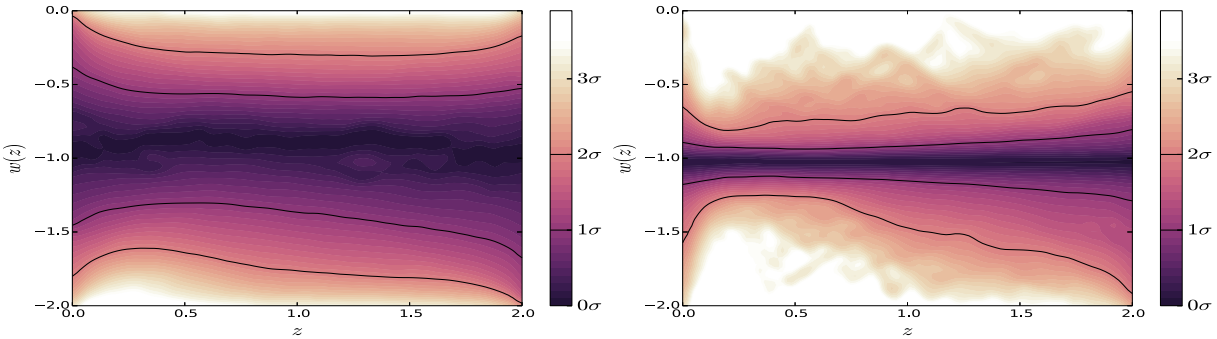


Figure 9. Summarising the DE model extension results for the constraints on the $w(z)$ plane. The five extension models, excluding ΛCDM, are weighted by their evidences to give a model averaged plane reconstruction (Parkinson & Liddle 2013; Planck Collaboration XX 2015), and plotted as in Fig. 8. When including ΛCDM, approximately 85 per cent of the central confidence interval region is contained in the line $w = -1$ due to the strength with which ΛCDM is favoured by the PORs, almost 2σ. The two plots show the prior space (left) contracting down to the POR averaged $w(z)$ plane reconstruction (right), as discussed in the text. It is clear that ΛCDM is well within the favourable region, with the 1σ contours easily containing $w = -1$.

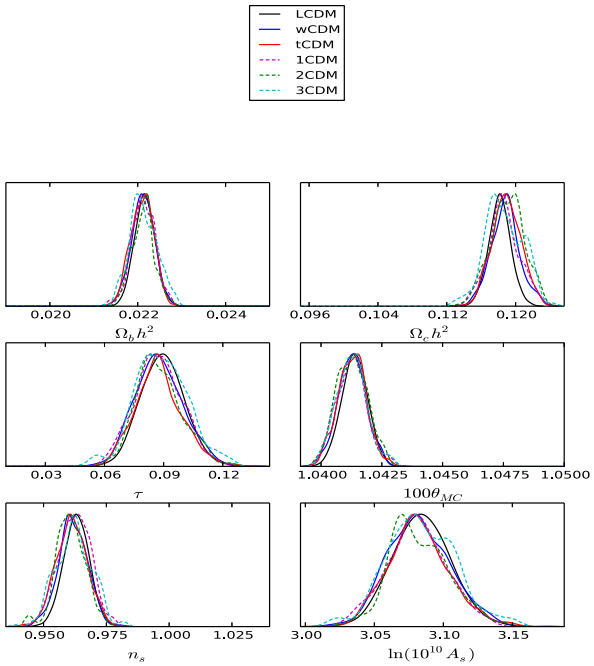


Figure 10. The CDM parameter 1D marginalized posteriors for each of the six models tested. As MULTINEST converges to the peak likelihood regions, the data points output to the chains file are more sparse for some of the models. Typically, ΛCDM had eight times more points than wCDM with which to accurately reconstruct these posteriors. The lower POR models had less still and this leads to a lower quality reconstruction for the less favoured models. Nevertheless, it is clear that the models agree well and there are no significant deviations from the ΛCDM values of the CDM parameters, as can be expected given the only slight deviation from $w = -1$ in each model.

algorithm and also that the error bars attained from repetitions are sufficiently small if needing to make judgments based on Jeffreys guideline. The toy model also highlighted the strength of the nodal reconstruction in identifying features in $y(x)$ plane reconstruction problems. We conclude that it is a useful tool for analysing the complexity supported by the data and add to the volume of literature using it (Vázquez et al. 2012a,b; Aslanyan et al. 2014; Planck Collaboration XX 2015).

Thereafter, taking the above considerations into account, the new method was used to attain PORs in a cosmological context where direct evaluation of evidences can be computationally demanding and problematic. We applied the nodal reconstruction technique to reconstruct the dark energy redshift-dependent equation of state parameter $w(z)$, analysing the dynamic behaviour supported by modern data sets in a search for deviations from the ΛCDM model ($w = -1$). This was principally an update on a paper using *WMAP* era data by Vázquez et al. (2012b). We concluded that ΛCDM is significantly favoured above any nodal reconstruction applied. Additionally, the model allowing w to vary as a constant is almost significantly disfavoured at -2.41 ± 0.03 log-units of the POR with respect to ΛCDM. We conclude that additional parameters are systematically disfavoured: increasing the complexity of the $w(z)$ reconstruction decreases PORs with respect to ΛCDM. The Occam's Razor effect penalizes additional parameters when using PORs to do model selection and, as ΛCDM is an excellent fit to current cosmological data, the addition of parameters to extended beyond ΛCDM adds less to the constraining power of the models than the Occam's factor penalizes.

The robustness of the results and methods were confirmed in several ways. Fig. 10 shows that the CDM parameters of each of the dark energy extension models agree well with the ΛCDM values, as is expected given that all models agree well with $w = -1$. Further, a potential problem in sampling the edges of priors in high dimensions was identified with MULTINEST when using constant efficiency mode, but through tracking an unconstrained parameter $\theta_{\rm uniform}$, it was shown to be insignificant given the final search parameters used. General robustness of the new method was confirmed too by repeating the calculation of $\Pr(n|\mathcal{D}, \mathcal{M})$ with different search parameters and showing that the value of $\Pr(n|\mathcal{D}, \mathcal{M})$ had converged with respect to algorithm tuning parameter.

Finally, the cosmological application demonstrated the strength of the new method, attaining PORs without needing evidence calculations and effectively dealing with parameter spaces of varying length. Errors on the PORs were attained through repeat runs with a faster sampling parameter setup which doubled to confirm that the PORs were converged and accurate. As such a robustness check is important for any parameter estimation or model selection problem, where an algorithm uses tuning parameters for the sampling, this approach should come at little extra cost in practice.

ACKNOWLEDGEMENTS

The authors thank Farhan Feroz for many useful discussions and insights, and also Ewan Cameron, Kirill Tchernyshyov and the journal referee for their very insightful additions. This work was performed using the Darwin Supercomputer of the University of Cambridge High Performance Computing Service (http://www.hpc.cam.ac.uk/), provided by Dell Inc. using Strategic Research Infrastructure Funding from the Higher Education Funding Council for England and funding from the Science and Technology Facilities Council. Parts of this work were undertaken on the COSMOS Shared Memory system at DAMTP, University of Cambridge operated on behalf of the STFC DiRAC HPC Facility, this equipment is funded by BIS National E-infrastructure capital grant ST/J005673/1 and STFC grants ST/H008586/1, ST/K00333X/1. SH and WH thank STFC for financial support.

REFERENCES

Akaike H., 1974, IEEE Trans. Autom. Control, 19, 716
Anderson L. et al., 2014, MNRAS, 441, 24
Aslanyan G., Price L. C., Abazajian K. N., Easther R., 2014, J. Cosmol. Astropart. Phys., 8, 52
Bayes M., Price M., 1763, Philosophical Trans. R. Soc. London, 53, 370
Bennett C. L. et al., 2013, ApJS, 208, 20
Brewer B. J., Donovan C. P., 2015, MNRAS, 448, 3206
Brewer B. J., Pártay L. B., Csányi G., 2011, Stat. Comput., 21, 649
Caldwell R., 2002, Phys. Lett. B, 545, 23
Caldwell R., Dave R., Steinhardt P., 1998, Phys. Rev. Lett., 80, 1582
Carlin B. P., Chib S., 1995, J. R. Stat. Soc. Ser. B, 57, 473
Clyde M. A., Berger J. O., Bullard F., Ford E. B., Jefferys W. H., Luo R., Paulo R., Loredo T., 2007, in Babu G. J., Feigelson E. D., eds, ASP Conf. Ser. Vol. 371, Statical Challenges in Modern Astronomy IV. Astron. Soc. Pac., San Francisco, p. 224
Fang W., Hu W., Lewis A., 2008, Phys. Rev. D, 78, 087303
Feroz F., Hobson M. P., 2008, MNRAS, 384, 449
Feroz F., Skilling J., 2013, in von Toussaint U., ed., AIP Conf. Proc. Vol. 1553, Bayesian Inference and Maximum Entropy Methods in Science and Engineering. Am. Inst. Phys., New York, p. 106
Feroz F., Hobson M. P., Bridges M., 2009, MNRAS, 398, 1601
Feroz F., Hobson M. P., Cameron E., Pettitt A. N., 2013, preprint (arXiv:1306.2144)
Gelman A., Meng X.-L., 1998, Stat. Sci., 13, 163
Goyder R., Lasenby A. N., 2004, MNRAS, 353, 338
Green P. J., 1995, Biometrica, 82, 711
Green D. A., 2011, Bull. Astron. Soc. India, 39, 289
Handley W. J., Hobson M. P., Lasenby A. N., 2015, MNRAS, 450, L61
Hobson M. P., McLachlan C., 2003, MNRAS, 338, 765
Howlett C., Lewis A., Hall A., Challinor A., 2012, J. Cosmol. Astropart. Phys., 2012, 27
Jeffreys S. H., 1961, The Theory of Probability. Oxford Univ. Press, Oxford
Lewis A., Bridle S., 2002, Phys. Rev. D, 66, 103511
Lewis A., Challinor A., Lasenby A., 2000, ApJ, 538, 473
Liddle A., Mukherjee P., Parkinson D., 2006, Astron. & Geophys., 47, 4.30
Lodewyckx T., Kim W., Lee M. D., Tuerlinckx F., Kuppens P., Wagenmakers E.-J., 2011, J. Math. Psychol., 55, 331
MacKay D. J. C., 2003, Information Theory, Inference and Learning Algorithms. Cambridge Univ. Press, Cambridge
Parkinson D., Liddle A. R., 2013, Stat. Anal. Data Min., 6, 3
Planck Collaboration XV, 2014a, A&A, 571, A15
Planck Collaboration XVI, 2014b, A&A, 571, A16
Planck Collaboration XVII, 2014c, A&A, 571, A17
Planck Collaboration XX, 2015, A&A, preprint (arXiv:1502.02114)
Ratra B., Peebles P., 1988, Phys. Rev. D, 37, 3406
Sahni V., 2005, in Papantonopoulos E., ed., Lecture Notes in Physics, Vol. 653, The Physics of the Early Universe. Springer, Berlin, p. 141
Schwarz G., 1978, Ann. Stat., 6, 461
Shafer D. L., Huterer D., 2014, Phys. Rev. D, 89, 063510
Sisson S. A., 2005, J. Am. Stat. Assoc., 100, 1077
Sivia D. S., Skilling J., 2006, Data Analysis: a Bayesian Tutorial. Oxford Univ. Press, Oxford
Skilling J., 2004, AIP Conf. Ser., 119, 1211
Skilling J., 2006, Bayesian Anal., 1, 833
Suzuki N. et al., 2012, ApJ, 746, 85
Tierney L., B. K. J., 1986, J. Am. Stat. Assoc., 81, 82
Trotta R., 2008, Contemp. Phys., 49, 71
Tsujikawa S., 2013, Class. Quantum Grav., 30, 214003
Vázquez J. A., Bridges M., Hobson M., Lasenby A., 2012a, J. Cosmol. Astropart. Phys., 2012, 6
Vázquez J. A., Bridges M., Hobson M., Lasenby A., 2012b, J. Cosmol. Astropart. Phys., 2012, 20
Verdinelli I., Wasserman L., 1995, J. Am. Stat. Assoc., 90, 614
Zhang H., 2009, preprint (arXiv:0909.3013)

APPENDIX A: LINE FITTING LIKELIHOOD

We aim to fit a parametric function $y = f(x)$ to a set of $j_{\max}$ data points $\{x_j, y_j\}$, where we have some knowledge of the errors on these measurements $\{\sigma_{x_j}, \sigma_{y_j}\}$ ($\{j = 1, \ldots, j_{\max}\}$). In order to fit the function, one needs to calculate the likelihood of observing the data $\{x_j, y_j\}$, given the function f, the observed errors and any additional assumptions we must make I:

$$\Pr(\{x_j, y_j\}|\{\sigma_{x_j}, \sigma_{y_j}\}, f, I). \tag{A1}$$

To model the 'error bars', we assume that each of the data points (x_j, y_j) is drawn from a separable Gaussian distribution with covariance $\mathrm{diag}(\sigma_{x_j}^2, \sigma_{y_j}^2)$. The distribution will be centred about some true value (X_j, Y_j), where these values are unknown and will need to be marginalized over as nuisance parameters in the final calculation. If each of these distributions are independent from each other, we arrive at the likelihood:

$$\Pr(\{x_j, y_j\}|\{X_j, Y_j\}, \{\sigma_{x_j}, \sigma_{y_j}\}) = \prod_{j=1}^{j_{\max}} \frac{1}{2\pi\sigma_{x_j}\sigma_{y_j}} \exp\left[-\frac{\left(x_j - X_j\right)^2}{2\sigma_{x_j}^2} - \frac{\left(y_j - Y_j\right)^2}{2\sigma_{y_j}^2}\right]. \tag{A2}$$

To marginalize out the nuisance parameters, we place our prior assumptions on them. We shall assume that the true X_j values are drawn uniformly in some range $X_- < X_j < X_+$, and we shall assume that the true Y_j obey the functional relationship: $Y_j = f(X_j)$. Given this, the probability distribution is

$$\Pr(\{X_j, Y_j\}|f, X_-, X_+) = \begin{cases} \frac{1}{X_+ - X_-} \prod_{j=1}^{j_{\max}} \delta\left[Y_j - f(X_j)\right] : X_- < X_j < X_+ \\ 0 : \text{otherwise}, \end{cases} \tag{A3}$$

where δ is the Dirac δ-function. Multiplying equations (A2) and (A3) together and marginalizing out $\{X_j, Y_j\}$ by integrating yields the likelihood:

$$\Pr(\{x_j, y_j\}|\{\sigma_{x_j}, \sigma_{y_j}\}, f, X_-, X_+) = \prod_{j=1}^{j_{\max}} \int_{X_-}^{X_+} dX_j \, \frac{\exp\left[-\frac{(x_j - X_j)^2}{2\sigma_{x_j}^2} - \frac{(y_j - f(X_j))^2}{2\sigma_{y_j}^2}\right]}{2\pi\sigma_{x_j}\sigma_{y_j}(X_+ - X_-)}. \tag{A4}$$

This procedure may be straightforwardly extended to consider correlated error bars where the covariance matrix of equation (A2) is no longer diagonal. One may also adjust equation (A3)

if some additional knowledge is known about the independent variables X_j. For further details, the reader is referred to Sivia & Skilling (2006).

APPENDIX B: EFFICIENT COMPUTING OF BAYES FACTORS

Using the data points in Fig. B1 to test the vanilla and Post(n) methods, we demonstrate that our new method may outperform the evidences approach in a systematic fashion that makes the approach desirable for common astrophysical and cosmological problems.

Running the nodal reconstruction technique with models of 1 internal node up to 13 internal nodes (3–15 total nodes), we obtain Bayes factors and timing results shown in Fig. B2. The timing data shows the number of posterior points, and thus likelihood calculations up to a factor of the PolyChord efficiency, that each method makes for each of the nodal reconstruction models (shaded plots), alongside the cumulative number of likelihood calculations of these models (line plots). Using the vanilla method, completing the evidence calculation for each model means that adding increasingly complex models is increasingly computationally expensive. In the Post(n) method, however, the model space is rapidly traversed from lower likelihood regions to higher likelihood regions, so that computationally expensive models with low likelihoods (or more correctly, with lower Bayes factors compared to other models in the space) are explored rapidly by the nested sampling algorithm. This is clearly identified by the fact that the Bayes factors and the number of likelihood calculations peak at the same model (four internal nodes) and tail off similarly for models on either side of this.

It is worth noting, however, that the Post(n) method performs more likelihood calculations for the most probable models, because the additional overhead of setting up the other parameters and populating their dimensions with live points (because throughout we use that $N_{\rm live} \propto N_{\rm dim}$) means that the algorithm progresses more slowly.

Astrophysical and cosmological problems where a number of models of increasing complexity are explored may therefore benefit from using this method. It is not guaranteed, however, as with the vanilla case one may have identified a drop off in the Bayes factors beyond $n = 8$ and stopped testing the more complex models thereafter. None the less, the Post(n) method could provide an efficient means of verifying the drop off (for example one might run the above with $\pi(n) = [4, 9, 10, 11, 12, 13]$ as a fast means of verifying the shape). Any gains in performance must be considered against the need for repetition of the algorithm to obtain an estimate of the error on the Bayes factors.

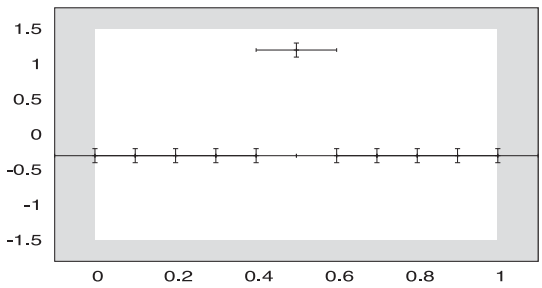


Figure B1. A set of 11 data points defining a *spike* in the x–y plane. We test this data set with models of 1 internal node up to 13 internal nodes (3–15 total nodes).

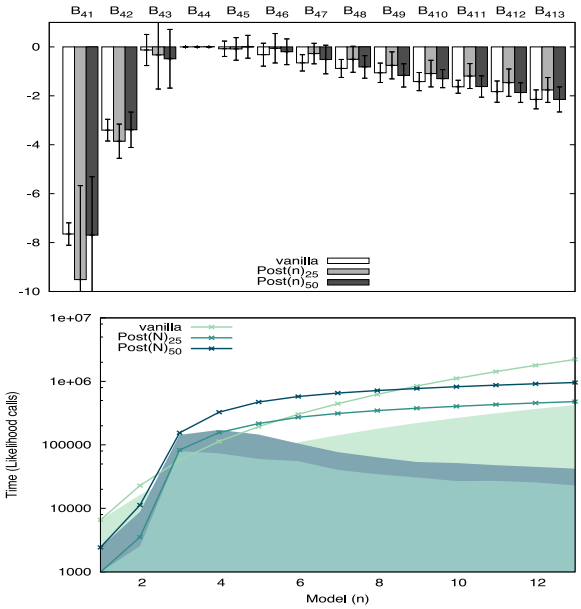


Figure B2. Bayes factors with respect to the most probable model (top) and timing data (bottom) for the vanilla method and the Post(n) method using $25N_{\rm dim}$ and $50N_{\rm dim}$ number of live points. Note that the large error bars on the data set in Fig. B1 allow models that underfit with less than three internal nodes (one at each vertex of the spike signal) to be probable. The timing data is measured by the number of likelihood calculations the algorithm makes. The shaded regions show the time taken on each nodal-reconstruction model for the vanilla (lightest colour plotted), Post$(n)_{25}$, and Post$(n)_{50}$ (darkest colour plotted) methods. Observe that the shapes of the Post(n) method timing data coincides with those of the Bayes factors, as explained in the text, and thus outperforms the vanilla method in obtaining Bayes factors accurately.

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