DISCLAIMER: COVID-19

The Covid-19 pandemic had a mixed effect on the following project. As soon as action was taken by the university, all international students were asked to leave. Different countries have had different policies in place. Unfortunately, because of the policies in Armenia, I had to spend a week in quarantine outside of my house, and without functional access to the CSD3 cluster. However, as the project is by nature theoretical, with a large portion of computational data, this did not preclude productivity, rather shifted priority away from cosmological simulations. The project lab-book was not affected either, it was kept and updated digitally, so COVID-19 had no bearing on the amount of work needed to maintain it (if anything it put everyone who maintained the lab-book digitally at an advantage).

Upon arriving at home, social distancing regulation was enforced via martial law. We are not (as of writing) permitted to occupy the same space as relatives, which meant that I was to live alone away from my spouse and parents. I could however ask for some of my personal belongings be transferred, which included a more convenient workstation, which also permitted more computationally expensive experiments to be conducted.

Unfortunately, resources on the CSD3 cluster became increasingly scarce. Firstly, the queue times lead to significant setbacks in the cosmological simulations. Additionally, because of increased network load in all countries, establishing and maintaining a stable connection to CSD3 became increasingly difficult. It did not lead to any results being cut, but only because our project lead to a significant optimisation of Cosmological data analysis. Had the speedup been less than by a factor of 20, we would not be able to obtain the final two figures.

Overall, Covid-19 placed obstacles, which were not insurmountable.

MNRAS **000**, 000–000 (0000) Preprint 12 June 2020 Compiled using MNRAS LATEX style file v3.0

Accelerated nested sampling in the context of cosmological parameter estimation

Examination ID: 8275R

12 June 2020

ABSTRACT

By extending previous work, we have found a method of using intuitive proposals with nested sampling to improve performance, accuracy and precision, called consistent posterior repartitioning. We have also developed a prescription called stochastic isometric mixture, of combining several intuitive proposals into a single model. This prescription allows nested sampling to make use of the most representative of the proposal priors. As a consequence, the sampling is hardened against prior imprinting, while also retaining most of its performance and accuracy. We demonstrate this by comparing full cosmological parameter estimations performed with the inference package `Cobaya` utilising the module `CLASS`, with and without our modifications. We demonstrate a run-time reduction in CSD3 Skylake compute-hours by a factor of 20, accompanied with increased precision. When precision normalised, the predicted performance uplift is by three orders of magnitude. The real-world performance uplift is such that cosmological inference was performed with a net precision gain on a personal computer within two days, provided a well-tuned proposal distribution. The findings are systematised, such that inference techniques that are similar to nested sampling in benefiting from our methods, can be identified. The scope of the findings suggests multiple improvements and an evolution of nested sampling.

Key words: Bayesian inference — automated posterior repartitioning — nested sampling — cosmology: miscellaneous — methods: statistical — methods: data analysis

1 INTRODUCTION

The standard model of the universe and its evolution in modern cosmology is the ΛCDM model (**?**), so named after the main components of the universe: the cosmological constant Λ and cold dark matter. It has six major independent[1] parameters: the physical baryon density $\Omega_b h^2$; the physical (cold) dark matter density $\Omega_c h^2$; the angular parameter $100\theta_s$; re-ionisation optical depth τ_{reio}; power spectrum slope n_s and amplitude $\ln(10^{10} A_s)$ (**?**).

The task of the present study is to develop better tools for evaluating the agreement of our observations from the Planck mission with ΛCDM, estimating the parameters in the process. In the language of Bayesian statistics[2], our goal is efficient Bayesian inference.

While said inference can be executed analytically in principle, it is often intractable. Multiple numerical algorithms exist to perform Bayesian inference: Metropolis-Hastings (**?**) in conjunction with the Gibbs sampler (**?**); Hybrid (Hamiltonian) Monte Carlo (**??**), and nested sampling (**?**). Most inference methods can benefit from proposals, so much so that these proposals are often provided with the Cosmological inference packages (**?**). Nested sampling is the exception, because it does not take proposals as separate input, and using them as priors may adversely affect the results. We have found a prescription for incorporating proposals safely: reducing the run-time, increasing the precision, while assuaging if not eliminating the aforementioned adverse effects.

We achieve this by extending *automatic power posterior re-partitioning* (**?**). The issues arising from improper usage of proposals as priors, are a variant of the problem of *unrepresentative priors* that we shall discuss in **??**. Hence, we find that posterior re-partitioning can be used to mitigate prior imprints. We found multiple extensions of posterior re-partitioning, with different trade-offs and potential. From

Table 1. A non-exhaustive list of major implementations of nested sampling.

Name	Publication
`MultiNest`	**?**
`PolyChord`	**?**
`nestle`	**?**
`dyNesty`	**?**

[1] there can be other equivalent parameter sextuplets.
[2] See **?** for comparison to frequentist statistics.

these extensions we have identified a combination that can provide significantly better results: performance, accuracy, precision, convenience, stability and imprint mitigation.

In the following section, we shall provide a brief primer on Bayesian inference and nested sampling, followed by an exploration of work by **?**. All work is our own from the third section onward, which includes the mathematical framework of consistent partitions, a few examples[3], along with descriptions of the underlying mechanism. We dedicate the final sections to practical demonstrations and a few suggested applications of our work.

2 THEORETICAL BACKGROUND

In this section we primarily focus on previous work, outlining the key elements of Bayesian inference (**?**) and automatic posterior repartitioning.

2.1 Bayesian inference

Hypothesis testing in Bayesian statistics requires said hypothesis to be formulated in terms of conditional probabilities, organising information in the following way.

A model $\mathcal{M}$ of a physical process, is parameterised by $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_n)$. New empirical observations of said process are encapsulated in the ***dataset*** $\mathfrak{D}$. The ***likelihood*** $\mathcal{L}$ of the parameters $\boldsymbol{\theta}$ is the probability of observing $\mathfrak{D}$, conditional on the configuration $\boldsymbol{\theta}$ and the model $\mathcal{M}$. The prior $\pi(\boldsymbol{\theta})$ is the probability of $\boldsymbol{\theta}$ assuming $\mathcal{M}$. It can be obtained from both previous datasets as well as constraints inherent to the model. The posterior is a probability of $\boldsymbol{\theta}$ that is conditional on $\mathcal{M}$ and the dataset $\mathfrak{D}$. The locus of all $\boldsymbol{\theta}$ for which the prior is both defined and non-zero defines the ***prior space*** Ψ. Finally, the ***evidence*** is the probability of the data $\mathfrak{D}$ assuming the model.

The interactions of probabilities of **??** is governed by **?**'s theorem:

$$\mathcal{L}(\boldsymbol{\theta}) \times \pi(\boldsymbol{\theta}) = \mathcal{Z} \times \mathcal{P}(\boldsymbol{\theta}). \tag{1}$$

Bayesian inference is the process of reconciling the model $\mathcal{M}$ represented in $\mathcal{L}$ and π, with observations $\mathfrak{D}$ represented in $\mathcal{L}$. A numerical algorithm that obtains $\mathcal{Z}$ and $\mathcal{P}$ from π and $\mathcal{L}$, is called a ***sampling algorithm*** or ***sampler***.

The convenient representation of π and $\mathcal{L}$ depends on the particulars of the sampler. For ***nested sampling*** (e.g. `PolyChord`, `MultiNest`) we delineate them indirectly: with the logarithm of the likelihood probability-density function $\ln \mathcal{L}(\boldsymbol{\theta})$, and ***prior quantile*** $C\{\pi\}(\boldsymbol{\theta})$. The latter, can be thought of as a coordinate transformation $C : \boldsymbol{u} \mapsto \boldsymbol{\theta}$ that maps a uniform distribution of $\boldsymbol{u}$ in a unit hypercube to $\pi(\boldsymbol{\theta})$ in Ψ. It is often obtained by inverting the cumulative distribution function of the prior.

For **?**' theorem to hold, the domains of all probability density functions need to be the same. Let $D(f)$ denote the domain of the probability density function f, i.e. where f is both defined and **non-zero**. Hence

$$D\{\pi\} \cap D\{\mathcal{L}\} = D\{\mathcal{P}\} \subset \Psi, \tag{2}$$

[3] We present only a small subset of consistent partitions designed during the project.

Table 2. Definitions of main quantities in Bayesian inference.

Term	Symbol	Definition
Prior	$\pi(\boldsymbol{\theta})$	$P(\boldsymbol{\theta}\|\mathcal{M})$
Likelihood	$\mathcal{L}(\boldsymbol{\theta})$	$P(\mathfrak{D}\|\boldsymbol{\theta} \cap \mathcal{M})$
Posterior	$\mathcal{P}(\boldsymbol{\theta})$	$P(\boldsymbol{\theta}\|\mathfrak{D} \cap \mathcal{M})$
Evidence	$\mathcal{Z}$	$P(\mathfrak{D}\|\mathcal{M})$

meaning the inference is possible only on a subset of the domain of prior and likelihood.

For each choice of $\mathcal{L}$ and π, there is a unique choice of $\mathcal{Z}$ and $\mathcal{P}$; equivalently they represent the same unique model $\mathcal{M}$, or partition it consistently. This correspondence is *surjective*, but not *injective*: many choices of $\mathcal{L}(\boldsymbol{\theta})$ and $\pi(\boldsymbol{\theta})$ may correspond to the same $\mathcal{P}(\boldsymbol{\theta})$ and $\mathcal{Z}$ (**?**). This remark is the cornerstone of our optimisation.

2.2 Nested Sampling

By noting that $\mathcal{P}$ is a probability, hence normalised, from **??** we obtain

$$\mathcal{Z} = \int_{\Psi} \mathcal{L}(\boldsymbol{\theta}) \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}. \tag{3}$$

Thus, **?**'s theorem reduces parameter estimation — obtaining $\mathcal{P}$ from π and $\mathcal{L}$, to integration (**?**). The naïve approach to obtaining $\mathcal{Z}$: e.g. uniformly rasterise Ψ, is intractable for hypotheses with $O(30)$ parameters (**?**). Integration is usually performed using Monte Carlo techniques, such as nested sampling.

The following is a short description of nested sampling (**?**). We begin, by picking n_{live} ***live points*** at random in Ψ. During each subsequent iteration, the point with the lowest likelihood is declared ***dead***, and another live point $\boldsymbol{\theta} \in \Psi$ is taken with a higher likelihood, based on the prior π and an implementation-dependent principle. Live points are thus gradually moved into regions of high likelihood. By tracking their locations and likelihoods, from a statistical argument we can approximate $\mathcal{Z}$ and its error for each iteration, and by **??**, $\mathcal{P}(\boldsymbol{\theta})$. We continue until a pre-determined fraction of the evidence associated to Ψ remains unaccounted for.

Not all parameter inference methods require obtaining $\mathcal{Z}$. Some methods, such as Hamiltonian Monte-Carlo (**?**), allow obtaining a normalised $\mathcal{P}$ directly. For such approaches, any consistent specification of π and $\mathcal{L}$ will lead to identically the same posterior, barring numerical errors. This is also true of methods that evaluate $\mathcal{Z}$ exactly. However, nested sampling allows uncertainty in $\mathcal{Z}$, which is controlled by π and $\mathcal{L}$. Thus, nested sampling, unlike, e.g. Metropolis-Hastings (**?**) is sensitive to the concrete definitions of prior and likelihood. While many choices of π and $\mathcal{L}$ correspond to the same $\mathcal{P}$ and $\mathcal{Z}$, the errors and nested sampling's time complexity are different for different specifications of π(**?**).

A probability density function $f(\boldsymbol{\theta})$ is said to be more ***informative*** than $g(\boldsymbol{\theta})$ if:

$$\mathcal{D}\{f, g\} > \mathcal{D}\{g, f\}. \tag{4}$$

This also highlights, that Kullback-Leilber divergence is not a metric on the space of distributions. However, being asymmetric lends itself well to considerations where such an asymmetry is natural: e.g. priors are not equivalent to posteriors, one comes after the other, and so $\mathcal{D}$ can be used to

quantify the "surprise" information obtained during inference.

The time complexity T of nested sampling satisfies

$$T \propto n_{\text{live}} \, \langle \mathcal{T}\{\mathcal{L}(\boldsymbol{\theta})\} \rangle \, \langle \mathcal{N}\{\mathcal{L}(\boldsymbol{\theta})\} \rangle, \tag{5}$$

where $\mathcal{T}\{f(\boldsymbol{\theta})\}$ represents time complexity of evaluating $f(\boldsymbol{\theta})$ and $\mathcal{N}\{f(\boldsymbol{\theta})\}$ — the quantity of such evaluations. Reducing n_{live} reduces the resolution of nested sampling, while $\mathcal{T}\{\mathcal{L}(\boldsymbol{\theta})\}$ is model-dependent. We can, however, reduce the number of likelihood evaluations, by providing a more informative prior. However, there is an associated risk, which precludes use of proposals as informative priors.

An important quantity for measuring the correctness of the obtained posterior is the ***Kullback-Leilber divergence*** $\mathcal{D}$ (**?**). For probability distributions $f(\boldsymbol{\theta})$ and $g(\boldsymbol{\theta})$, it is defined as:

$$\mathcal{D}\{f, g\} = \int_{\Psi} f(\boldsymbol{\theta}) \ln \frac{f(\boldsymbol{\theta})}{g(\boldsymbol{\theta})} d\boldsymbol{\theta}. \tag{6}$$

It is a pre-metric on the space of probability distributions: it is nil if and only if $f(\boldsymbol{\theta}) = g(\boldsymbol{\theta})$, but is not symmetric. This is convenient for defining a representation hierarchy. The statement: f represents g better than h is equivalent to

$$\mathcal{D}\{f, g\} < \mathcal{D}\{h, g\}. \tag{7}$$

Specifically, distribution h is said to be unrepresentative of g if a uniform distribution f represents g better than h does.

The representation of $\mathcal{P}$ and π is crucial for nested sampling's correctness and performance. For example, assuming the same likelihood, if π_0 and π_1 are equally informative, but π_0 is more representative of $\mathcal{P}$, then the inference with π_0 will terminate more quickly than with π_1, (more accurate, also).

Similarly, if π_1 is more informative than π_2, but equally as representative, nested sampling will terminate with π_1 faster than with π_2, and the result will be more precise. In detail, if $\pi_1(\boldsymbol{\theta})$ is more similar to $\mathcal{P}(\boldsymbol{\theta})$, then points drawn with PDF $\pi_1(\boldsymbol{\theta})$ are more likely to lie in $\boldsymbol{\theta}$ regions of high $\mathcal{P}(\boldsymbol{\theta})$, leading to fewer iterations.

Posteriors $\mathcal{P}_1$ and $\mathcal{P}_2$ obtained with the priors π_1 and π_2 are different, because of **??**. In fact, the posterior $\mathcal{P}_1$ will be more informative than $\mathcal{P}_2$, and more similar to π_1. This we call ***prior imprinting***.

Imprinting is desirable if the informative prior π_1 is the result of multiple inferences over multiple datasets. However, even in such a case imprinting limits the information obtainable from $\mathfrak{D}$. The risk of getting no usable data from the inference is sufficient to prefer uniform priors even when more information is available. This is exasperated in case of proposals, which is why they are almost never used with nested sampling. The issue is that the algorithm has no room to consult the proposal distributions outside of the prior. Using a prior taken out of "thin air", with nested sampling is recipe for disaster. However, in the next section we shall discuss how one can mitigate these issues.

2.3 Power posterior repartitioning

From this section onward we shall adopt the following notation. π and $\mathcal{L}$ with similar annotations (index, diactrics),

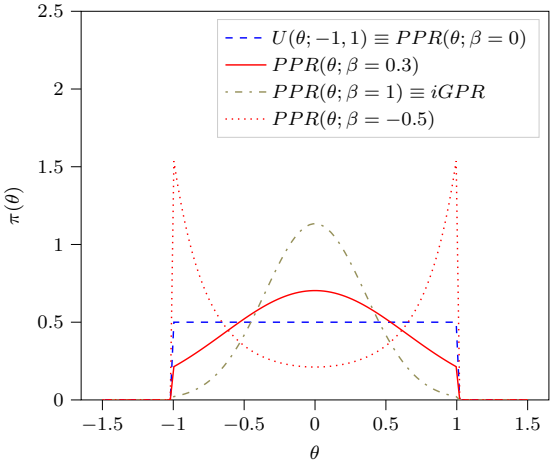


Figure 1. Demonstration of $\hat{\pi}(\theta; \beta)$ for different values of β in one dimension. Note that we've assumed that the original $\pi(\boldsymbol{\theta})$ distribution is a truncated Gaussian, i.e. zero outside the region $(-1, 1)$, which manifests as changes in curvature at the boundaries. The area under curves for different β is normalised to unity as in **??**.

belong to the same specification of the model. Models using the uniform prior are special, in that they obtain the most accurate posterior and evidence. They are represented with an over-bar (the plot of a uniform prior in 1D is a horizontal line). Hats delineate the consistent partitions, that incorporate the proposal (the hat represents the peak(s) often present in informative proposals).

We are working under the assumption that $\pi(\boldsymbol{\theta})$ is an informative, unrepresentative prior. We want to obtain correct posterior $\bar{\mathcal{P}}$ but without using a uniform, universally representative reference prior $\bar{\pi}$, because it is often the least informative. To avoid loss of precision and mitigate prior imprinting, **?** have proposed introducing the parameter β to control the breadth of the informative prior:

$$\hat{\pi}(\boldsymbol{\theta}; \beta) = \frac{\pi(\boldsymbol{\theta})^{\beta}}{Z(\beta)\{\pi\}}, \tag{8}$$

(see **??**) where $Z(\beta)\{\pi\}$ — a functional of $\pi(\boldsymbol{\theta})$ is a normalisation factor for $\mathcal{P}(\boldsymbol{\theta})$, i.e.

$$Z(\beta)\{\pi\} = \int_{\Psi} \pi(\boldsymbol{\theta})^{\beta} d\boldsymbol{\theta}. \tag{9}$$

In their prescription, the likelihood changes to

$$\hat{\mathcal{L}}(\boldsymbol{\theta}; \beta) = \mathcal{L}(\boldsymbol{\theta}) Z(\beta)\{\pi\} \cdot \pi^{1-\beta}(\boldsymbol{\theta}). \tag{10}$$

The new parameter β is treated as any other non-derived parameter of the original theory.

Note, that $\mathcal{L}(\boldsymbol{\theta})\pi(\boldsymbol{\theta}) = \hat{\mathcal{L}}(\boldsymbol{\theta})\hat{\pi}(\boldsymbol{\theta})$ by construction. Thus, from **??** the posterior and evidence corresponding to $\hat{\mathcal{L}}(\boldsymbol{\theta}; \beta)$ and $\hat{\pi}(\boldsymbol{\theta}; \beta)$ will be the same as $\mathcal{P}(\boldsymbol{\theta})$ and $\mathcal{Z}$, which correspond to the original $\pi(\boldsymbol{\theta})$ and $\mathcal{L}(\boldsymbol{\theta})$.

If the informative prior $\pi(\boldsymbol{\theta})$ is less representative of the posterior $\bar{\mathcal{P}}(\boldsymbol{\theta})$, error in $\hat{\mathcal{Z}}$ is larger. Hence, while we don't violate **??** directly, $\bar{\mathcal{Z}}$ can be more different from $\mathcal{Z}$ while remaining within margin of error, and similarly $\mathcal{P}(\boldsymbol{\theta}) \neq \bar{\mathcal{P}}(\boldsymbol{\theta})$. This is where the new parameter comes into play. $\hat{\pi}$ may become representative for some value of $\beta = \beta_R$. Values β

close to β_R correlate with higher likelihoods, thus the sampler prefers them. Hence, the system will converge to a state where $\mathcal{P}(\boldsymbol{\theta})$ is represented in $\hat{\pi}(\boldsymbol{\theta};\beta)$[4]. As a consequence, we reduced the errors and obtained the same result as we would have with a less informative but more representative prior.

? dubbed this ***automatic power posterior repartitioning*** (PPR) because the choice of $\beta \to \beta_R$ is automatic. It mitigates the loss precision and thus accuracy for unrepresentative informative priors π, by sacrificing performance.

3 THEORETICAL DISCOVERIES

3.1 The trouble with proposals

Nested sampling is different from Metropolis-Hastings-Gibbs and many other Markov-Chain Monte Carlo methods. Often, such algorithms are designed with a separate input that is the proposal: an initial guess that guides the algorithm towards the right answer. For nested sampling no such provisions are in place. The only location where such information can be used is the prior. Thus, to understand why one can't use proposals directly, we must first address why informative priors are avoided.

From **??**, we can see that changing only the prior π necessarily leads to changes in both $\mathcal{P}$ and $\mathcal{Z}$. For example if π is a Gaussian centered at $\boldsymbol{\theta} = \boldsymbol{\mu}_\pi$ and $\mathcal{L}$ is a Dirac δ-function peaked at $\boldsymbol{\theta} = \boldsymbol{\mu}_\mathcal{L}$, with $\boldsymbol{\mu}_\pi$ sufficiently far from $\boldsymbol{\mu}_\mathcal{L}$ then the posterior will necessarily have peaks at both $\boldsymbol{\mu}_\pi$ and $\boldsymbol{\mu}_\mathcal{L}$. This is an example of prior imprinting and is a necessary part of a Bayesian view of statistics. For a Bayesian, the prior information is no less valuable than the information inferred from the dataset $\mathfrak{D}$, and the posterior represents *all* of our best knowledge.

The problem however, is the *prejudiced sampler*. Because nested sampling chooses live points with probability proportional to the prior, the probability of a point being drawn from the likelihood peak can be made arbitrarily small. In fact, if $\boldsymbol{\mu}_\mathcal{L}$ and $\boldsymbol{\mu}_\pi$ are separated by more than five standard deviations of the prior Gaussian, thirty million samples will be drawn from $\boldsymbol{\mu}_\pi$ before a single point is drawn on the circle containing $\boldsymbol{\theta} = \boldsymbol{\mu}_\mathcal{L}$.

An apt analogy can be drawn with the Venera-14 mission (**?**). Upon landing, due to a number of unfortunate coincidences, the lander took its one and only measurement of Venusian soil from one of its own lens caps. As a result, we have obtained objectively correct information from Venus: a sample of an object on its surface. However, the efficiency of this measurement of the compressibility of rubber leaves much to be desired.

Before **?** the best solution was to use a uniform prior that included both $\boldsymbol{\mu}_\pi$ and $\boldsymbol{\mu}_\mathcal{L}$. The computational cost of inference is so high that the risk of gaining nothing from a dataset is untenable. Thus discarding all prior information in hopes of inferring some from the dataset is preferable to using the information in π.

Thus, proposals are not even considered for use with nested sampling. Since proposals may be crude approximations, we may obtain far worse than no new information. Any potential benefit in performance or precision is far outweighed by the unreliable posterior. We do, however, have one method of mitigating these problems — automatic posterior repartitioning (**?**). In the following sections we shall expand our arsenal of methods of avoiding these pitfalls and incorporating proposals into nested sampling-based inference.

[4] Technically we obtain $\hat{\mathcal{P}}(\boldsymbol{\theta};\beta)$ which, when marginalised over β, yields $\mathcal{P}(\boldsymbol{\theta}) = \int \hat{\mathcal{P}}(\boldsymbol{\theta};\beta)d\beta$ — the correct posterior.

3.2 Intuitive proposals and accelerated convergence

Consider the following premise: we're given a model $\mathcal{M}$, for which our prior π is not the uniform $\bar{\pi}(\boldsymbol{\theta})$. This usually means that from other sources, e.g. other inferences, physical reasoning, etc., we know that

$$\pi(\boldsymbol{\theta}) = f(\boldsymbol{\theta};\boldsymbol{\mu},\boldsymbol{\Sigma}), \tag{11}$$

which is representative of the posterior $\bar{\mathcal{P}}(\boldsymbol{\theta})$. Here, the probability density function f is parameterised by $\boldsymbol{\mu}$ in its location and $\boldsymbol{\Sigma}$ its breadth. In order to obtain the same result as one would have with the less informative uniform prior $\bar{\pi}(\boldsymbol{\theta})$, one needs to correct the likelihood $\mathcal{L}$. Recall, that the reason why PPR obtains the same posterior $\bar{\mathcal{P}}(\boldsymbol{\theta}) = \hat{\mathcal{P}}(\boldsymbol{\theta})$ as one would have using $\bar{\pi}(\boldsymbol{\theta}) = \text{Const.}$ is because $\hat{\mathcal{L}}(\boldsymbol{\theta};\beta)$ and $\hat{\pi}(\boldsymbol{\theta};\beta)$ are a ***consistent (re)partitioning*** of $\bar{\mathcal{Z}}$ and $\mathcal{P}(\boldsymbol{\theta})$. That is:

$$\int_\Psi \hat{\mathcal{L}}(\hat{\boldsymbol{\theta}})\hat{\pi}(\hat{\boldsymbol{\theta}})d\hat{\boldsymbol{\theta}} = \int_\Psi \bar{\pi}(\boldsymbol{\theta})\bar{\mathcal{L}}(\boldsymbol{\theta})d\boldsymbol{\theta} = \bar{\mathcal{Z}}, \tag{12}$$

where in the case of PPR $\hat{\boldsymbol{\theta}} = (\theta_1, \theta_2, \ldots, \theta_n, \beta)$. **??** holds if

$$\hat{\mathcal{L}}(\boldsymbol{\theta};\beta)\hat{\pi}(\boldsymbol{\theta};\beta) = \bar{\mathcal{L}}(\boldsymbol{\theta})\bar{\pi}(\boldsymbol{\theta}) \tag{13}$$

for all β, by **??**. Note that **?** have used **??** as the primary expression. Following their convention, we shall sometimes refer to consistent partitions as posterior repartitioning, rather than evidence repartitioning.

By using a more informative prior in this way, we accelerates convergence, because each iteration obtains a larger evidence estimate, so fewer are needed to reach the termination point (See **??**). There is a competing mechanism: the evidence estimates accumulate fewer errors, so inference proceeds longer before the precision loss triggers termination (**??**). Thus repartitioning reaches a more precise result quicker. Of course the obtained precision can be sacrificed to further accelerate inference.

3.2.1 Example: intuitive proposal posterior repartitioning

Suppose that one has obtained the posterior $\mathcal{P}(\boldsymbol{\theta})$ from a different inference. This could be nested sampling with a uniform prior, or Hamiltonian Monte Carlo, or a theoretical approximation. Under these circumstances,

$$\hat{\pi}(\boldsymbol{\theta}) = f(\boldsymbol{\theta},\boldsymbol{\mu},\boldsymbol{\Sigma}) = \mathcal{P}(\boldsymbol{\theta}), \tag{14a}$$

is an informative prior that represents our knowledge, but might not represent the posterior. We call it an ***(intuitive) proposal***. However, we wish to avoid prejudicing the sampler and use the (uniform) reference prior $\bar{\pi}(\boldsymbol{\theta})$, with reference likelihood $\bar{\mathcal{L}}(\boldsymbol{\theta})$.

To obtain with $\hat{\pi}(\boldsymbol{\theta})$ the same posterior and evidence as one would have with $\bar{\pi}(\boldsymbol{\theta})$ and $\bar{\mathcal{L}}(\boldsymbol{\theta})$, the partitioning of

the (evidence) needs to be ***consistent*** with the reference. Specifically:

$$\hat{\mathcal{L}}(\boldsymbol{\theta}) = \frac{\bar{\pi}(\boldsymbol{\theta})\bar{\mathcal{L}}(\boldsymbol{\theta})}{f(\boldsymbol{\theta};\boldsymbol{\mu},\boldsymbol{\Sigma})}. \tag{14b}$$

This we call ***intuitive proposal posterior***[5] ***repartitioning*** (iPPR). It is the fastest possible and the least robust consistent partitioning scheme. While we have technically addressed the change in $\mathcal{P}$ due to a different prior, we have not addressed the problem of $\hat{\pi}$ being (potentially) unrepresentative of $\bar{\mathcal{P}}$. In the example already considered in **??**, we will have reduced prior imprinting, but not all addressed the prejudice. The probability of sampling from the true likelihood peak is still minuscule. By contrast, we have seen that automatic power posterior repartitioning can mitigate both issues. What iPPR lacks, is a mechanism for extending its representation. Rather than attempt a modification akin to power partitioning, in **??** we shall provide this mechanism as completely external to iPPR and unleash its potential.

3.3 General automatic posterior repartitioning

In this section, we look at the family of prescriptions similar to PPR and iPPR called consistent partitioning. We note which schemes are more useful for the task of accelerating nested sampling without biasing the posterior. We begin by noting, that **??** alone does not guarantee the correct posterior and evidence.

We shall consider a general consistent partitioning $\hat{\pi}, \hat{\mathcal{L}}$ with re-parametrisation $\hat{\boldsymbol{\theta}}$. Because $\boldsymbol{\theta} \neq \hat{\boldsymbol{\theta}}$, generally, the posterior $\mathcal{P}(\hat{\theta})$ would not have the same functional form as $\bar{\mathcal{P}}(\boldsymbol{\theta})$. Nonetheless, if inverting the parametrisation from $\hat{\boldsymbol{\theta}}$ to $\boldsymbol{\theta}$ is possible, and under that procedure $\hat{\mathcal{P}}$ maps to $\mathcal{P}$, we shall say that $\hat{\mathcal{P}}$ is marginalised to $\mathcal{P}$. Thus, the correct posterior is one that marginalises to $\bar{\mathcal{P}}$. We shall often use $\hat{\mathcal{P}}(\hat{\theta})$ and $\mathcal{P}(\boldsymbol{\theta})$ that it marginalises to, interchangeably.

We can rigorously prove[6], that the following conditions are necessary for a consistent partitioning to yield the correct posterior and evidence through Bayesian inference.

(i) **Consistency**. The partitioning is consistent i.e. satisfies **??**.

(ii) **Representation**. In prior hyperspace $\hat{\Psi} \supset \Psi$ there exists a subspace $\Psi_R \subset \hat{\Psi}$, such that for all $\hat{\boldsymbol{\theta}} \in \Psi_R$, $\mathcal{P}(\boldsymbol{\theta})$ is represented in $\hat{\pi}(\hat{\boldsymbol{\theta}})$. In other words, the re-parameterised prior includes a representative configuration.

(iii) **Convergence**. The sampling favours representative configurations $\hat{\theta} \in \Psi_R$.

(iv) **Objectivity**. The prior bias (towards $\hat{\pi}(\hat{\boldsymbol{\theta}})$) is weaker than the posterior bias (towards $\hat{\mathcal{P}}(\hat{\boldsymbol{\theta}})$).

Note that these properties are contingent on the sampling algorithm. For example, in the case of inference by integration $\mathcal{Z}$ using uniform rasterisation, all properties follow from **??**. Not so for a class of algorithms that estimate $\mathcal{Z}$ by controlled error propagation and approximation, e.g. nested sampling. Thus, understanding the circumstances wherein these conditions are violated, may clarify the circumstances where both PPR and iPPR fail to produce the expected result.

Firstly, they satisfy **??** by construction. iPPR satisfies **??** if and only if $\hat{\pi}(\boldsymbol{\theta})$ represented the correct posterior to begin with, in which case $\Psi_R = \Psi$. **??** follows from the correctness proof of nested sampling (**?**), and **??** if **??** is also satisfied. In **??** we have shown that PPR satisfies **??**, where $\Psi_R = \{\beta = \beta_R = \text{Const.}\}$, if β_R exists. There's always at least one: $\Psi_R = \text{Locus}\{\beta_R = 0\} \cap \Psi$, but we are interested in values of $\beta_R > 0$, as such priors are more informative. In that section we have provided an intuitive explanation for why PPR has **??**.

However, this does not guarantee the correct posterior, indeed in **??**, we see that both θ_0 and θ_2 marginalised posteriors are offset from the correct result obtained using $\bar{\pi}(\boldsymbol{\theta}) = Const.$. This is an illustration, of the importance of **??**, as the test case **??** was constructed to violate it specifically.

3.4 Isometric mixtures of repartitioning schemes

In this section we shall consider two methods of combining several proposals (consistent partitions) into one (consistent partition). Identifying the posterior to which points in Ψ correspond to by **??**, as a metric, the mixture is isometric: the metric in the new, parameterised space marginalises to the original metric $\mathcal{P}$.

3.4.1 Additive isometric mixtures

Consider m consistent repartitioning schemes of the same posterior $\bar{\mathcal{P}}(\boldsymbol{\theta})$:

$$\hat{\mathcal{L}}_1(\boldsymbol{\theta})\hat{\pi}_1(\boldsymbol{\theta}) = \hat{\mathcal{L}}_2(\boldsymbol{\theta})\hat{\pi}_2(\boldsymbol{\theta}) = \ldots = \hat{\mathcal{L}}_m(\boldsymbol{\theta})\hat{\pi}_m(\boldsymbol{\theta}). \tag{15}$$

Their ***isometric mixture***, is a consistent partitioning that involves information from each constituent prior.

For example: an ***additive mixture*** **??**, defined as

$$\hat{\pi}(\boldsymbol{\theta};\boldsymbol{\beta}) = \sum_i \beta_i \hat{\pi}_i(\boldsymbol{\theta}), \tag{16a}$$

$$\hat{\mathcal{L}}(\boldsymbol{\theta};\boldsymbol{\beta}) = \frac{\sum_i \beta_i \hat{\pi}_i(\boldsymbol{\theta})\hat{\mathcal{L}}_i(\boldsymbol{\theta})}{\sum_i \beta_i \hat{\pi}_i(\boldsymbol{\theta})}, \tag{16b}$$

parameterised by $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_m)$ where each $\beta_i \in [0, 1]$. It is itself a consistent partitioning, i.e. ***isometric***, if and only if $\sum_i \beta_i = 1$.

Isometric mixtures are an attempt to relax some of the limitations imposed by power posterior repartitioning. Firstly, all proposals in PPR have to be linked by a power relation. This class always includes a uniform prior, but not, for example, a "wedding cake" prior (stepped uniform prior). Additive mixtures permit such proposals. Moreover, in additive isometric mixtures, any consistent partitions are compatible provided the set union of their domains matches Ψ.

However, additive mixtures have limited utility: they are slow, difficult to implement and susceptible to numerical instability more than any other consistent partitioning[7]. We can, however do much better.

[5] More accurately evidence repartitioning, which is equivalent in simple cases.

[6] Albeit in more than 5,000 words.

[7] These claims shall be substantiated in a more detailed publication.

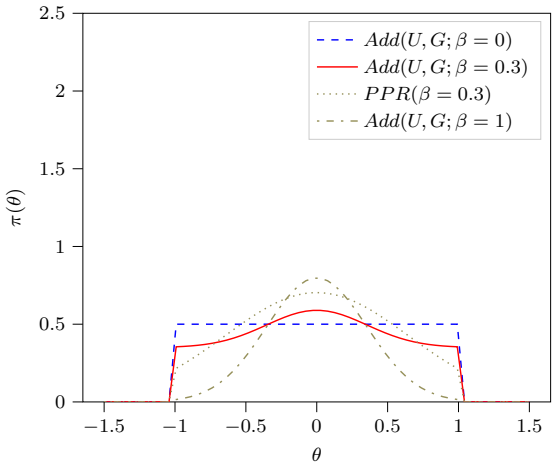


Figure 2. An additive isometric mixture of a Gaussian proposal and a uniform reference. Power-Gaussian added for comparison.

3.4.2 Stochastic superpositional isometric mixtures

One major problem with additive mixtures lies in the definition of $\hat{\mathcal{L}}$. Instead of having to evaluate only one of the constituent likelihoods, we are forced to evaluate all of them. This sets a lower bound on time complexity

$$\mathcal{T}\{\hat{\mathcal{L}}\} = o\left(\max_i \mathcal{T}\{\mathcal{L}_i\}\right), \tag{17}$$

which is the average case when the likelihoods $\mathcal{L}_i$ are all related to the same reference (e.g. $\bar{\mathcal{L}}$) with only minor corrections computed asynchronously to account for different proposals. If $\mathcal{L}_i$ and $\mathcal{L}_j$ have no common computations to reuse, the average case time complexity is $o\left[\mathcal{T}(\mathcal{L}_i) + \mathcal{T}(\mathcal{L}_j)\right]$.

Another issue is that the overall likelihood depends on the prior PDFs of the constituents. This is problematic since nested sampling requires specification of the prior via its quantile (**???**). Function inversion is not linear with respect to addition, so the quantile of the weighted sum needs to be evaluated for each type of mixture individually. For a linear combination of uniform priors, evaluating the quantile can be performed analytically, but not in case of two Gaussians or a Gaussian mixed with a uniform. By contrast, the quantile of PPR with an uncorrelated[8] Gaussian proposal is found in closed form.

We thus try to avoid mathematical operations that require evaluation of all of the constituents' priors/likelihoods. This naturally leads to deterministic prior branching, which circumvents the difficulties with determining the quantile of the mixture. If the probability of said choice can be tuned using a parameter, it can be made part $\hat{\boldsymbol{\theta}}$ similarly to β in PPR. This provides the mechanism needed for **??**.

Hence, we propose that a ***superpositional mixture***, defined via the following parametrisation:

$$\hat{\pi}(\boldsymbol{\theta}; \boldsymbol{\beta}) = \begin{cases} \hat{\pi}_1(\boldsymbol{\theta}) & \text{with probability } \beta_1, \\ \quad\vdots & \\ \hat{\pi}_n(\boldsymbol{\theta}) & \text{with probability } (1 - \sum_i^m \beta_i), \end{cases} \tag{18a}$$

$$\hat{\mathcal{L}}(\boldsymbol{\theta}; \boldsymbol{\beta}) = \begin{cases} \hat{\mathcal{L}}_1(\boldsymbol{\theta}) & \text{with probability } \beta_1, \\ \quad\vdots & \\ \hat{\mathcal{L}}_m(\boldsymbol{\theta}) & \text{with probability} (1 - \sum_i^m \beta_i). \end{cases} \tag{18b}$$

is isometric, if and only if

$$\hat{\pi}(\boldsymbol{\theta}; \boldsymbol{\beta}) = \hat{\pi}_i(\boldsymbol{\theta}) \Leftrightarrow \hat{\mathcal{L}}(\boldsymbol{\theta}; \boldsymbol{\beta}) = \hat{\mathcal{L}}_i(\boldsymbol{\theta}; \boldsymbol{\beta}), \tag{18c}$$

that is, the branches are chosen consistently.

The **??** is satisfied, if any of the priors $\hat{\pi}$ represented the posterior. The **??** is satisfied similarly to PPR: the likelihood is determined by $\hat{\theta} \supset \boldsymbol{\beta}$, so $\boldsymbol{\beta}$s that lead to higher likelihoods are favoured, ergo configurations representing $\mathcal{P}$ are preferred.

Superpositional mixtures have multiple advantages when compared with additive mixtures. Crucially, only one of $\mathcal{L}_i$ is evaluated each time $\hat{\mathcal{L}}$ is evaluated. As a result, ignoring the overhead of branch choice, the worst-case time complexity is the same if not better than the best case for additive mixtures. This has vast implications discussed in **??**.

The superpositional mixture's branch choice must be external to and independent from the likelihoods and priors. For example, the prior quantile of the mixture must branch into either of the component prior quantiles. As a result, the end user doesn't need to perform any calculations beyond the proposal quantiles themselves.

There can be many implementations of a superpositional mixture. A natural first choice would be a quantum computer, where the $\hat{\pi}$ and $\hat{\mathcal{L}}$ are represented by m level systems entangled with each other (consistent branching) and a classical computer (to evaluate $\mathcal{L}$ and π). However, we can also attain an implementation using only computational methods via stochastic deterministic choice based on $\boldsymbol{\theta}$.

The ***stochastic superpositional (isometric) mixture*** of consistent partitioning (SSIM) ensures branch consistency by requiring

$$\hat{\pi}(\boldsymbol{\theta}; \boldsymbol{\beta}) = \hat{\pi}_{F(\boldsymbol{\theta}; \boldsymbol{\beta})}(\boldsymbol{\theta}; \boldsymbol{\beta}), \tag{19}$$

where $F : (\boldsymbol{\theta}, \boldsymbol{\beta}) \mapsto i \in \{1, 2, \ldots, m-1\}$. In our implementation it is a niche-apportionment random number generator (sometimes called the broken stick model), seeded with the numerical `hash` of the vector $\boldsymbol{\theta}$, illustrated in **??**.

Superpositional mixtures are superior in robustness and ease of implementation. They do, nevertheless, come with one drawback. As a result of branching, the likelihood $\hat{\mathcal{L}}$ visible to the sampler, is no longer continuous (**??**). Thus a nested sampling implementation that relied on said continuity will have undefined behaviour. `PolyChord`'s slice sampling seems not affected by the discontinuity, but there may be other samplers that are.

[8] not so for a correlated Gaussian. Nonetheless, every correlated covariance matrix can be diagonalised, and included in the reparametrisation.

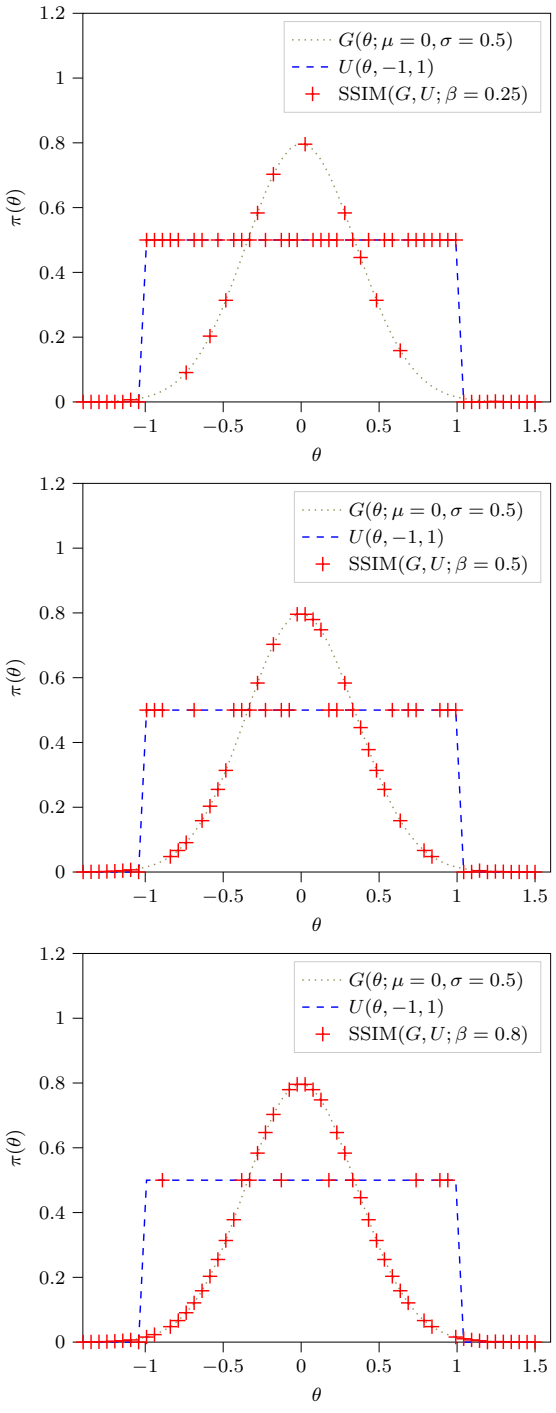


Figure 3. An example of mixture repartitioning. The mixture is not normalised to emphasise the coincidence of values with both the uniform distribution and a Gaussian. β controls the probability of belonging to the Gaussian in the stochastic mixture. Additionally, the resolution is deliberately reduced, to contrast behaviour of all three at the truncation boundary.

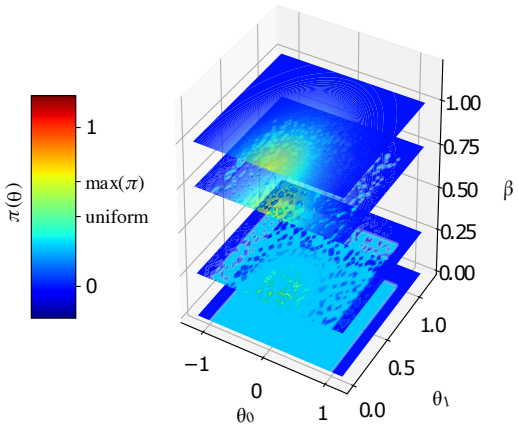


Figure 4. An illustration of SSIM in two dimensions. Colour represents the value of $\pi(\boldsymbol{\theta})$. As a result of nested sampling, nucleation of the representative phase is dynamically favoured.

3.5 On notation and mental models

It is opportune time to discuss a subtlety that we have previously neglected. chen-ferroz-hobson originally named the technique automatic posterior repartitioning, which evokes a clear mental model. Assuming that the original definitions of π and $\mathcal{L}$ were a partitioning of only the posterior, a new value of β produces a new partitioning, thus it re-partitions the posterior. The extra parameter is a time-like object, with a clear direction of evolution, in that any change to its value causes a re-partitioning of the model.

While this mental model had served well for the purposes of solving the unrepresentative prior problem, it is severely limiting to the effect of introducing proposals.

The first ineptitude of the mental model is that the expression "re-partitioning" implies the mutability of the posterior. It is not mutable. In fact, the posterior that we obtained via re-partitioning has a strict functional dependence on the parameter, which is strictly a different function. Meaningful information is lost when we project the repartitioned result to the original prior space, albeit only a Bayesian would regard it as such.

A second deeper problem is that the notation inherently puts impetus on the posterior. In reality automatic posterior repartitioning is a necessary, but insufficient condition for consistent partitioning. As long as no coordinate transformation is performed, the difference is negligible. However, for more complicated cases, e.g. re-sizeable prior space schemes, the posterior repartitioning is under-determined. A naive extension doesn't and indeed can't produce the expected result, if one considers an extension similar to

$$\pi(\theta)\mathcal{L}(\theta) = \hat{\pi}(\theta)\hat{\mathcal{L}}(\theta) \tag{20}$$

one shall obtain nonsense. One can prove (by considering a reference prior space from which all prior spaces of the same dimensionality derive via coordinate transformation), that the correct expression is actually one that preserves the evidence differential element.

What we propose is a much more general world-view

and a more accurate and expressive model. A consistent partitioning involves specifying a hyperspace that includes the original prior space. The partitioning into π and $\mathcal{L}$ is done once only, when the Bayesian inference problem is set up. The original posterior is a function in the original prior (sub)space. The posterior we obtain as a result, is the original in some projections, the evidence to which it corresponds is also the same as the original.

One might object that this is not a good model for the superpositional mixture, as the dynamical analogy would be much more appropriate, as the parameters really only control the partitioning. This point is partially valid. I would advocate seeing superpositions as an extension into a hilbert space of vectors that are themselves spaces. Not easy to imagine, but to someone fluent in Quantum theory, not a challenge. A better analogy would be to imagine the spaces for each individual prior side by side, and have a few parameters that control the relative "heights" of these spaces, or activation energy for diffusion. This is a middle-ground that retains the generality of treating the entire problem in a hyperspace, but also has a dynamical analogy.

Arguments can be made either way, but an important consideration is to have a model that gives accurate predictions first, and is easy to imagine second.

4 MEASUREMENTS AND METHODOLOGY

Our measurements have to ascertain three key points. First we must prove that the consistent partitions obtain sensible estimates of $\mathcal{P}$ and $\mathcal{Z}$ and document the circumstances when they don't. We shall then need to measure the performance uplift that can be attained while preserving the accuracy and precision of the sampling. Lastly, we must attempt to apply this to real-world example: Cosmological parameter estimation.

For performance, we shall adopt the weighted accounting approach (**?**) for measuring time complexity in units of $\mathcal{N}\{\mathcal{L}\}$, and reducing all quantities to their long-run averages. Consequently, all of the partitions' overheads associated with internal implementation details are ignored. This is to ensure fairness in comparing power repartitioning to a stochastic mixture[9].

We shall use Kullback-Leibler divergence in two contexts. First, $\mathcal{D}\{\pi, \mathcal{P}\}$ — a measure of information obtained from the dataset ignoring the prior, is used to gauge performance (as seen in **??**).

We also need a method of comparing posteriors to determine their accuracy. This is the Second use: $\mathcal{D}\{\mathcal{P}, \bar{\mathcal{P}}\}$ quantifies the correctness of the obtained posterior, where $\bar{\mathcal{P}}$ is the posterior obtained using a $\bar{\pi}(\boldsymbol{\theta}) = \text{Const}$. We also use $\mathcal{Z}$ to gauge correctness.

From **??**, errors in $\mathcal{P}$ are necessarily caused by errors in estimating $\mathcal{Z}$, and is the crucial reason why nested sampling is sensitive to partitioning in the first instance. Moreover, the character of error in $\mathcal{Z}$ indicates the type of error in $\mathcal{P}$. A higher than expected evidence $\mathcal{Z}$ is indicative of inconsistent partitioning, where the likelihood was not re-scaled to accommodate a more informative prior (**??**). A less than expected $\mathcal{Z}$ indicates that the regions of high $\mathcal{L}$ were not probed sufficiently. Usually, this is accompanied by bias imprinting as with PPR in **??**.

[9] SSIM has far less overhead

Table 3. Typical values of posterior-to-reference-posterior Kullback-Leibler divergence $\mathcal{D}\{\mathcal{P}, \bar{\mathcal{P}}\}$ for the runs shown in **??**. The inconsistent re-sizeable uniform had not been given an improper normalisation of $\hat{\mathcal{L}} = \mathcal{L}$. It is of type ***Re-sizeable uniform***.

Scheme	$\mathcal{D}\{\mathcal{P}, \bar{\mathcal{P}}\}$	$\mathcal{Z}$
Uniform	0.018	-62.70 ± 0.30
Analytical	0.000	-62.72 ± 0.00
R	0.724	-54.8 ± 0.90
PPR	0.011	-62.73 ± 0.01
$SSIM(U, G)$	0.007	-62.72 ± 0.01
$SSIM(U, G, R)$	0.696	-57.70 ± 0.30

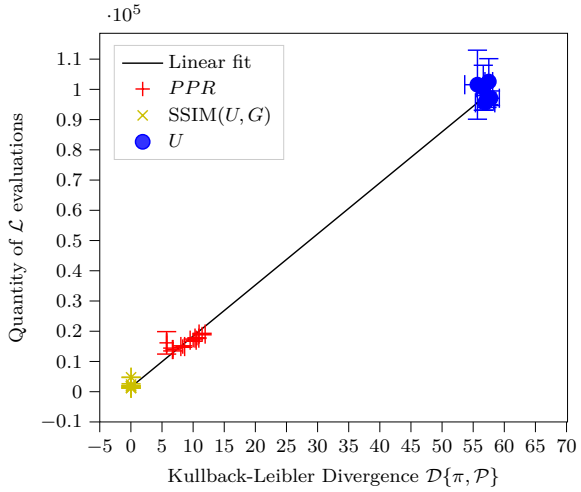


Figure 5. Scaling of number of likelihood calls with Kullback-Leibler divergence $\mathcal{D}\{\pi, \mathcal{P}\}$ With colinear offsets varying from $10\boldsymbol{\mu}$ to $300\boldsymbol{\mu}$. The best fit line is $[(1.5 \pm 0.2)\mathcal{D} + (1.7 \pm 0.1)] \cdot 10^3$ with determination coefficient $R^2 = 0.85$ which indicates that $\mathcal{D}$ is a reliable performance indicator for `PolyChord`.

When constructing the test cases, we shall use on no more than three-dimensional models with Gaussian likelihoods, as they are sufficiently general to share similarities with cosmological inference, while also being practical to investigate under small perturbations. For this purpose, we use a uniform baseline prior, and a Gaussian likelihood:

$$\ln \mathcal{L}(\boldsymbol{\theta}) = \ln \mathcal{L}^{\max} - \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{\theta} - \boldsymbol{\mu}), \qquad (21)$$

where the covariance matrix $\boldsymbol{\Sigma}$, specifies the extent of the peak, and the vector $\boldsymbol{\mu}$ — the location. $\mathcal{L}^{\max}$ is the normalisation factor, which we keep implicit, for convenience.

$\boldsymbol{\Sigma}$ is assumed diagonal, without loss of generality. While $\boldsymbol{\Sigma}$ can be singular, this usually means a redundancy in the parametrisation, which can be fixed (by turning the strongly correlated parameters derived). Otherwise it is positive semi-definite, and symmetric, meaning that the it can be diagonalised via change into its eigen-basis. Counter-intuitively, this basis must not be made part of the quantile. It is applied before computations involving correlated Gaussians, and reversed afterwards. This is a consequence of the

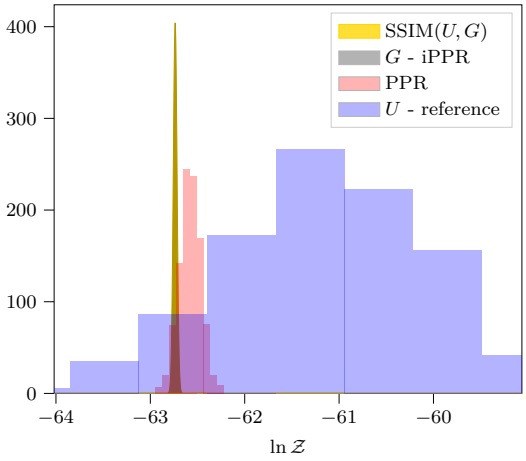


Figure 6. An illustration of the histograms for the last 1000 evidence estimates of different types of consistent partitioning. SSIM is a stochastic superposition of Gaussian iPPR (G), uniform (U). The likelihood of R — a resizeable-bounds uniform prior partition was not properly re-scaled to illustrate the effects of inconsistent partitioning. **??**.

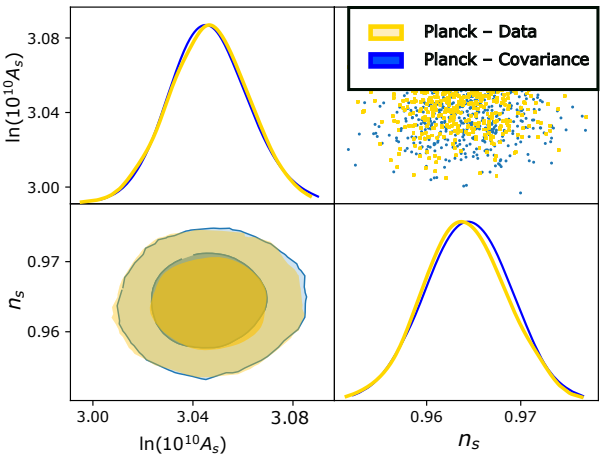


Figure 7. An example of a posterior obtained with PPR, based on Planck parameter covariance matrix, compared with the Planck posterior chains. The differences in the distributions indicate variance across different inference runs. $\mathcal{D}\{\mathcal{P}, \bar{\mathcal{P}}\} \approx 0.01$. The deviation is due to a different (smaller) number of live points used, and the difference between the correct likelihood and its approximation using a Gaussian.

extra Jacobian brought on by the difference between **??** and **??**. Essentially by applying the transformation globally the unit hypercube becomes a parallelopiped, which is the result of neglecting the Jacobian associated to the linear transformation.

To simulate imperfections we consider translational offsets between the proposal prior and the model likelihood. The main test posterior is thus

$$\bar{\mathcal{P}}(\boldsymbol{\theta}) = G(\boldsymbol{\theta}; \boldsymbol{\mu} = (1, 2, 3), \boldsymbol{\Sigma} = \mathbb{1}_3), \tag{22}$$

truncated to a cube of side length[10] $a = 1.2 \cdot 10^9$. The corresponding evidence (**??**) is $\ln \mathcal{Z} \approx -62.7$. The quantile of this Gaussian distribution is the one that enters iPPR and PPR's priors as well as the reference likelihood. All other test cases are dervied from this Gaussian either via re-scaling, deformation of variances, or translation.

The choice of the prior scale: $a = O(10^9)$, is to ensure that the series are not affected by run-to-run variance, even with a reduced number of live points. This has the added benefit of simulating an unbounded uniform prior numerically, as it is near the numerical limits. Also, any error in re-scaling the likelihood (e.g. **??**) leading to an inconsistent partition would not be obvious or as clean with a smaller prior boundary. Lastly, this allowed us to test the hypothesis that both stochastic mixtures and power posterior repartitioning can effectively remove the burn-in stage altogether. Last but not least, under such circumstances, stochastic mixtures are put at the greatest disadvantage. In the average case, approximately half the original live points are drawn from the proposal distribution and half from the uniform. The probability of finding the offset posterior peak is thus minuscule for large offsets. By contrast, In the average case the original live points with a Gaussian power posterior are drawn from a twice broad Gaussian.

5 RESULTS AND DISCUSSION.

The first test was to ensure that the repartitioning was implemented correctly, so the Gaussians entering all Gaussian repartitioning schemes were given with identical variance and mean. Results are shown in **??** and **??**.

The second class of tests involved deforming the prior Gaussians. Both SSIM (iPPR and uniform) and PPR were resilient with respect to re-scaling and anisotropic deformation of the likelihood, obtaining $\mathcal{D}\{\mathcal{P}, \bar{\mathcal{P}}\} \leq 0.03$. iPPR coped with situations where $\mathcal{P}$ was narrower than π, while failing in the opposite case: $\mathcal{D}\{\mathcal{P}, \bar{\mathcal{P}}\} \geq 5.5$, when $\mathcal{D}\{\pi, \mathcal{P}\} = 5.5$ and $\Sigma = 0.3 \times \mathbb{1}_3$.

The final test was with regards to translational offsets. The results are shown in **??????**. In **??**, we see that the amount of information extracted from PPR increases with increased offset. However, it does so sub-linearly, which combined with **??**, renders suspect the validity of the posteriors obtained using PPR and SSIM. However, **??** shows that only PPR is adversely affected.

The posterior to posterior Kullback-Leibler divergence remained stable and less than 0.3 for the stochastic mixture and the reference. Power repartitioning fluctuated considerably, ensuring that no suitable plot could be produced. This suggests instability towards perturbations, and unpredictability of the accuracy of the posterior. However, none of the values reached the prior to posterior divergence. This suggests that at no offset was the posterior entirely obtained

[10] The value 1.2 was chosen because it is the shortest non-machine representable floating point number, whose inverse is also not machine representable. This causes numerical instability in the uniform prior probability density function and quantile (at the boundaries). This was chosen for tests of boundary effects, which had to be removed from the project, because of volume constraints.

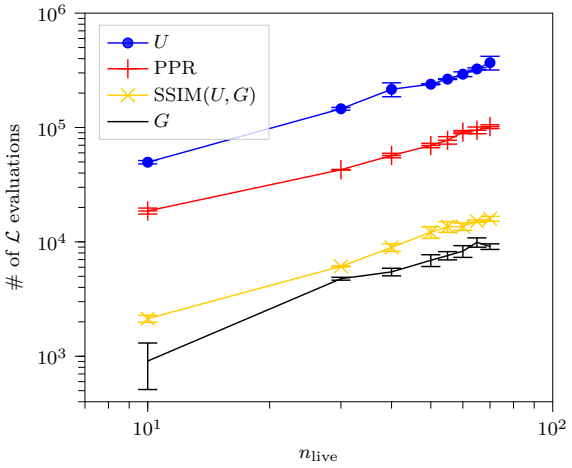


Figure 8. number of $\mathcal{L}$ evaluations as a function of the number of live points. U is the reference uniform, and G is the pure Gaussian proposal. $\max \mathcal{D}\{\mathcal{P}, \bar{\mathcal{P}}\} < 1.5$, meaning all participating consistent partitions obtained the correct posterior. The number of evaluations scales as $k \cdot n_{\rm live}^{1.1\pm0.2}$, where k reduces for faster repartitioning schemes.

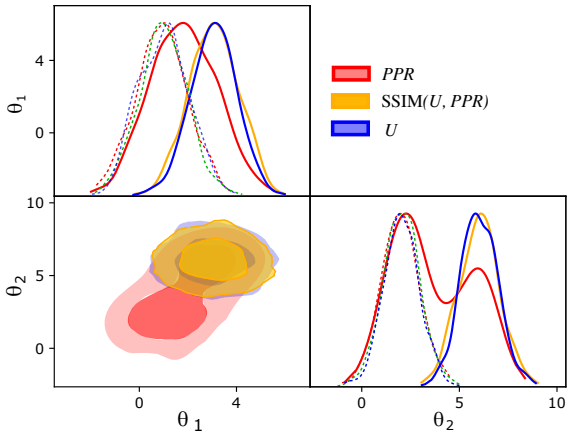


Figure 9. An illustration of offsets affecting $\mathcal{P}$ under various repartitioning schemes. Dotted series represent the prior imprint. The reference uniform and the stochastic mixture agree with the analytical posterior: Gaussian peak at $\boldsymbol{\theta} = (4, 6, 8)$.

from the prior. As a result, power repartitioning may still be useful for unrepresentative informative priors, that are not proposals, as **?** have shown.

A special case is that shown in **??**, in a reduced size bounding box $a = 2 \times 10^3$. The main notable feature is the inaccuracy of the posterior obtained by PPR. If the offset is small — $O(2\sigma)$, the posterior is shifted. With a larger offset, e.g. $O(4\sigma)$, two peaks can be resolved. Both errors are caused by incorrect evidence (see **??**) PPR: $\ln \mathcal{Z} \approx -25.4 \pm 2$, vs uniform reference $\ln \mathcal{Z} = -22.7 \pm 0.4$ and SSIM, $\ln \mathcal{Z} = -22.5 \pm 0.3$. There are two key observations to be made: the evidence is still within reasonable variance from the reference, and its estimated error is large. As a result, while we haven't obtained the right information, we know that something went wrong.

This is not at variance with **?**'s observations, as they do not have a comparable test case. All of the numerical test cases were restricted to two physical parameters, while we extended it to three. The example given required considerable fine-tuning to be reproducible[11], as larger or smaller offsets often lead to correct convergence some of the time. Another hint at why power repartitioning may have been affected more than a stochastic mixture can be gleaned from **??**. By noticing that the correct evidence is still within one standard deviation of the estimate obtained using power repartitioning we can suggest, that the result is less precise. So the unusual shape of the marginalised posterior, is the result of loss of precision. The inaccurate posterior is within margin of error of the analytical result,

It is worthwhile to consider the impact of such a scenario occurring during practical use of Bayesian inference. If either of the posterior looks as PPR's marginalised posteriors in **??**, the researcher performing the inference has the following options:

(i) accept the posterior as is
(ii) accept the posterior, but as a less credible result
(iii) reject the PPR result entirely, and perform a run with only a uniform prior
(iv) readjust the PPR mean and variance using the posterior, and re-run
(v) combine PPR with SSIM in mixture with a uniform prior

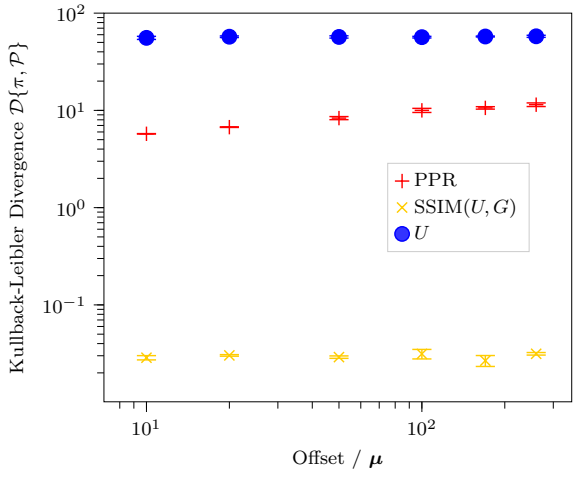
?? is a last resort. **??** is adequate for low accuracy applications, provided errors are properly estimated using e.g. `nestcheck` (**?**). From **??**, we see that the performance uplift allows for **??** to be more efficient than **??**, albeit marginally so.

This is where our technique is most useful: one obtains, as we've shown in **??**, a more accurate $\mathcal{P}(\boldsymbol{\theta})$, by using PPR from within SSIM. This results in a repartitioning scheme that is on average slower than PPR (by approximately 18% extra $\mathcal{L}$ evaluations) within margin of run-to-run variance of PPR (approximately 20%)[12], which is an order of magnitude less than **????** on this page and on the current page would afford. That said, using the proposal directly is faster still **??**.

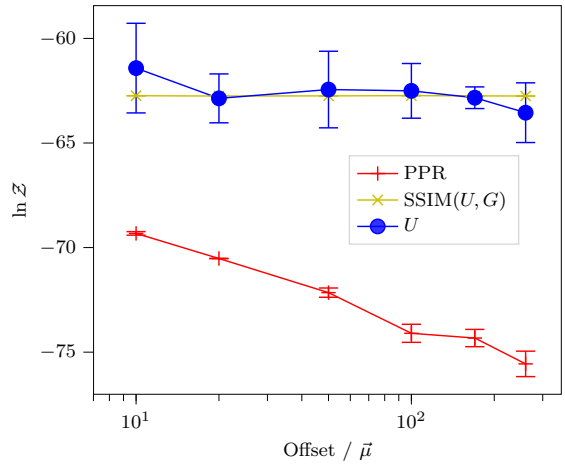
Lastly, ***posterior mass*** — a measure of convergence speed (**?**), is often used in diagnosing nested sampling. Typical examples of posterior mass for a run with $\pi = $ Const. and runs accelerated by posterior repartitioning are given in **??**. Notice that the repartitioned series has a longer extinction phase, as a result of introducing extra nuisance parameters. Also, the confidence intervals on each parameter between the uniform and the repartitioned run are identical, signifying that we have not lost precision.

[11] Too much free time in quarantine.

[12] Comparison with **??** may be miselading, as the error margins there correspond to exact coincidence, while the case in question involves an offset of 6μ.



(a) Kullback-Leibler divergence $\mathcal{D}$ for different offsets: Gaussian peaks displaced from $\boldsymbol{\mu}$ by Offset$\times\boldsymbol{\mu}$. Notice that the faster repartitioning methods produce a lower value of $\mathcal{D}$. The divergence $\mathcal{D}$ scales sub-linearly with the offset.



(b) An illustration of offsets affecting $\mathcal{Z}$. The true value is constant, mirrored by the mixture: SSIM of PPR and reference uniform. PPR alone produces incorrect evidence, consistent with **??**. Tighter errorbars on SSIM are consistent with our observations from **??**.

Figure 10. Illustrations of effects of offsets on the correctness **??** and performance **??** of nested sampling under consistent posterior repartitioning.

5.1 Cosmological Simulations.

After an initial run of `Cobaya` (**?**), we have obtained the marginalised posteriors of all the key parameters of the ΛCDM model, as well as the nuisance parameters.

First, we have performed an inference using the Planck (**?**) dataset, with the ΛCDM model. The results of our initial run are presented in **??**. From these data, under the assumption that the parameters' posteriors are a correlated Gaussian distribution, we extract the means $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$.

We use a stochastic mixture of a uniform prior and a single Gaussian obtained from the posterior samples of a run with a uniform prior, which we patch into `Cobaya`'s interface to `PolyChord` (**?**). The posteriors of two runs with identical settings (save live point number) are given in **??**.

Table 4. Accuracy metrics for Cosmology runs using Cobaya.

Prior	**Device**	$\mathcal{D}\{\mathcal{P}, \bar{\mathcal{P}}\}$	$\ln \mathcal{Z}$	n_{live}
Uniform	CSD3	0.000	-1432.8 ± 0.8	108
SSIM(U, G)	CSD3	0.2	-1433.6 ± 0.1	100
iPPR(G)	CSD3	0.4	-1433.8 ± 0.05	100
SSIM(U, G)	PC	0.25	-1433.5 ± 0.2	50

Table 5. Performance metrics for Cosmology runs using Cobaya. t is the time from beginning of sampling, to output. Starred series were extrapolated linearly. Precision normalisation assumes errors in $\mathcal{Z}$ scale as n_{live}^{-1}.

Prior	**Device used**	t**/(hrs)**	$\mathcal{N}\{\mathcal{L}\}$	n_{live}
Uniform	CSD3	32.2	480000	108
SSIM(U, G)	CSD3	1.7	90000	100
SSIM(U, G)	PC	50	49000	50
Uniform	PC*	912	240000	50
Uniform	CSD3*	224	3 360 000	700

Firstly, notice that the posteriors have a significant overlap. Each plot on the diagonal of **??** is a Gaussian, agreeing with the results of the reference run to within less than 1/10-th of a standard deviation. However SSIM predicts a deformed (non-ellipsoidal) covariance of the ΛCDM parameters.

The deformations are present in all posteriors that used a Gaussian proposal, which indicates that the deformations are systematic. The deformities are not caused by finite-grain size in the stochastic mixture, as the Gaussian proposal has them, and to a greater extent. The mixing portion parameter β, has converged to a mean of $\langle\beta\rangle = 0.82$, which indicates that the Gaussian proposal was not fully the most representative, but also that the later stages of sampling were dominated by the Gaussian proposal. Despite the appearance, however, **??** shows that the posteriors between SSIM and non-SSIM runs are not significantly different ($\mathcal{D} < 0.3$). Moreover the evidence is within one standard deviation and more precise with SSIM by a factor of 8.

While this might indicate a higher accuracy than obtainable with a pure uniform prior, one must exercise caution. While we can eliminate some potential systematic errors, a more conclusive analysis is needed.

With accuracy out of the way, **??**, highlights a significant improvement in performance. Using SSIM offers a reduction of run-time by a factor of 19. By exploiting increased precision one can reduce the number of live points, and gain a further reduction of run-time by a factor of 37. Further improvements are attainable by reducing the precision criterion and terminating early. Conversely, to obtain similar precision to SSIM, assuming sub-linear scaling with n_{live}, one would need to extend the duration of the inference to 912 hours ≈ 40 days. Assuming that errors in evidence scale as $n_{\text{live}}^{-1/2}$ the time would be then of the order of a year.

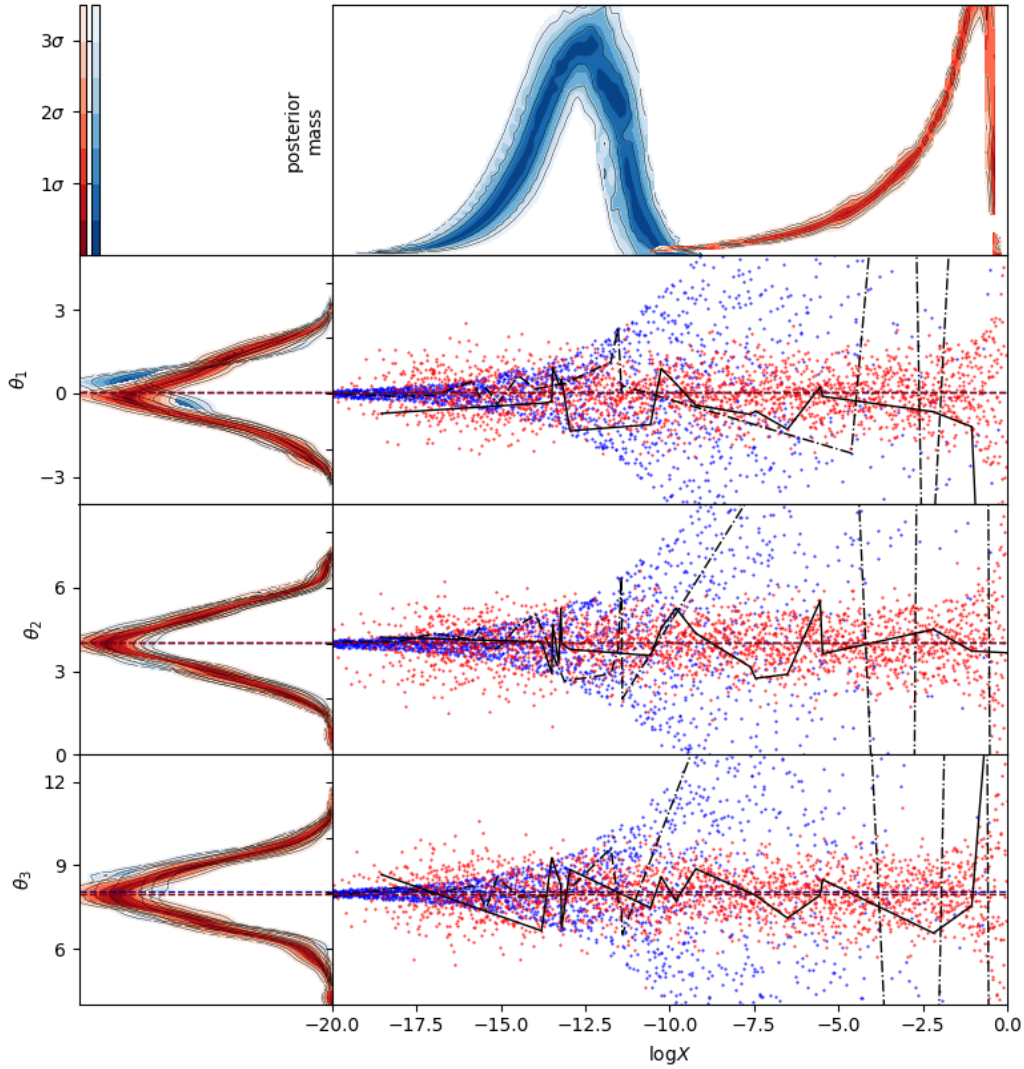


Figure 11. plot of the evolution of nested sampling. The red series corresponds to SSIM of iPPR, while the blue series — to a reference uniform. The horizontal axis of plots in the second column is $\ln X$, where $X(\mathcal{L}) \in [0, 1]$ is the fraction of the prior with likelihood greater than $\mathcal{L}$. The top plot is the relative posterior mass. In row i the $\mathcal{P}(\theta_i)$ is plotted. Confidence intervals represented with color intensity. The reference values for the model parameters are $\theta = (0, 4, 8)$

6 CONCLUSIONS

6.1 Results

The project's purpose has been to investigate the performance increase attainable by algorithmic optimisations of the inputs to nested sampling.We have identified a class of methods based on work by **?**, called consistent partitions, fit for this purpose. We have shown that each consistent partition can accelerate nested sampling when given an informative proposal. We have developed stochastic superpositional isometric mixing (SSIM), to combine several proposals, into one. When used with nested sampling, SSIM produces more

precise and accurate posteriors, faster than any individual consistent partition.

We have established the following advantages in using SSIM over PPR: SSIM admits multiple types of proposal priors, while PPR admits only one; it permits a broader class of proposals, for example: with differing domains, while PPR — only if the domains of the proposals coincide. SSIM is abstract: the prior quantile is a superposition of the constituent priors' quantiles. By contrast, PPR prior quantile needs to be calculated by the end user for each type of proposal. The calculation is non-trivial for non-Gaussian proposals. SSIM supports an unbiased reference (uniform) prior exactly. PPR tends to an unbiased reference as $\beta \to 0$, but is only truly unbiased if $\beta = 0$, with negligible probability. SSIM, like PPR, prefers the prior that leads to a higher likelihood, but unlike PPR, this does not lead to the total exclusion of less-representative priors.

As a result, faster, but more fragile consistent partitions (e.g. iPPR), in conjunction with a standard uniform prior can exceed more robust but slower PPR in precision accuracy and speed. When applied to real-world cosmological parameter estimation, our strategy of using SSIM of Uniform and iPPR resulted in a significant performance increase, reducing the run-time requirements of `Cobaya` by a factor of 30.

6.2 Further refinements

As of now, the system can be adapted to work with virtually any nested sampler in existence. All that one needs is a pseudo random number generator that can be seeded with the coordinates to produce a deterministic spread.

6.3 Applications

The obtained results are general. They can be applied in any area of any science that relies on Bayesian inference using nested sampling, e.g. astronomy (**?**). SSIM should be considered for high-performance compute applications in COVID-19 research (e.g. **??**), as inference in this field is both time and resource-intensive, while also being time-critical. It may prove useful for agent-based simulations, with complex Likelihood functions (**?**), similar to Cosmology. Identifying causal links between policies and incidence of Covid 19 cases, for example is described by 49 parameters.

Note that the asymptotic worst-case time complexity of superpositional mixtures liberates one to use as many complex models as one likes. For example: consider two libraries providing a likelihood for ΛCDM, one which makes multiple approximations (fast), and one which performs the full calculation (slow). By using the two in a superpositional mixture, one shall obtain a speedup compared to the slow run of nested sampling. This is because the slow likelihood is evaluated only some of the time. It will only be comparable to the pure slow run if the fast prior were utterly unrepresentative of the results, which itself is a valuable insight. This may be of particular interest for further refining `CLASS` and `Cobaya`, as the time complexity of computing the likelihood is the bottleneck of modern cosmological code.

Nested sampling can also be applied to inference-related problems, such as reinforcement learning (**?**). The process of training a neural network involves estimating connection strengths between nodes of said network. Normally, this is achieved via negative feedback: connections correlated with the desired behaviour are reinforced, and vice versa (**?**). This problem maps neatly onto Bayesian inference when identifying connections strengths as parameters of a model, and likelihood — correlation with desired behaviour. Most neural networks are trained with uniform priors.

We may also extend Bayesian analysis to ***consistent Bayesian meta-analysis***. Consider data obtained from multiple physical processes that are described in one theory with an overlapping set of parameters θ. As of now, we only perform separate analyses of each experiment. However, SSIM allows us to combine these models, and naturally represents consistency in the posteriors of the shared parameters. As an example, all of the estimates of the age of the universe may be obtained in one fell swoop from all the available models and data. This will have the bonus of highlighting datasets that are incompatible with the overall conclusion, allowing us to re-evaluate the experimental data as needed[13].

This is related to the issue of discordant datasets (**?**), and Bayes factor as a method of combining datasets. The idea is not new: usage of evidence as the sole judge of consistency between a model and a dataset had been discussed as long as the subject of Bayesian inference exists. Multiple metrics had been proposed e.g. **?**.

However, we propose a different delineation of datasets. Instead of considering the results of some early experiments as parts of the prior, and considering their agreement with newer observations only, we propose clearing the prior of anything but the theoretical constraints violation of which would lead to the theory being disproved. For example, if our theory predicts no negative-mass dark matter, our prior is uniform in the positive Ω_c. The data that used to be part of the prior inextricably, are now considered proposals. In Bayesian meta-analysis, our prior is a stochastic mixture of all previous observations of dark matter and the aforementioned constrained uniform prior. To clarify, this does not imply a mixture of just two priors. If the existence of dark matter can be (and was) inferred from n datasets, then our mixture is of as many as $n+1$ priors, and would consist of the posteriors of the analysis of the experiments used as proposals. The joint likelihood is suitably programmed. Due to the consistent branching, there is no "cross-talk" between likelihoods. However, the marginalised posteriors would indicate the best fit parameter distributions and take consistency and precision of different observations into account. Effectively, this synthesises data into a coherent model, without artificially splitting the model into different experimental datasets, and requiring manual reconcilliation.

The posteriors for the branch probabilities would be a measure of the consistency of specific experiments. If nested sampling chose to ignore e.g. the Type IA supernova datasets, it may suggest that such experiments are systematically inconsistent with other observations. It is much better than attempting to reconcile the discrepant datasets

[13] Additional, more detailed explanations shall be published in a paper submitted to the ***Monthly Notices of the Royal Astronomical Society***.

manually, as people are prone to fallacies. Moreover, for experiments for which data is still prserved, can be continuously integrated into a joint posterior. This may reveal cases where data was doctored to fit a particular conclusion. In such cases, the marginalised posteriors will show unusual covariances, and be outliers in the analysis.

In conclusion, the new methodology of combining information from many priors shows great promise in the field of Bayesian inference. It has demonstrably reduced the runtime of some of the most complex problems: that of Cosmological Parameter Estimation. A rich field of research awaits those courageous-enough to follow. It is ours but to point the way.

APPENDIX A: CODE

All code used to generate the plots, the framework for systematising consistent partitons as well as the configurations of `Cobaya` for cosmological simulations can be found on Github (**?**). In a separate repository (**?**) is the version of Cobaya with our modifications, which was used to produce the figures overleaf.

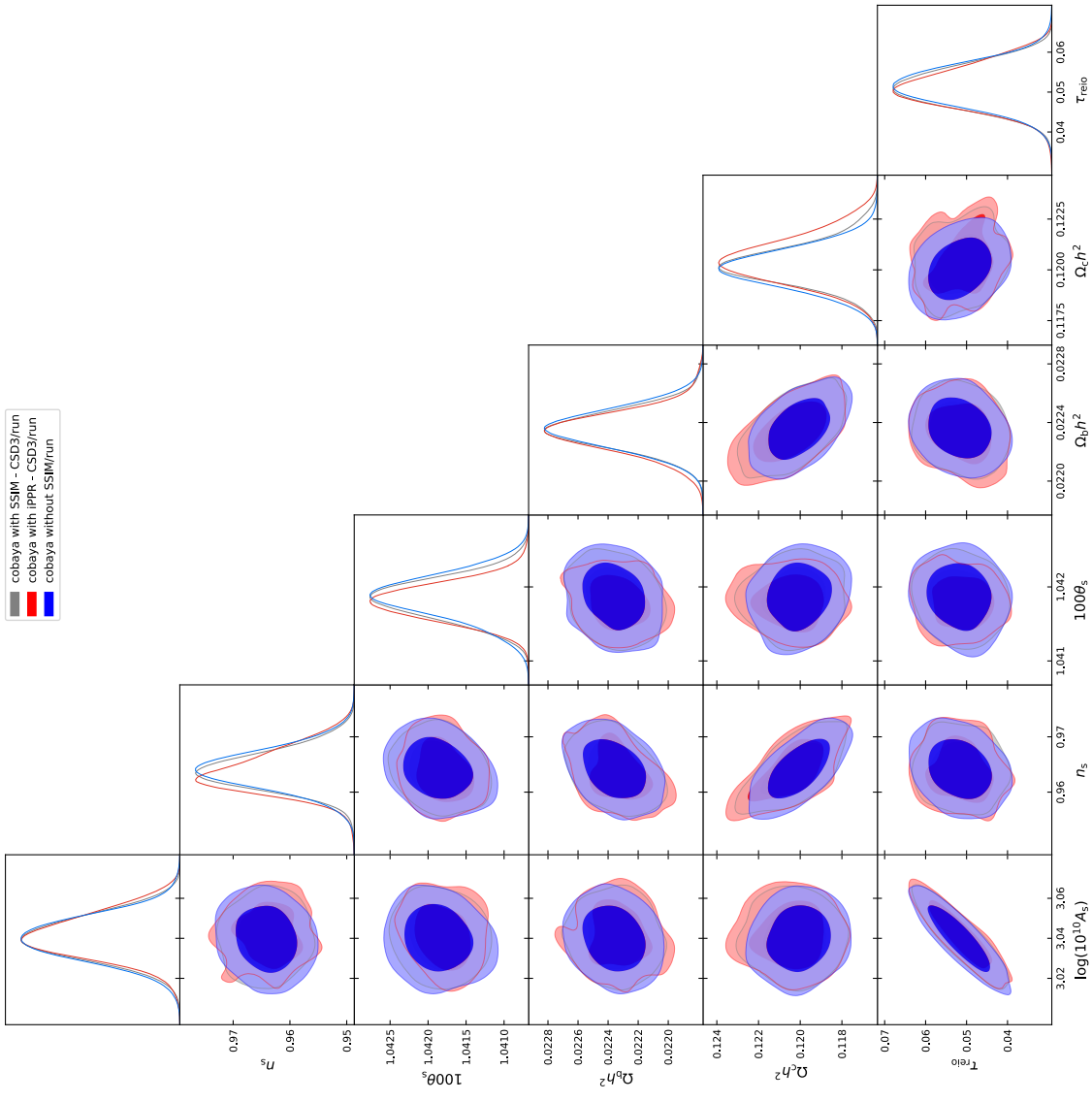


Figure A1. The marginalised posteriors for `Cobaya` + `Class` on CSD3 with $n_{\rm live} = 100$. The Reference uniform is red , while SSIM is blue . With the exception of $n_{\rm s}$ and $\Omega_{\rm c}$, all parameters are more tightly constrained. iPPR added to rule out finite-grain-size effects for partially representative priors.

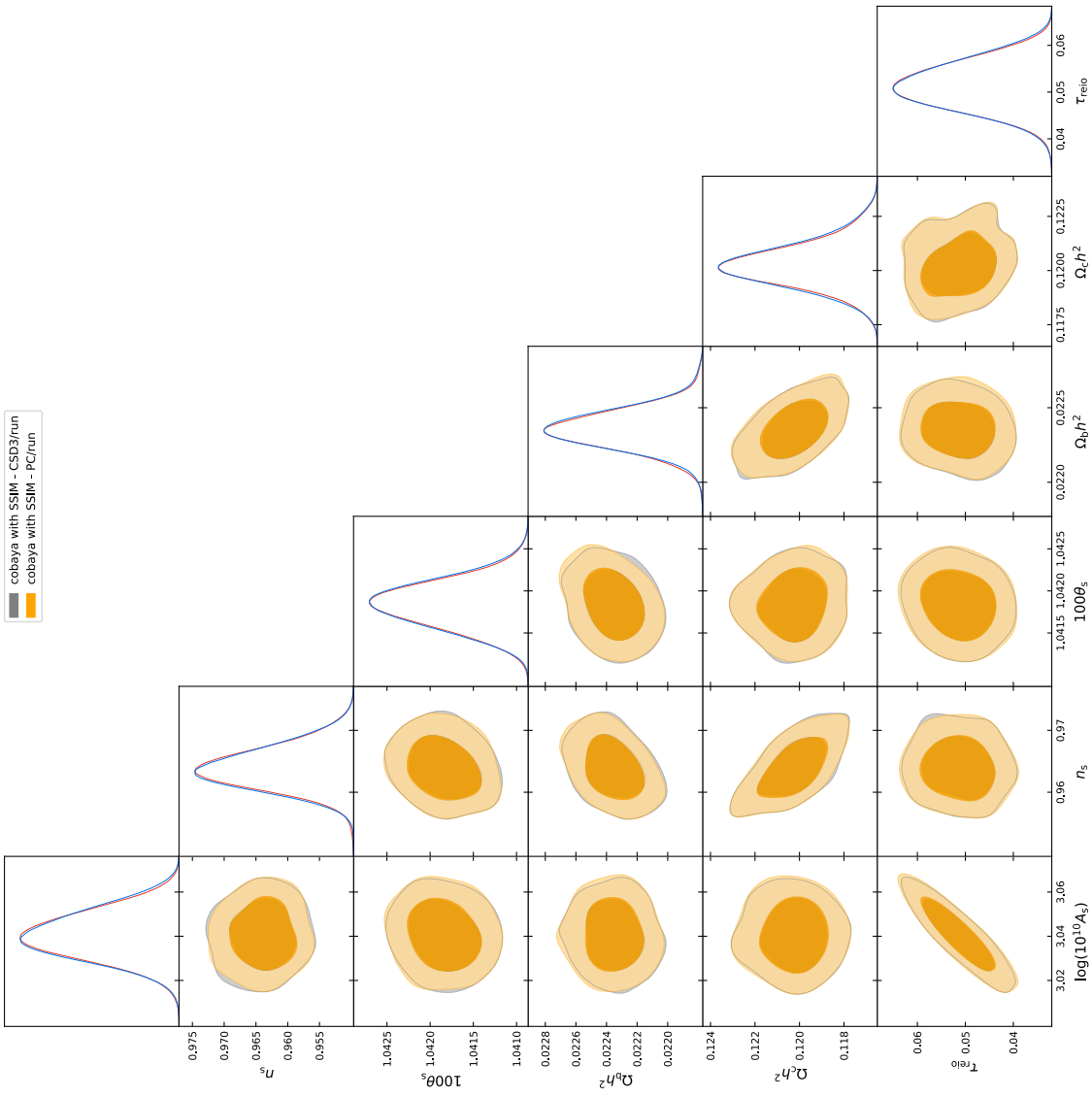


Figure A2. The marginalised posteriors for `Cobaya` + `Class` on CSD3 with $n_{\rm live} = 100$ vs PC $n_{\rm live} = 50$.