VERSION NOVEMBER 27, 2019
Preprint typeset using LaTeX style openjournal v. 09/06/15

IMPORTANCE NESTED SAMPLING AND THE MULTINEST ALGORITHM

F. FEROZ[1] E-MAIL: F.FEROZ@MRAO.CAM.AC.UK, M.P. HOBSON[1], E. CAMERON[2] AND A.N. PETTITT[3]
[1] ASTROPHYSICS GROUP, CAVENDISH LABORATORY, JJ THOMSON AVENUE, CAMBRIDGE, UK
[2] BIG DATA INSTITUTE, LI KA SHING CENTRE FOR HEALTH INFORMATION AND DISCOVERY, UNIVERSITY OF OXFORD, UK
[3] SCHOOL OF MATHEMATICAL SCIENCES (STATISTICAL SCIENCE), QUEENSLAND UNIVERSITY OF TECHNOLOGY (QUT), BRISBANE, AUSTRALIA

Bayesian inference involves two main computational challenges. First, in estimating the parameters of some model for the data, the posterior distribution may well be highly multi-modal: a regime in which the convergence to stationarity of traditional Markov Chain Monte Carlo (MCMC) techniques becomes incredibly slow. Second, in selecting between a set of competing models the necessary estimation of the Bayesian evidence for each is, by definition, a (possibly high-dimensional) integration over the entire parameter space; again this can be a daunting computational task, although new Monte Carlo (MC) integration algorithms offer solutions of ever increasing efficiency. Nested sampling (NS) is one such contemporary MC strategy targeted at calculation of the Bayesian evidence, but which also enables posterior inference as a by-product, thereby allowing simultaneous parameter estimation and model selection. The widely-used MULTINEST algorithm presents a particularly efficient implementation of the NS technique for multi-modal posteriors. In this paper we discuss importance nested sampling (INS), an alternative summation of the MULTINEST draws, which can calculate the Bayesian evidence at up to an order of magnitude higher accuracy than 'vanilla' NS with no change in the way MULTINEST explores the parameter space. This is accomplished by treating as a (pseudo-)importance sample the totality of points collected by MULTINEST, including those previously discarded under the constrained likelihood sampling of the NS algorithm. We apply this technique to several challenging test problems and compare the accuracy of Bayesian evidences obtained with INS against those from vanilla NS.

Keywords: Bayesian methods, model selection, data analysis, Monte Carlo methods

1. INTRODUCTION

The last two decades in astrophysics and cosmology have seen the arrival of vast amounts of high quality data. To facilitate inference regarding the physical processes under investigation, Bayesian methods have become increasingly important and widely used (see e.g. Trotta 2008 for a review). In such applications, the process of Bayesian inference may be sensibly divided into two distinct categories: parameter estimation and model selection. Parameter estimation is typically achieved via MCMC sampling methods based on the Metropolis-Hastings algorithm and its variants, such as slice and Gibbs sampling (see e.g. Mackay 2003). Unfortunately, these methods can be highly inefficient in exploring multi-modal or degenerate distributions. Moreover, in order to perform Bayesian model selection (Clyde et al. 2007), estimation of the Bayesian 'evidence', or marginal likelihood, is needed, requiring a multi-dimensional integration over the prior density. Consequently, the computational expense involved in Bayesian model selection is typically an order of magnitude greater than that for parameter estimation, which has undoubtedly hindered its use in cosmology and astroparticle physics to-date.

Nested sampling (NS; Skilling 2004, 2006; Sivia and Skilling 2006) is a contemporary Monte Carlo (MC) method targeted at the efficient calculation of the evidence, yet which allows posterior inference as a by-product, providing a means to carry out simultaneous parameter estimation and model selection (and, where appropriate, model averaging). Feroz and Hobson (2008) and Feroz et al. (2009) have built on the NS framework by introducing the now-popular MULTINEST algorithm, which is especially efficient in sampling from posteriors that may contain multiple modes and/or degeneracies. This technique has already greatly reduced the computational cost of Bayesian parameter estimation and model selection and has successfully been applied to numerous inference problems in astrophysics, cosmology and astroparticle physics (see e.g. Feroz et al. 2009, 2010, 2011*a*,*b*; Bridges et al. 2009; Graff et al. 2012; White and Feroz 2010; Kipping et al. 2012; Karpenka et al. 2012, 2013; Strege et al. 2013; Teachey and Kipping 2018).

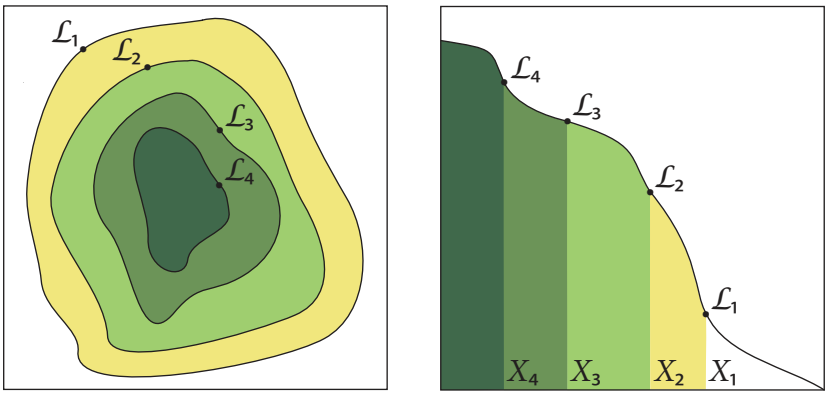
In this paper, we discuss importance nested sampling (INS), an alternative summation of the draws from MULTINEST's exploration of the model parameter space with the potential to increase its efficiency in evidence computation by up to a order-of-magnitude. Version (v3.0) of MULTINEST, which implements INS in addition to the vanilla NS scheme of previous versions, is available at `https://github.com/farhanferoz/MultiNest`.

The outline of this paper is as follows. We give a brief introduction to Bayesian inference in Sec. 2 and describe nested sampling along with the MULTINEST algorithm in Sec. 3. The INS technique is discussed in Sec. 4 and is applied to several test problems in Sec. 5. We summarize our findings in Sec. 6. Finally, in Appendix A we discuss the relationship between INS and other contemporary MC schemes, in Appendix B we give a detailed account of the convergence properties of INS within the MULTINEST algorithm, and in Appendix C we present a brief measure-theoretic commentary on vanilla NS.

2. BAYESIAN INFERENCE

Bayesian inference provides a principled approach to the inference of a set of parameters, $\boldsymbol{\Theta}$, in a model (or hypothesis), H, for data, $\mathbf{D}$. Bayes' theorem states that

$$\Pr(\boldsymbol{\Theta}|\mathbf{D}, H) = \frac{\Pr(\mathbf{D}|\,\boldsymbol{\Theta}, H)\,\Pr(\boldsymbol{\Theta}|H)}{\Pr(\mathbf{D}|H)}, \tag{1}$$



(a) (b)

FIG. 1.— Cartoon illustrating (a) the posterior of a two dimensional problem; and (b) the transformed $\mathcal{L}(X)$ function where the prior volumes, X_i, are associated with each likelihood, $\mathcal{L}_i$.

where $\Pr(\mathbf{\Theta}|\mathbf{D}, H) \equiv P(\mathbf{\Theta}|\mathbf{D})$ is the posterior probability density of the model parameters, $\Pr(\mathbf{D}|\mathbf{\Theta}, H) \equiv \mathcal{L}(\mathbf{\Theta})$ the likelihood of the data, and $\Pr(\mathbf{\Theta}|H) \equiv \pi(\mathbf{\Theta})$ the parameter prior. The final term, $\Pr(\mathbf{D}|H) \equiv \mathcal{Z}$ (the Bayesian evidence), represents the factor required to normalize the posterior over the domain of $\mathbf{\Theta}$ given by:

$$\mathcal{Z} = \int_{\Omega_{\mathbf{\Theta}}} \mathcal{L}(\mathbf{\Theta})\pi(\mathbf{\Theta})d\mathbf{\Theta}. \tag{2}$$

Being independent of the parameters, however, this factor can be ignored in parameter inference problems which can be approximated by taking samples from the unnormalized posterior only, using standard MCMC methods (for instance).

Model selection between two competing models, H_0 and H_1, can be achieved by comparing their respective posterior probabilities given the observed dataset as follows:

$$R = \frac{\Pr(H_1|\mathbf{D})}{\Pr(H_0|\mathbf{D})} = \frac{\Pr(\mathbf{D}|H_1)\Pr(H_1)}{\Pr(\mathbf{D}|H_0)\Pr(H_0)} = \frac{\mathcal{Z}_1}{\mathcal{Z}_0}\frac{\Pr(H_1)}{\Pr(H_0)}. \tag{3}$$

Here $\Pr(H_1)/\Pr(H_0)$ is the prior probability ratio for the two models, which can often be set to unity in situations where there is no strong *a priori* reason for preferring one model over the other, but occasionally requires further consideration (as in the Prosecutor's Fallacy; see also Feroz et al. 2008, 2009 for key astrophysical examples). It can be seen from Eq. (3) that the Bayesian evidence thus plays a central role in Bayesian model selection.

As the average of the likelihood over the prior, the evidence is generally larger for a model if more of its parameter space is likely and smaller for a model with large areas in its parameter space having low likelihood values, even if the likelihood function is sharply peaked. Thus, the evidence may be seen both as penalizing 'fine tuning' of a model against the observed data and as an automatic implementation of Occam's Razor.

3. NESTED SAMPLING AND THE MULTINEST ALGORITHM

Nested sampling estimates the Bayesian evidence by transforming the multi-dimensional evidence integral over the prior density into a one-dimensional integral over an inverse survival function (with respect to prior mass) for the likelihood itself. This is accomplished by considering the survival function, $X(\lambda)$, for $\mathcal{L}(\mathbf{\Theta})$, dubbed "the prior volume" here; namely,

$$X(\lambda) = \int_{\{\mathbf{\Theta}:\mathcal{L}(\mathbf{\Theta})>\lambda\}} \pi(\mathbf{\Theta})d\mathbf{\Theta}, \tag{4}$$

where the integral extends over the region(s) of parameter space contained within the iso-likelihood contour, $\mathcal{L}(\mathbf{\Theta}) = \lambda$. Recalling that the expectation value of a non-negative random variable may be recovered by integration over its survival function (a result evident from integration by parts) we have (unconditionally):

$$\mathcal{Z} = \int_0^\infty X(\lambda)d\lambda. \tag{5}$$

When $\mathcal{L}(X)$, the inverse of $X(\lambda)$, exists (i.e., when $\mathcal{L}(\mathbf{\Theta})$ is a continuous function with connected support; Chopin and Robert 2010) the evidence integral may thus be further rearranged as:

$$\mathcal{Z} = \int_0^1 \mathcal{L}(X)dX. \tag{6}$$

Indeed, if $\mathcal{L}(X)$ were known exactly (and Riemann integrable[1]), by evaluating the likelihoods, $\mathcal{L}_i = \mathcal{L}(X_i)$, for a deterministic sequence of X values,

$$0 < X_N < \cdots < X_2 < X_1 < X_0 = 1, \tag{7}$$

as shown schematically in Fig. 1, the evidence could in principle be approximated numerically using only standard quadrature methods as follows:

$$\mathcal{Z} \approx \hat{\mathcal{Z}} = \sum_{i=1}^{N} \mathcal{L}_i w_i, \tag{8}$$

where the weights, w_i, for the simple trapezium rule are given by $w_i = \frac{1}{2}(X_{i-1} - X_{i+1})$. With $\mathcal{L}(X)$ typically unknown, however, we must turn to MC methods for the probabilistic association of prior volumes, X_i, with likelihood contours, $\mathcal{L}_i = \mathcal{L}(X_i)$, in our computational evidence estimation.

3.1. *Evidence estimation*

Under the default nested sampling algorithm the summation in Eq. (8) is performed as follows. First $N_{\rm live}$ 'live' points are drawn from the prior, $\pi(\mathbf{\Theta})$, and the initial prior volume, X_0, is set to unity. At each subsequent iteration, i, the point with lowest likelihood value, $\mathcal{L}_i$, is removed from the live point set and replaced by another point drawn from the prior under the constraint that its likelihood is higher than $\mathcal{L}_i$. The prior volume contained within this region at the $i^{\rm th}$ iteration, is thus a random variable distributed as $X_i = t_i X_{i-1}$, where t_i follows the distribution for the largest of $N_{\rm live}$ samples drawn uniformly from the interval $[0, 1]$ (i.e., $\Pr(t) = N_{\rm live} t^{N_{\rm live}-1}$). This sampling process is repeated until (effectively) the entire prior volume has been traversed; the live particles moving through *nested* shells of constrained likelihood as the prior volume is steadily reduced. The mean and standard deviation of $\log t$, which governs the geometrical exploration of the prior volume, are:

$$E[\log t] = -1/N_{\rm live}, \quad \sigma[\log t] = 1/N_{\rm live}. \tag{9}$$

Since each draw of $\log t_i$ is independent here, after i iterations the prior volume will shrink down as $\log X_i \approx -(i \pm \sqrt{i})/N_{\rm live}$. Thus, one may take $X_i \approx \exp(-i/N_{\rm live})$.

3.2. *Stopping criterion*

The NS algorithm should terminate when the expected evidence contribution from the current set of live points is less than a user-defined tolerance. This expected remaining contribution can be estimated (cautiously) as $\Delta\mathcal{Z}_i = \mathcal{L}_{\max} X_i$, where $\mathcal{L}_{\max}$ is the maximum likelihood value amongst the current set of live points (with X_i the expected value of remaining prior volume, as before).

3.3. *Posterior inferences*

Although the NS algorithm is designed specifically to estimate the Bayesian evidence, inferences from the posterior distribution can be easily obtained using the final live points and the full sequence of discarded points from the NS process, i.e., the points with the lowest likelihood value at each iteration of the algorithm. Each such point is simply assigned the importance weight,

$$p_i = \frac{\mathcal{L}_i w_i}{\sum_j \mathcal{L}_j w_j} = \frac{\mathcal{L}_i w_i}{\hat{\mathcal{Z}}}, \tag{10}$$

from which sample-based estimates for the key posterior parameter summaries (e.g. means, standard deviations, covariances and so on) may be computed[2]. (As a self-normalizing importance sampling estimator the asymptotic variance of these moments is of course dependent upon both the similarity between the NS path and the target and the accuracy of $\hat{\mathcal{Z}}$ itself; cf. Hesterberg 1995.) Readers unfamiliar with importance sampling (IS) ideas may refer to Liu (2008) for an insightful overview of this topic and its application to diverse branches of modern science (including statistical physics, cell biology, target tracking, and genetic analysis).

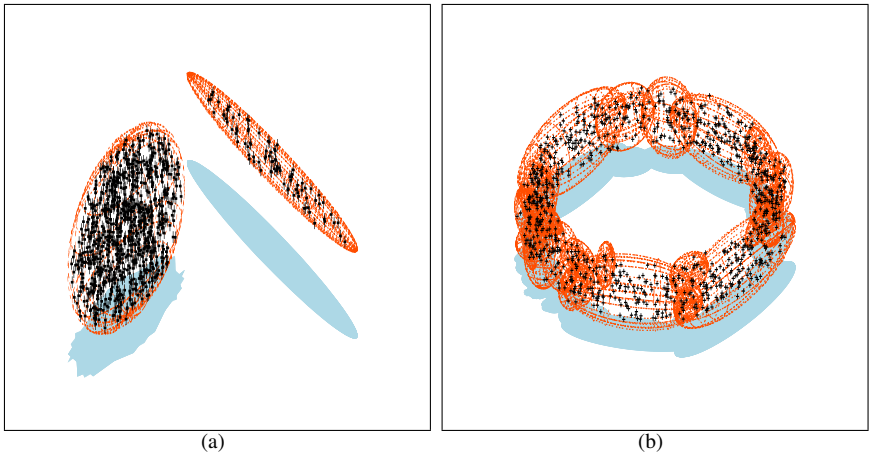
3.4. *Practical implementations of nested sampling*

The main challenge in implementing the computational NS algorithm is to draw unbiased samples efficiently from the likelihood-constrained prior. John Skilling, originally proposed to use the Markov Chain Monte Carlo (MCMC) method for this purpose (Skilling 2004, 2006; Sivia and Skilling 2006). One such implementation (Veitch and Vecchio 2010), with specific proposal distributions for the MCMC step, has been used successively in gravitational wave searches.

In astrophysics in particular, rejection sampling schemes have been successfully employed to draw samples from the likelihood-constrained prior. It was first proposed in the COSMONEST package (Mukherjee et al. 2006) through the use of ellipsoidal rejection sampling scheme and was shown to work very well for uni-modal posterior distributions. This method was improved upon in COSMOCLUST package (Shaw et al. 2007) through the use of a clustering scheme to deal with multi-modal distributions. MULTINEST was then proposed with several innovations to make ellipsoidal rejection sampling more robust in dealing with multi-modal distributions. Other methods employing ellipoidal rejection sampling scheme within Nested Sampling framework include the DIAMONDS (Corsaro and De Ridder 2014) and DYNESTY (Speagle 2019) packages.

[1] We give a brief measure-theoretic formulation of NS in Appendix C.

[2] Some relevant commentary on this aspect of NS with regard to Lemma 1 of Chopin and Robert (2010) appears in Appendix C.



(a) (b)

FIG. 2.— Illustrations of the ellipsoidal decompositions returned by MULTINEST. The points given as input are overlaid on the resulting ellipsoids. Here 1000 points were sampled uniformly from: (a) two non-intersecting ellipsoids; and (b) a torus.

One particular problem with rejection sampling schemes is the exponential reduction in sampling efficiency with increasing dimensionality of the problem. In order to address this issue, a slice sampling method has been employed to draw unbiased samples efficiently from the likelihood-constrained prior in the POLYCHORD (Handley et al. 2015a,b) package.

Another algorithm to increase the efficiency of Nested Sampling through the variable number of live points is the "Dynamic Nested Sampling" method (Higson et al. 2018) which has been used in the DYNESTY (Speagle 2019) package.

3.5. MULTINEST *algorithm*

The MULTINEST algorithm (Feroz and Hobson 2008; Feroz et al. 2009) addresses this problem of drawing unbiased samples from the likelihood-constrained prior, through an ellipsoidal rejection sampling scheme. At each iteration, i, the full set of $N_{\rm live}$ live points is enclosed within a set of (possibly overlapping) ellipsoids and the desired replacement point sought from within their union. The ellipsoidal decomposition of the live point set is performed through an expectation-minimisation algorithm such that the sum of volumes of the ellipsoids is minimised with the additional constraint that the total volume enclosed by the ellipsoids is at least X_i/f. Again $X_i \approx \exp(-i/N_{\rm live})$ is the expected prior volume, while $0 < f \leq 1$ is a user defined value for the target efficiency (the ratio of points accepted to points sampled). Thus, f is analogous to the (inverse of the) "enlargement factor" introduced by Mukherjee et al. (2006) into their pioneering ellipsoid-based NS code; the larger the target f the faster the algorithm runs, but the greater the chance of some ellipsoids failing to cover the full $\mathcal{L} > \mathcal{L}_i$ volume (biasing the vanilla NS estimates, though not necessarily the INS estimates, as we discuss later).

The MULTINEST ellipsoidal decomposition algorithm thus allows substantial flexibility in the geometry of its posterior exploration; with bent and/or irregularly-shaped posterior modes typically broken into a relatively large number of small 'overlapping' ellipsoids and smooth, near-Gaussian posterior modes kept whole (or broken into relatively few ellipsoids), as shown in Fig. 2. It thereby automatically accommodates elongated, curving degeneracies while maintaining high efficiency for simpler problems. MULTINEST also specifically enables the identification of distinct modes by isolating non-overlapping subsets of the ellipsoidal decomposition; so identified, these distinct modes can then be evolved independently.

Once the ellipsoidal bounds have been created at a given iteration of the MULTINEST algorithm a new point is drawn uniformly from the union of these ellipsoids as follows. If there are L ellipsoids at iteration i, a particular ellipsoid is chosen with probability p_l given as:

$$p_l = V_l / V_{\rm tot}, \tag{11}$$

where $V_{\rm tot} = \sum_{l=1}^{L} V_l$, from which a single point is then drawn uniformly and checked against the constraint $\mathcal{L} > \mathcal{L}_i$. If satisfied the point is accepted with probability $1/q$, where q is the number of ellipsoids the new point lies in (in order to take into account the possibility of non-empty intersections), otherwise it is rejected (but saved for INS summation) and the process is repeated with a new random choice of ellipsoid.

In higher dimensions, most of the volume of an ellipsoid lies in its outer shells and therefore any overshoot of the ellipsoidal decomposition relative to the true iso-likelihood surface can result in a marked drop in sampling efficiency. In order to maintain the sampling efficiency for such high dimensional problems, MULTINEST can also operate in a 'constant efficiency mode'. In this mode, the total volume enclosed by the ellipsoids is no longer linked with the expected prior volume X_i by requiring the total ellipsoidal volume to be at least X_i/f, instead the total volume enclosed by the union of ellipsoids is adjusted such that the sampling efficiency is as close to the user defined target efficiency f as possible while keeping every live point enclosed in at least one ellipsoid. Despite the increased chance of the fitted ellipsoids encroaching *within* the constrained-likelihood volume (i.e., missing regions of parameter space for which $\mathcal{L} > \mathcal{L}_i$), past experience has shown (e.g. Feroz et al. 2009) this constant efficiency mode may nevertheless produce reasonably accurate posterior distributions for parameter estimation purposes. The vanilla NS evidence values, however, cannot be relied upon in this mode, with the expectation being a systematic over-estimation of the model evidence. Interestingly, the same is not *strictly* true of the INS evidence estimates, which use the NS technique only for posterior exploration (not evidence summation); though we will note later some important caveats for its error estimation.

In the rest of this paper, we refer to the mode in which MULTINEST links the volume of the ellipsoidal decomposition with the expected prior volume as its 'default' mode, and we specifically highlight instances where 'constant efficiency mode' has been trialled.

4. IMPORTANCE NESTED SAMPLING

Though highly efficient in its approximation to the iso-likelihood contour bounding the live particle set at each iteration, the ellipsoidal rejection sampling scheme used by MULTINEST ultimately discards a significant pool of sampled points failing to satisfy the NS constraint, $\mathcal{L} > \mathcal{L}_i$, for which the likelihood has nevertheless been evaluated at some computational cost. In order to redress this final inefficiency the technique of importance nested sampling (INS) has recently been proposed by Cameron and Pettitt (2013) as an alternative summation for the Bayesian evidence in this context. In particular, INS uses all the points drawn by MULTINEST, or any other ellipsoidal rejection sampling algorithm, at each iteration regardless of whether they satisfy the constraint $\mathcal{L} > \mathcal{L}_i$ or not. The relationship of INS to existing MC schemes is summarised in Appendix A.

4.1. *Pseudo-importance sampling density*

One begins by defining the following pseudo-importance sampling density:

$$g(\boldsymbol{\Theta}) = \frac{1}{N_{\rm tot}} \sum_{i=1}^{N_{\rm iter}} \frac{n_i E_i(\boldsymbol{\Theta})}{V_{{\rm tot},i}}, \tag{12}$$

where $N_{\rm iter}$ is the total number of iterations (ellipsoidal decompositions) performed by MULTINEST, n_i the number of points collected at the $i^{\rm th}$ iteration (with total, $N_{\rm tot} = \sum_{i=1}^{N_{\rm iter}} n_i$), $V_{{\rm tot},i}$ the total volume enclosed in the union of ellipsoids at the $i^{\rm th}$ iteration, and $E_i(\boldsymbol{\Theta})$ an indicator function returning 1 when $\boldsymbol{\Theta}$ lies in the $i^{\rm th}$ ellipsoidal decomposition and 0 otherwise. We call $g(\boldsymbol{\Theta})$ here a *pseudo*-importance sampling density since it is of course defined only *a posteriori* to our sampling from it, with the consequence that all $\boldsymbol{\Theta} \sim E_{j>i}(\boldsymbol{\Theta})$ are to some (ideally negligible) extent dependent on all previous $\boldsymbol{\Theta} \sim E_{j\leq i}(\boldsymbol{\Theta})$ (some important implications of which we discuss in Sec. 4.3 below). The heritage of this technique lies with the reverse logistic regression strategy of Geyer (1994) and the "biased sampling" framework of Vardi (1985). Another term that has been used in place of pseudo-importance sampling is "recycling of past draws" (e.g. Cornuet et al. 2012).

If at each iteration, the ellipsoidal decomposition would consist of only one ellipsoid then $V_{{\rm tot},i}$ is simply the geometric volume of the ellipsoid at iteration i. MULTINEST, however, may enclose its live points in a set of possibly overlapping ellipsoids. An analytical expression for calculating the volume in the overlapped region of ellipsoids is not available and therefore we estimate the volume occupied by the union of ellipsoids through the following MC method. Whenever an ellipsoidal decomposition is constructed, we draw M points ($\boldsymbol{\Theta}'_m$, $m = 1, 2, \ldots, M$) from it as follows: for each draw we first pick an ellipsoid with probability $V_l / \sum_{l=1}^{L} V_l$, where V_l are the volumes of the L ellipsoids in the decomposition; a point $\boldsymbol{\Theta}'_m$ is then drawn uniformly from the chosen ellipsoid and we calculate, q_m, the number of ellipsoids it lies in. The volume in the union of ellipsoids is then:

$$V_{\rm tot} \approx \hat{V}_{\rm tot} = \frac{M}{\sum_{m=1}^{M} q_m} \sum_{l=1}^{L} V_l. \tag{13}$$

We note that this Monte Carlo procedure does not require any evaluations of the likelihood function, and thus is not computationally demanding.

4.2. *Evidence estimation and posterior samples*

As an alternative to the canonical NS summation given by Eq. (8) the Bayesian evidence can instead be estimated with reference to the above pseudo-importance sampling density as:

$$\hat{\mathcal{Z}} = \frac{1}{N_{\rm tot}} \sum_{k=1}^{N_{\rm tot}} \frac{\mathcal{L}(\boldsymbol{\Theta}_k)\pi(\boldsymbol{\Theta}_k)}{g(\boldsymbol{\Theta}_k)}. \tag{14}$$

Moreover, each one of the $N_{\rm tot}$ points collected by MULTINEST can be assigned the following estimator of its posterior probability density:

$$P(\boldsymbol{\Theta}) = \frac{\mathcal{L}(\boldsymbol{\Theta})\pi(\boldsymbol{\Theta})}{N_{\rm tot} g(\boldsymbol{\Theta})}. \tag{15}$$

Since the importance nested sampling scheme does not rely on the ellipsoidal decomposition fully enclosing the region(s) satisfying the constraint $\mathcal{L} > \mathcal{L}_i$, it can also achieve accurate evidence estimates and posterior summaries from sampling done in the constant efficiency mode of MULTINEST.[3] However, as we discuss shortly, the utility of this feature is often limited by ensuing difficulties in the estimation of uncertainty for such constant efficiency mode evidence estimates.

From a computational perspective we note that in a naïve application of this scheme it will be necessary to store $N_{\rm tot}$ points, $\boldsymbol{\Theta}_k$, along with the likelihood, $\mathcal{L}(\boldsymbol{\Theta}_k)$, and prior probability, $\pi(\boldsymbol{\Theta}_k)$, for each, *as well as* all relevant information describing the ellipsoidal decompositions (centroids, eigen-values and eigen-vectors) at each iteration. Even with a Cholesky factorization of

[3] The reasons for this are described in detail in Appendix B; but in brief we note that like 'ordinary' importance sampling the only fundamental constraint on the $g(\boldsymbol{\Theta})$ of the INS scheme is that its support enclose that of the posterior, which we ensure by drawing our first set of points from the prior support itself, $E_1(\boldsymbol{\Theta}) = 1$ whenever $\pi(\boldsymbol{\Theta}) > 0$.

the eigen-vectors, storing the latter may easily result in excessive memory requirements. However, since in the MULTINEST algorithm the prior volume, and consequently the volume occupied by the bounding ellipsoids, shrinks at each subsequent iteration one can confidently assume $E_i(\mathbf{\Theta}) = 1$ for all points drawn at iterations $j > i$. At a given iteration then, one needs only to check if points collected from previous iterations lie in the current ellipsoidal decomposition and add the contribution to $g(\mathbf{\Theta})$ coming from the current iteration as given in Eq. (12). This results in an enormous reduction in memory requirements as information about the ellipsoidal decomposition from previous iterations no longer needs to be stored.

At each iteration, MULTINEST draws points from the ellipsoidal decomposition, but in order to take account of the volume in the overlaps between ellipsoids, each point is accepted only with probability $1/q$ where q is the number of ellipsoids in which the given point lies. Rather than discarding all these rejected points, which would be wasteful, we include them by dividing the importance sampling weights as given in Eq. (12), in three components:

$$g(\mathbf{\Theta}) = g_1(\mathbf{\Theta}) + g_2(\mathbf{\Theta}) + g_3(\mathbf{\Theta}). \tag{16}$$

Assuming that the point $\mathbf{\Theta}$ was drawn at iteration i, g_1, g_2 and g_3 are the contributions to importance weight for $\mathbf{\Theta}$ coming from iteration i, iterations before i and iterations after i respectively. Thus, g_1 is calculated as follows:

$$g_1(\mathbf{\Theta}) = \frac{q n_i}{N_{\rm tot} V_{{\rm tot},i}}, \tag{17}$$

where q is the number of ellipsoids at iteration i in which point $\mathbf{\Theta}$ lies, while g_2 is calculated as follows:

$$g_2(\mathbf{\Theta}) = \frac{1}{N_{\rm tot}} \sum_{j=1}^{i-1} \frac{n_j}{V_{\rm tot}, j}, \tag{18}$$

where $V_{{\rm tot},j}$ is volume occupied by the union of ellipsoids at iteration j as given in Eq. (13). Here we have assumed that ellipsoids shrink at subsequent iterations and therefore points drawn at iteration i lie inside the ellipsoidal decompositions of previous iterations as discussed earlier. Finally, g_3 is calculated as follows:

$$g_3(\mathbf{\Theta}) = \frac{1}{N_{\rm tot}} \sum_{j=i+1}^{n_{\rm iter}} \frac{n_j E_j(\mathbf{\Theta})}{V_{{\rm tot},j}}. \tag{19}$$

4.3. *Evidence error estimation*

As discussed by Skilling (2004) (and by Feroz and Hobson 2008; Feroz et al. 2009 for the specific case of MULTINEST) repeated summation of the NS draws under random sampling of the associated X_i (governed by t_i; cf. Sec. 3) allows one to estimate the error on the NS evidence approximation from just a single run (whereas many other MC integration techniques, such as thermodynamic integration, require repeat runs to achieve this). Provided that the parameter space has been explored with sufficient thoroughness (i.e., the $N_{\rm live}$ point set has evolved through all the significant posterior modes), the reliability of this evidence estimate was demonstrated in Feroz and Hobson (2008). Importantly, such a single run error estimate can also be calculated for the INS scheme as described below.

Under ordinary (as opposed to pseudo-) importance sampling the unbiased estimator for the asymptotic variance of the evidence estimate here, $\widehat{\rm Var}[\hat{\mathcal{Z}}]$, would be given as follows:

$$\widehat{\rm Var}[\hat{\mathcal{Z}}] = \frac{1}{N_{\rm tot}(N_{\rm tot}-1)} \sum_{k=1}^{N_{\rm tot}} \left[\frac{\mathcal{L}(\mathbf{\Theta}_k)\pi(\mathbf{\Theta}_k)}{g(\mathbf{\Theta}_k)} - \hat{\mathcal{Z}} \right]^2, \tag{20}$$

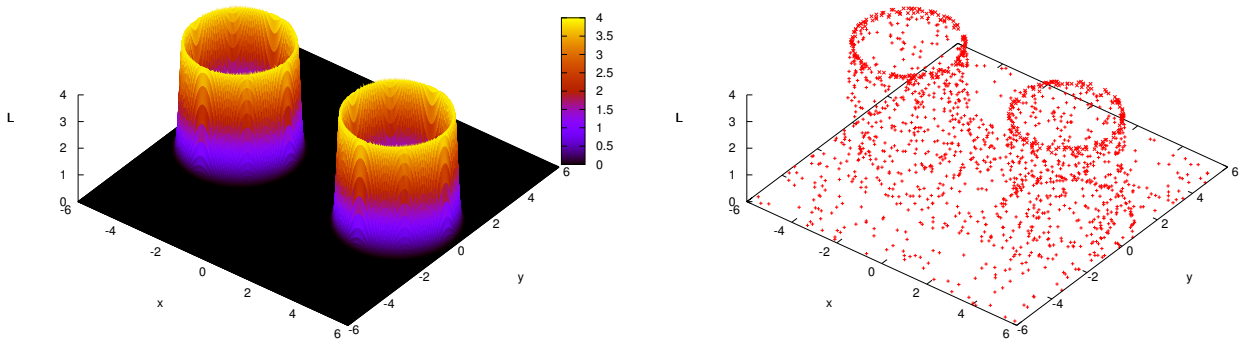
with $\hat{\mathcal{Z}}$ given by Eq. (14).

With the draws from MULTINEST representing our *a posteriori* constructed $g(\mathbf{\Theta})$ not in fact an independent, identically distributed sequence from this pseudo-importance sampling function, the above uncertainty estimate is, unfortunately, not strictly applicable here. In particular, with the placement of subsequent ellipses, $E_{j>i}$, dependent on the position of the live particles drawn up to the present step, i, so too are the subsequently drawn $\mathbf{\Theta}_{j>i}$. However, when MULTINEST is run in its default mode, such that we strongly govern the maximum rate at which the volume of the successive E_i can shrink we can be confident that our sampling becomes ever more nearly independent and that the dominant variance component is indeed given in Eq. (20). Our reasoning behind this is explained in detail in Appendix B. On the other hand, when MULTINEST is being run in 'constant efficiency mode' we recommended for the user to check (via repeat simulation) that the INS evidence is stable (with respect to its error estimate) for reasonable variation in $N_{\rm live}$ and/or f.

5. APPLICATIONS

In this section we apply the MULTINEST algorithm with INS described above to three test problems to demonstrate that it indeed calculates the Bayesian evidence much more accurately than vanilla NS. These test examples are chosen to have features that resemble those that can occur in real inference problems in astro- and particle physics.

5.1. *Test problem 1: Gaussian shells likelihood*



(a) (b)

FIG. 3.— Test problem 1: (a) two-dimensional plot of the likelihood function defined in Eqs. (21) and (22); (b) dots denoting the points with the lowest likelihood at successive iterations of the MULTINEST algorithm.

D	$N_{\rm live}$	f	$N_{\rm like}$ default	$N_{\rm like}$ ceff
2	300	0.30	4,581	3,871
5	300	0.30	8,922	7,882
10	300	0.05	73,342	76,255
20	300	0.05	219,145	163,234
30	500	0.05	604,906	548,501
50	500	0.01	10,531,223	5,290,550

TABLE 1
DIMENSIONALITY (D) OF PROBLEM, NUMBER OF LIVE POINTS ($N_{\rm live}$), TARGET EFFICIENCY (f) AND THE TOTAL NUMBER OF LIKELIHOOD EVALUATIONS ($N_{\rm like}$) IN DEFAULT AND CONSTANT EFFICIENCY (CEFF) MODES OF MULTINEST FOR TEST PROBLEM 1, DISCUSSED IN SEC. 5.1.

In this section, we apply MULTINEST with and without INS to sample from a posterior containing multiple modes with pronounced (curving) degeneracies in relatively high dimensions. Our test problem here is the same one used in Allanach and Lester (2008); Feroz and Hobson (2008); Feroz et al. (2009). The likelihood function of this problem is defined as,

$$\mathcal{L}(\theta) = {\rm circ}(\theta; \mathbf{c}_1, r_1, w_1) + {\rm circ}(\theta; \mathbf{c}_2, r_2, w_2), \tag{21}$$

where

$${\rm circ}(\theta; \mathbf{c}, r, w) = \frac{1}{\sqrt{2\pi w^2}} \exp\left[-\frac{(|\theta - \mathbf{c}| - r)^2}{2w^2}\right]. \tag{22}$$

In two dimensions, this distribution represents two well separated rings, centred on the points $\mathbf{c}_1$ and $\mathbf{c}_2$ respectively, each of radius r and with a Gaussian radial profile of width w (see Fig. 3).

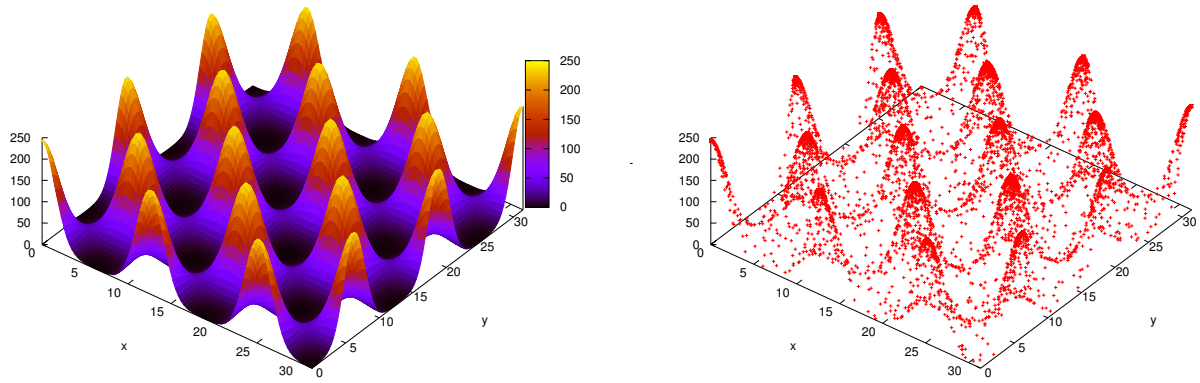
We investigate the above distribution up to a 50-dimensional parameter space θ. In all cases, the centres of the two rings are separated by 7 units in the parameter space, and we take $w_1 = w_2 = 0.1$ and $r_1 = r_2 = 2$. We make r_1 and r_2 equal, since in higher dimensions any slight difference between these two values would result in a vast difference between the volumes occupied by the rings and consequently the ring with the smaller r value would occupy a vanishingly small fraction of the total probability volume, making its detection almost impossible. It should also be noted that setting $w = 0.1$ means the rings have an extremely narrow Gaussian profile. We impose uniform priors $\mathcal{U}(-6, 6)$ on all the parameters. For the two-dimensional case, with the parameters described above, the likelihood is shown in Fig. 3.

Table 1 lists the total number of live points ($N_{\rm live}$) and target efficiency (f) used and the total number of likelihood evaluations ($N_{\rm like}$) performed by MULTINEST in default and constant efficiency (ceff) modes. The volume of the parameter space increases exponentially with the dimensionality D, therefore we need to increase $N_{\rm live}$ and/or decrease f with D, in order to get accurate estimates of $\log(\mathcal{Z})$. The true and estimated values of $\log(\mathcal{Z})$ are listed in Table 2.

It can be seen from Table 2 that $\log(\hat{\mathcal{Z}})$ values obtained by MULTINEST with and without INS and in both default and constant efficiency modes are consistent with the true $\log(\mathcal{Z})$ for $D \leq 20$, the only exception being the $\log(\hat{\mathcal{Z}})$ from constant efficiency mode with INS which is $\sim 6\sigma$ away from the analytical $\log(\mathcal{Z})$. We attribute this to the heightened potential for underestimation of the INS uncertainties in constant efficiency mode discussed in Sec. 4 and Appendix B. For $D \geq 30$ however, the $\log(\hat{\mathcal{Z}})$ values obtained by MULTINEST without INS start to become inaccurate, with constant efficiency mode again giving more inaccurate results as expected. These inaccuracies are caused by inadequate numbers of live points used to cover the region satisfying the constraint $\mathcal{L} > \mathcal{L}_i$ at each iteration i. However, with the same values for $N_{\rm live}$ and f, and indeed with the same set of points,

D	Analytical	MULTINEST without INS		MULTINEST with INS	
		default	ceff	default	ceff
2	-1.75	-1.61 ± 0.09	-1.71 ± 0.09	-1.72 ± 0.02	-1.69 ± 0.02
5	-5.67	-5.42 ± 0.15	-5.78 ± 0.15	-5.67 ± 0.03	-5.87 ± 0.03
10	-14.59	-14.55 ± 0.23	-14.83 ± 0.23	-14.60 ± 0.03	-14.58 ± 0.03
20	-36.09	-35.90 ± 0.35	-35.99 ± 0.35	-36.11 ± 0.03	-36.06 ± 0.03
30	-60.13	-59.72 ± 0.35	-59.43 ± 0.34	-60.09 ± 0.02	-59.90 ± 0.02
50	-112.42	-110.69 ± 0.47	-108.96 ± 0.46	-112.37 ± 0.01	-112.18 ± 0.01

TABLE 2
THE TRUE AND ESTIMATED $\log(\mathcal{Z})$ FOR TEST PROBLEM 1, DISCUSSED IN SEC. 5.1, AS A FUNCTION OF THE DIMENSIONS D OF THE PARAMETER SPACE, USING MULTINEST WITH AND WITHOUT INS AND IN ITS DEFAULT AND CONSTANT EFFICIENCY MODES.



(a) (b)

FIG. 4.— Test problem 2: (a) two-dimensional plot of the likelihood function defined in Eq. 23; (b) dots denoting the points with the lowest likelihood at successive iterations of the MULTINEST algorithm.

INS returns $\log(\hat{\mathcal{Z}})$ values which are consistent with the true $\log(\mathcal{Z})$ in default MULTINEST mode and off by at most ~ 0.2 in the constant efficiency mode. The error estimate on $\log(\hat{\mathcal{Z}})$ from INS in the constant efficiency mode might indicate that $\log(\hat{\mathcal{Z}})$ has a large bias but as discussed in Sec. 4.3, these error estimates are reliable only when the importance sampling distribution is guaranteed to give non-vanishing probabilities for all regions of parameter space where posterior distribution has a non-vanishing probability as well. This is much more difficult to accomplish in a 50D parameter space. In addition to this, the approximations we have made to calculate the volume in the overlapped region of ellipsoids are expected to be less accurate in higher dimensions. Therefore, it is very encouraging that INS can obtain $\log(\hat{\mathcal{Z}})$ to within 0.2 units for a very challenging 50D problem with just 500 live points. We should also notice the number of likelihood evaluations in constant efficiency mode starts to become significantly smaller than in the default mode for $D \geq 20$.

5.2. *Test problem 2: egg-box likelihood*

We now demonstrate the application of MULTINEST to a highly multimodal two-dimensional problem, for which the likelihood resembles an egg-box. The un-normalized likelihood is defined as:

$$\mathcal{L}(x, y) = \exp\left\{\left[2 + \cos\left(\frac{x}{2}\right)\cos\left(\frac{y}{2}\right)\right]^5\right\}, \tag{23}$$

and we assume a uniform prior $\mathcal{U}(0, 10\pi)$ for both x and y.

A plot of the log-likelihood is shown in Fig. 4 and the prior ranges are chosen such that some of the modes are truncated, making it a challenging problem for identifying all the modes as well as to calculate the evidence accurately. The true value of the log-evidence is $\log Z = 235.856$, obtained by numerical integration on a very fine grid, which is feasible for this simple two-dimensional example.

It was shown in Feroz et al. (2009) that MULTINEST can explore the parameter space of this problem efficiently, and also calculate the evidence accurately. Here we demonstrate the accuracy of the evidence obtained with MULTINEST using the INS summation. For low-dimensional problems, results obtained with the constant efficiency mode of MULTINEST agree very well with the ones obtained with the default mode, we therefore only discuss the default mode results in this section.

We use 1000 live points with target efficiency $f = 0.5$. The results obtained with MULTINEST are illustrated in Fig. 4, in which the dots show the points with the lowest likelihood at successive iterations of the nested sampling process. MULTINEST required $\sim 20,000$ likelihood evaluations and obtained $\log(\hat{\mathcal{Z}}) = 235.837 \pm 0.008$ (235.848 ± 0.078) with (without) INS, which

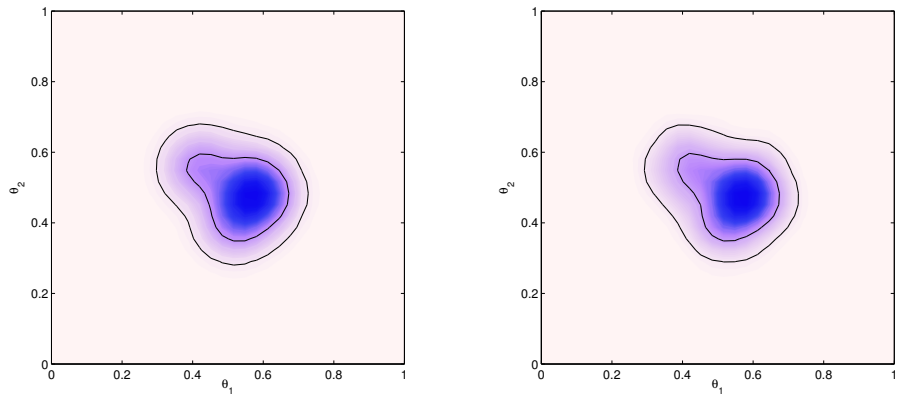


FIG. 5.— Test problem 3: Marginalized posterior distribution in the first 2 dimensions of the 16D Gaussian mixture model discussed in Sec. 5.3. Panel (a) shows the analytical distribution while panel (b) shows the distribution obtained from MULTINEST. The contours represent the 68% and 95% Bayesian credible regions.

compares favourably with the true value given above. In each case, the random number seed was the same, so the points sampled by MULTINEST were identical with and without INS. In order to check if the error estimates on $\log(\hat{\mathcal{Z}})$ are accurate, we ran 10 instances of MULTINEST in both cases, each with a different seed and found the mean and standard deviation of $\log(\hat{\mathcal{Z}})$ to be 235.835 ± 0.009 (235.839 ± 0.063) with (without) INS. In both cases, the standard error agrees with the error estimate from just a single run. There is, however, some indication of bias in the $\log(\hat{\mathcal{Z}})$ value evaluated with INS, which lies $\sim 2\sigma$ away from the true value. This is most likely due to the approximations used in calculating the volume in the overlapped region of ellipsoids, as discussed in Sec. 4. Nonetheless, the absolute value of the bias is very low (~ 0.02), particularly compared with the accuracy (~ 0.5) to which log-evidence values are usually required in practical applications.

5.3. *Test problem 3: 16D Gaussian mixture model*

Our next test problem is the same as test problem 4 in Weinberg et al. (2013) which is a mixture model of four randomly-oriented Gaussian distributions with their centers uniformly selected from the hypercube $[0.5 - 2\sigma, 0.5 + 2\sigma]^D$ with D being the dimensionality of the problem and the variance σ^2 of all four Gaussians is set to 0.003. Weights of the Gaussians are distributed according to a Dirichlet distribution with shape parameter $\alpha = 1$. We impose uniform priors $\mathcal{U}(0, 1)$ on all the parameters. The analytical posterior distribution for this problem, marginalized in the first two dimensions is shown in Fig. 5(a).

The analytical value of $\log(\mathcal{Z})$ for this problem is 0, regardless of D. We set $D = 16$ and used 300 live points with target efficiency $f = 0.05$. The marginalized posterior distribution in the first two dimensions, obtained with the default mode of MULTINEST with INS is shown in Fig. 5(b). The posterior distribution obtained from the constant efficiency mode is identical to the one obtained from the default and therefore we do not show it. In the default mode MULTINEST performed $208,978$ likelihood evaluations and returned $\log(\hat{\mathcal{Z}}) = -0.03 \pm 0.01$ (0.39 ± 0.27) with (without) INS. In the constant efficiency mode, $158,219$ likelihood evaluations were performed and $\log(\hat{\mathcal{Z}}) = 0.21 \pm 0.01$ (0.25 ± 0.27) with (without) INS.

5.4. *Test problem 4: 20D Gaussian-LogGamma mixture model*

Our final test problem is the same as test problem 2 in Beaujean and Caldwell (2013), in which the likelihood is a mixture model consisting of four identical modes, each of which is a product of an equal number of Gaussian and LogGamma 1D distributions, centred at $\theta_1 = \pm 10$, $\theta_2 = \pm 10$, $\theta_3 = \theta_4 = \cdots = \theta_D = 0$ in the hypercube $\theta \in [-30, 30]^D$, where D is the (even) dimensionality of the parameter space. Each Gaussian distribution has unit variance. The LogGamma distribution is asymmetric and heavy-tailed; its scale and shape parameters are both set to unity. We impose uniform priors $\mathcal{U}(-30, 30)$ on all the parameters. The analytical marginalised posterior distribution in the subspace (θ_1, θ_2) is shown in Fig. 6(a).

We set $D = 20$, for which the analytical value of the log-evidence is $\log(\mathcal{Z}) = \log(60^{-20}) = -81.887$. To be consistent with test problem 3, which is of similar dimensionality, we again used 300 live points with a target efficiency $f = 0.05$ (note these values differ from those used in Beaujean and Caldwell (2013), who set $N_{\rm live} = 1000$ and $f = 0.3$ in the standard vanilla NS version of MULTINEST). The marginalized posterior in the first two dimensions, obtained in the default mode of MULTINEST with INS is shown in Fig. 6(b), and is identical to the corresponding analytical distribution, recovering all four modes with very close to equal weights. The posterior distribution obtained from the constant efficiency mode is identical to the one obtained from the default and therefore we do not show it. In the default mode MULTINEST performed 2,786,538 likelihood evaluations and returned $\log(\hat{\mathcal{Z}}) = -81.958 \pm 0.008$ (-78.836 ± 0.398) with (without) INS. In both cases, we see that, for this more challenging problem containing multi-dimensional heavy-tailed distributions, the log-evidence estimates are substantially biased, with each being $\sim 8\sigma$ from the true value. Nonetheless, we note that the estimate using INS is much more accurate than that obtained with vanilla NS, and differs from the true value by only ~ 0.1 units, which is much smaller than the accuracy required in most practical applications. As one might expect, however, the log-evidence estimates obtained in constant

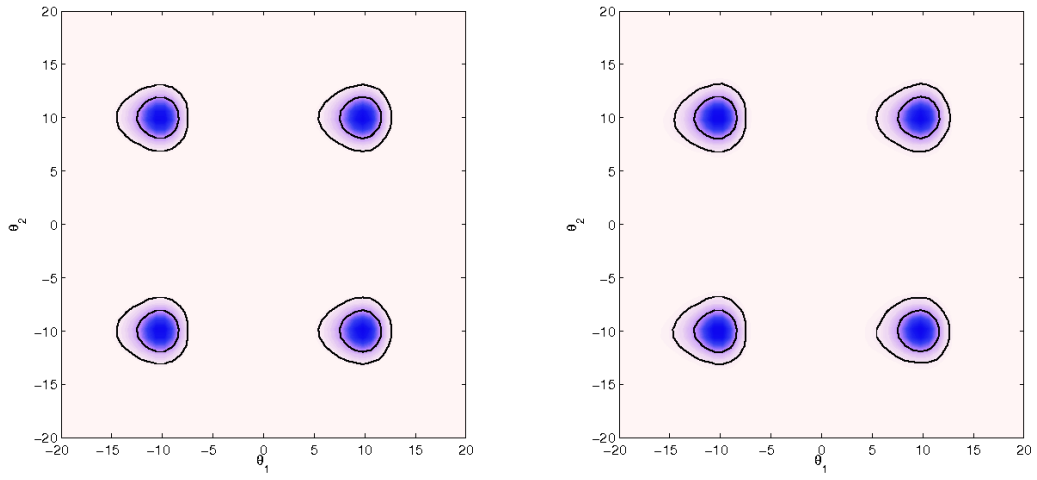


FIG. 6.— Test problem 4: Marginalized posterior distribution in the first 2 dimensions of the 20D Gaussian-LogGamma mixture model discussed in Sec. 5.4. Panel (a) shows the analytical distribution while panel (b) shows the distribution obtained from MULTINEST. The contours represent the 68% and 95% Bayesian credible regions.

efficiency mode are somewhat poorer and show a significant bias. In this mode, $297,513$ likelihood evaluations were performed and $\log(\hat{\mathcal{Z}}) = -82.383 \pm 0.010$ (-71.635 ± 0.376) with (without) INS.

6. SUMMARY AND DISCUSSION

With the availability of vast amounts of high quality data, statistical inference is increasingly playing an important role in cosmology and astroparticle physics. MCMC techniques and more recently algorithms based on nested sampling have been employed successfully in a variety of different areas. The MULTINEST algorithm in particular has received much attention in astrophysics, cosmology and particle physics owing to its ability to efficiently explore challenging multi-modal distributions as well as to calculate the Bayesian evidence.

In this paper we have discussed further development of the MULTINEST algorithm, based on the implementation of the INS scheme recently proposed by Cameron and Pettitt (2013). INS requires no change in the way MULTINEST explores the parameter space, but can calculate the Bayesian evidence at up to an order-of-magnitude higher accuracy than vanilla nested sampling. Moreover, INS also provides a means to obtain reasonably accurate evidence estimates from the constant efficiency mode of MULTINEST. This is particularly important, as the constant efficiency mode enables MULTINEST to explore higher-dimensional spaces (up to $\sim 50D$) much more efficiently than the default mode. Higher evidence accuracy from INS could potentially allow users to use fewer live points $N_{\rm live}$ or higher target efficiency f to achieve the same level of accuracy as vanilla nested sampling, and therefore to speed-up the analysis by several factors. We recommend that users should always check that their posterior distributions are stable with reasonable variation of $N_{\rm live}$ and f. A slight drawback of INS is increased memory requirements. As the importance sampling distributions given in Eqs. 18 and 19 change for every point at each iteration, all the points need to be saved in memory. However, with $N_{\rm live} \leq 1000$ the increased memory requirements should be manageable on most modern computers.

Finally, we give some recommendations for setting the number of live points $N_{\rm live}$ and target efficiency f, which determine the accuracy and computational cost of running the MULTINEST algorithm, with or without INS. Generally, the larger the $N_{\rm live}$ and lower the f, the more accurate are the posteriors and evidence values but the higher the computational cost. For multi-modal problems, $N_{\rm live}$ is particularly important as it determines the effective sampling resolution. If it is too small, certain modes, in particular the ones occupying a very small prior mass, can be missed. Experience has shown that the accuracy of evidence is more sensitive to f than $N_{\rm live}$. In general, for problems where accuracy of evidence is paramount, we suggest f to be no larger than 0.3 in the 'default' mode. In 'constant efficiency mode', we suggest f to be no larger than 0.1 in all cases. Generally, a value of $N_{\rm live}$ in lower hundreds is sufficient. For very low dimensional problems $N_{\rm live}$ can even be in tens. However, for highly multi-modal problems, one may need to set $N_{\rm live}$ to be in a few thousands. It is always advisable to increase $N_{\rm live}$ and reduce f to check if the posteriors and evidence values are stable as function of $N_{\rm live}$ and f.

ACKNOWLEDGEMENTS

This work was performed on COSMOS VIII, an SGI Altix UV1000 supercomputer, funded by SGI/Intel, HEFCE and PPARC, and the authors thank Andrey Kaliazin for assistance. The work also utilized 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. FF is supported by a Research Fellowship from the Leverhulme and Newton Trusts. EC is supported by an Australian Research Council (ARC) Discovery Grant. ANP is supported by an ARC Professorial Fellowship.

A. RELATION OF INS TO EXISTING MONTE CARLO SCHEMES

We review here the heritage of INS amongst the wider family of pseudo-importance sampling, NS, and adaptive Monte Carlo algorithms, for which limited convergence proofs have yet been achieved.

As described in Cameron and Pettitt (2013) the initial idea for INS arose from the study of recursive marginal likelihood estimators, as characterised by Reverse Logistic Regression (RLR; Geyer 1994; Chen and Shao 1997; Kong et al. 2003) and the Density of States (DoS; Habeck 2012; Tan et al. 2012). In these (*equivalent*; cf. Cameron and Pettitt 2013) schemes, the marginal likelihood is sought by pooling (or 'losing the labels' on) a series of draws from a pre-specified set of largely unnormalised importance sampling densities, bridging (at least crudely) the prior and posterior; after this a maximum-likelihood-based estimator is used to infer the relative normalisation of each bridging density in light of the 'missing' label information. As emphasised by Kong et al. (2003), these recursive algorithms may, in turn, be seen as deriving from the 'biased sampling' results of Vardi (1985) and collaborators (e.g. Gill et al. 1988), who give consistency and Central Limit Theorem proofs for this deterministic (i.e., non-adaptive), *pseudo-importance sampling* procedure under simple connectedness/non-separability conditions for the supports of the bridging sequence.

Developed in parallel to the recursive estimators described above, the Deterministic Multiple Mixture Sampling scheme (DMMS; Veach and Guibas 1995; Owen and Zhou 2000) applies much the same strategy, but for a given sequence of strictly *normalised* importance sampling proposal densities; hence, the motivation for 'losing the labels' here becomes simply the reduction of variance in the resulting estimator with respect to that achievable by allocating the same number of draws to ordinary importance sampling from each proposal separately. [We will discuss further the limiting variance of a simple 'losing the labels' estimator, as relevant to INS, in Appendix B.] At the expense of introducing an *intractable* (but asymptotically-diminishing) *bias*, Cornuet et al. (2012) have recently constructed a yet more efficient extension called Adaptive Multiple Importance Sampling (AMIS) in which the sequence of importance sampling proposal densities is refined adaptively towards the target at runtime. In particular, each proposal density after the first is chosen (from a given parametric family) in a manner dependent on the weighted empirical distribution of draws from all previous densities. As suggested by their given numerical examples this approach appears superior to other adaptive importance sampling schemes (e.g. the cross-entropy method; cf. Rubinstein and Kroese 2004) in which the past draws from sub-optimal proposals are ultimately discarded.

In our opinion, despite its genesis in the study of RLR/DoS, INS may perhaps most accurately be viewed as a 'descendent' of this AMIS methodology; the key difference being that INS builds an efficient mixture proposal density for marginal likelihood estimation via the NS pathway (Sec. 3), whereas AMIS aims to iterate towards such a form chosen from within a pre-specified family of proposal densities. In other words, in INS the proposal densities represented by our sequence of ellipsoidal decompositions should share (by design) a near-equal 'importance' in the final mixture, while in AMIS those draws from the earlier proposals are expected to become increasingly insignificant as the later proposals achieve refinement towards their target.

As acknowledged by Cornuet et al. (2012), the inherent dependence structure of the pseudo-importance weighted terms entering the final AMIS summation—owing entirely in this approach to the dependence structure of the corresponding sequence of proposal densities—renders intractable the demonstration of a *general* consistency for the algorithm. Indeed even the elegant and detailed solution presented by Marin et al. (2012) is reliant on key modifications to the basic procedure, including that the number of points drawn from each successive density grows significantly at each iteration (incompatible with INS; and seemingly at odds too with Cornuet et al.'s original recommendation of heavy sampling from the first proposal), as well as numerous assumptions regarding the nature of the target and proposal families. In light of this historical background we will therefore give particular attention in the following Appendix to the variance reduction benefits of the theoretically-problematic 'losing the labels' strategy as employed in INS, before sketching a rough proof of consistency thereafter (albeit under some strong assumptions on the asymptotic behaviour of the EM plus k-means algorithm employed for ellipsoidal decomposition with respect to the target density; which may be difficult, if not impossible, to establish in practice).

Finally, before proceeding it is worth mentioning briefly the heritage of INS with respect to vanilla NS (Skilling 2004, 2006). As described in Sections 3 and 4, the original MULTINEST code (Feroz and Hobson 2008; Feroz et al. 2009) was designed for estimation of $\mathcal{Z}$ via the NS pathway (Skilling 2006) with the challenge of constrained-likelihood sampling tackled via rejection sampling within a series of ellipsoidal decompositions bounding the evolving live point set. In contrast to AMIS (and INS) the convergence properties of the simple NS algorithm are well understood; in particular, Chopin and Robert (2010) have derived a robust CLT for nested sampling, and both Skilling (2006) and Keeton (2011) give insightful discussions. Despite the value of this ready availability of a CLT for vanilla NS, our experience (cf. Sec. 5 of the main text) is that by harnessing the information content of the otherwise-discarded draws from MULTINEST's ellipsoidal rejection sampling the INS summation does ultimately yield in practice a substantially more efficient approximation to the desired marginal likelihood, the reliability of which is moreover adequately estimable via the prescription given subsequently.

B. CONVERGENCE BEHAVIOUR OF INS

In this Appendix we discuss various issues relating to the *asymptotic convergence* of the INS marginal likelihood estimator (14), which we denote here by $\hat{\mathcal{Z}}^{\mathrm{INS}}$, towards the true marginal likelihood, $\mathcal{Z}$, as the sample size (controlled by the size of the live point set, N_{live}) approaches infinity. We begin by considering the intriguing role of pseudo-importance sampling for variance reduction within certain such schemes; this step, ironically, is itself primarily responsible for the intractable bias of the complete algorithm. With this background in mind we can at last outline a heuristic argument for the consistency of INS and consider a break down of its variance into distinct terms of transparent origin.

To be precise, we will investigate here the asymptotic convergence behaviour of the INS estimator with ellipsoidal decompositions almost exactly as implemented in MULTINEST, a detailed description of which is given in the main text (Sections 3 & 4).

For reference, we take

$$\hat{\mathcal{Z}}^{\mathrm{INS}} \equiv \frac{1}{N_{\mathrm{tot}}} \sum_{i=1}^{c\times N_{\mathrm{live}}} \sum_{k=1}^{n_i} \frac{\mathcal{L}(\boldsymbol{\Theta}_k^{(i)})\pi(\boldsymbol{\Theta}_k^{(i)})}{g(\boldsymbol{\Theta}_k^{(i)})},$$
$$= \frac{1}{N_{\mathrm{tot}}} \sum_{k=1}^{N_{\mathrm{tot}}} \frac{\mathcal{L}(\boldsymbol{\Theta}_k)\pi(\boldsymbol{\Theta}_k)}{g(\boldsymbol{\Theta}_k)}, \tag{24}$$

with $N_{\mathrm{tot}} \equiv \sum_{i=1}^{c\times N_{\mathrm{live}}} n_i$, $g(\boldsymbol{\Theta}) \equiv \frac{1}{N_{\mathrm{tot}}} \sum_{i=1}^{c\times N_{\mathrm{live}}} \frac{n_i E_i(\boldsymbol{\Theta})}{V_{\mathrm{tot},i}}$, and $c\times N_{\mathrm{live}}$ a fixed stopping proxy for the total number of ellipsoidal decompositions required to meet our actual flexible stopping criterion (cf. Sec. 3.2). Each collection of $\boldsymbol{\Theta}_{k=1,\ldots,n_i}^{(i)}$ here is assumed drawn uniformly from the corresponding ellipsoidal decomposition (of the live particle set), $E_i(\cdot)$, with volume, $V_{\mathrm{tot},i}$, until the discovery of a single point, say $\boldsymbol{\Theta}_{n_i}^{(i)}$, with $\mathcal{L}(\boldsymbol{\Theta}_{n_i}^{(i)}) > \mathcal{L}_{i-1}$. This new constrained-likelihood point serves, of course, as the replacement to the $\mathcal{L}_{i-1}$ member of the NS live point set against which the next, $E_{i+1}(\cdot)$, decomposition is then defined.

The equality in (24) highlights the fact that having pooled our draws from each $E_i(\cdot)$ into the pseudo-importance sampling function, $g(\cdot)$, we may proceed to 'lose the labels', $^{(i)}$, on these as in, e.g., Reverse Logistic Regression or "biased sampling". Note also that we suppose $E_1(\cdot)$ is fixed to the support of the prior itself (to ensure that the support of $\mathcal{L}(\boldsymbol{\Theta})\pi(\boldsymbol{\Theta})$ is contained within that of $g(\boldsymbol{\Theta})$), and that we must sample an initial collection of N_{live} live points from the prior as well to populate the original live particle set in advance of our first constrained-likelihood exploration.

Finally, we neglect in the ensuing analysis any uncertainty in our $V_{\mathrm{tot},i}$ since, although these are in fact estimated also via (simple MC) simulation, without the need for likelihood function calls in this endeavour we consider the cost of arbitrarily improving their precision effectively negligible.

B.1. *Motivation for 'losing the labels' on a normalised pseudo-importance sampling mixture*

The effectiveness of the so-called 'losing the labels' strategy for marginal likelihood estimation via the recursive pathway can be easily appreciated for the typical RLR/DoS case of multiple unnormalised bridging densities, since by allowing for, e.g., the use of tempered Monte Carlo sampling we immediately alleviate to a large extent the burdens of importance sampling proposal design (cf. Hesterberg 1995). However, its utility in cases of strictly normalised mixtures of proposal densities as encountered in DMMS and INS is perhaps surprising. Owen and Zhou (2000) give a proof that, under the DMMS scheme, the asymptotic variance of the final estimator will not be very much worse than that achievable under ordinary importance sampling from the optimal distribution alone. However, as the INS sequence of ellipsoids is not designed to contain a single optimal proposal, but rather to function 'optimally' as an ensemble we focus here on demonstrating the strict ordering (from largest to smallest) of the asymptotic variance for (I) ordinary importance sampling under each mixture component separately, (II) ordinary importance sampling under the true mixture density itself, and (III) *pseudo*-importance sampling from the mixture density (i.e., 'losing the labels').

Consider a grossly simplified version of INS in which, at the nth iteration (it is more convenient here to use n rather than i as the iteration counter), a single random point is drawn independently from each of n labelled densities, $h_{k,n}(\cdot)$ $(k = 1, 2, \ldots, n)$, with identical supports matching those of the target, $f(\cdot)$. We denote the resulting set of n samples by $\boldsymbol{\Theta}_{k=1,2,\ldots,n}^{(n)}$. The three key simplifications here are: (I) that the draws are independent, when in MULTINEST they are inherently dependent; (II) that the supports of the $h_{k,n}(\cdot)$ match, when in fact the ellipsoidal decompositions, $E_n(\cdot)$, of MULTINEST have generally nested supports (though one *could* modify them appropriately in the manner of defensive sampling; Hesterberg 1995); and (III) that a single point is drawn from each labelled density, when in fact the sampling from each $E_n(\boldsymbol{\Theta})$ under MULTINEST follows a negative binomial distribution for one $E_n(\boldsymbol{\Theta}) \cap \{\mathcal{L}(\boldsymbol{\Theta}) > \mathcal{L}_{n-1}\}$ 'success'. Suppose also now that the unbiased estimator, $\hat{\mathcal{Z}}_k^{(n)} = f(\boldsymbol{\Theta}_k^{(n)})/h_{k,n}(\boldsymbol{\Theta}_k^{(n)})$, for the normalizing constant belonging to $f(\cdot)$, namely $\mathcal{Z} = \int f(\boldsymbol{\Theta})d\boldsymbol{\Theta}$, in such single draw importance sampling from each of the specified $h_{k,n}(\cdot)$ has finite (but non-zero) variance (cf. Hesterberg 1995), i.e.,

$$\sigma_{k,n}^2 = \int \frac{f(\boldsymbol{\Theta})^2}{h_{k,n}(\boldsymbol{\Theta})} d\boldsymbol{\Theta} - \mathcal{Z}^2, \qquad 0 < \sigma_{k,n}^2 < \infty, \tag{25}$$

and that together our $\hat{\mathcal{Z}}_k^{(n)}$ satisfy Lindeberg's condition such that the CLT holds for this triangular array of random variables (cf. Billingsley 1995).

Now, if we would decide to **keep the labels**, k, on our independent draws from the sequence of $h_{k,n}(\cdot)$ then supposing no prior knowledge of any $\sigma_{k,n}^2$ (i.e., no prior knowledge of how close each proposal density might be to our target) the most sensible option might be to take as a 'best guess' for $\mathcal{Z}$ the (unweighted) sample mean of our individual $\hat{\mathcal{Z}}_k^{(n)} = f(\boldsymbol{\Theta}_k^{(n)})/h_{k,n}(\boldsymbol{\Theta}_k^{(n)})$, that is,

$$\hat{\mathcal{Z}}_{\mathrm{labelled}} = \frac{1}{n} \sum_{k=1}^{n} \frac{f(\boldsymbol{\Theta}_k^{(n)})}{h_{k,n}(\boldsymbol{\Theta}_k^{(n)})}. \tag{26}$$

With a common mean and finite variances for each, this sum over a triangular array converges (in distribution) to a univariate

Normal with mean, $\mathcal{Z}$, and variance,

$$\sigma^2_{\text{labelled}} = \frac{s_n^2}{n} = \frac{1}{n}\sum_{k=1}^{n}\left[\int f(\boldsymbol{\Theta})^2/h_{k,n}(\boldsymbol{\Theta})d\boldsymbol{\Theta} - \mathcal{Z}^2\right], \tag{27}$$

here we use the abbreviation, $s_n^2 = \sum_{k=1}^n \sigma_{k,n}^2$.

On the other hand, if we would instead decide to **lose the labels** on our independent draws we might then follow Vardi's method and imagine each $\boldsymbol{\Theta}_k^{(n)}$ to have come from the mixture distribution, $g(\boldsymbol{\Theta}) = \frac{1}{n}\sum_{j=1}^n h_{j,n}(\boldsymbol{\Theta})$, for which the alternative estimator,

$$\hat{\mathcal{Z}}_{\text{unlabelled}} = \frac{1}{n}\sum_{k=1}^{n}\frac{f(\boldsymbol{\Theta}_k^{(n)})}{g(\boldsymbol{\Theta}_k^{(n)})}, \tag{28}$$

may be derived. To see that the $\hat{\mathcal{Z}}_{\text{unlabelled}}$ estimator so defined is in fact unbiased we let $\hat{\mathcal{Z}}_k^{(n)\prime} = f(\boldsymbol{\Theta}^{(n)})/h_{k,n}(\boldsymbol{\Theta}^{(n)})$ for $\boldsymbol{\Theta}^{(n)} \sim h_{k,n}(\cdot)$ and observe that

$$\begin{aligned}
\mathrm{E}[\hat{\mathcal{Z}}_{\text{unlabelled}}] &= \frac{1}{n}\sum_{k=1}^{n}\mathrm{E}[\hat{\mathcal{Z}}_k^{(n)\prime}], \\
&= \frac{1}{n}\sum_{k=1}^{n}\int\frac{f(\boldsymbol{\Theta})h_{k,n}(\boldsymbol{\Theta})}{g(\boldsymbol{\Theta})}d\boldsymbol{\Theta}, \\
&= \int\frac{f(\boldsymbol{\Theta})}{g(\boldsymbol{\Theta})}\left[\frac{1}{n}\sum_{k=1}^{n}h_{k,n}(\boldsymbol{\Theta})\right]d\boldsymbol{\Theta} = \mathcal{Z}.
\end{aligned} \tag{29}$$

For n iid samples drawn faithfully from the mixture density, $g(\cdot)$, we would expect (via the CLT) that the estimator $\hat{\mathcal{Z}}_{\text{unlabelled}}$ will converge (in distribution) once again to a univariate Normal with mean, $\mathcal{Z}$, but with alternative variance, $\sigma^{2\ [g(\cdot),\text{true}]}_{\text{unlabelled}} = \frac{1}{n}\int f(\boldsymbol{\Theta})^2/g(\boldsymbol{\Theta})d\boldsymbol{\Theta} - \mathcal{Z}^2$. However, for the pseudo-importance sampling from $g(\cdot)$ described above, in which we instead pool an explicit sample from each of its separate mixture components, the asymptotic variance of $\hat{\mathcal{Z}}_{\text{unlabelled}}$ is significantly smaller again. In particular,

$$\begin{aligned}
\sigma^{2\ [g(\cdot),\text{pseudo}]}_{\text{unlabelled}} &= \sum_{k=1}^{n}\mathrm{Var}[\hat{\mathcal{Z}}_k^{(n)\prime}/n], \\
&= \frac{1}{n^2}\sum_{k=1}^{n}\{\mathrm{E}[(\hat{\mathcal{Z}}_k^{(n)\prime})^2] - \mathrm{E}[\hat{\mathcal{Z}}_k^{(n)\prime}]^2\}, \\
&= \frac{1}{n^2}\sum_{k=1}^{n}\int\frac{f(\boldsymbol{\Theta})^2}{g(\boldsymbol{\Theta})^2}h_{k,n}(\boldsymbol{\Theta})d\boldsymbol{\Theta} - \frac{1}{n^2}\sum_{k=1}^{n}\mathrm{E}[\hat{\mathcal{Z}}_k^{(n)\prime}]^2, \\
&= \frac{1}{n}\int\frac{f(\boldsymbol{\Theta})^2}{g(\boldsymbol{\Theta})}d\boldsymbol{\Theta} - \frac{1}{n^2}\sum_{i=1}^{n}\{(\mathrm{E}[\hat{\mathcal{Z}}_k^{(n)\prime}] - \mathcal{Z}) + \mathcal{Z}\}^2, \\
&= \sigma^{2\ [g(\cdot),\text{true}]}_{\text{unlabelled}} - \frac{1}{n^2}\sum_{i=1}^{n}\{\mathrm{E}[\hat{\mathcal{Z}}_k^{(n)\prime}] - \mathcal{Z}\}^2, \\
&\leq \sigma^{2\ [g(\cdot),\text{true}]}_{\text{unlabelled}},
\end{aligned} \tag{30}$$

with equality achieved only in the trivial case that all mixture components are identical. The variance reduction here in the pseudo-importance sampling framework relative to the true importance sampling case derives of course from the effective replacement of multinomial sampling of the mixture components by fixed sampling from their expected proportions.

Comparing now the asymptotic variances of our labelled and unlabelled estimators we can see that the latter is (perhaps surprisingly) *always* smaller than the former, i.e.,

$$\sigma^{2\ [g(\cdot),\text{true}]}_{\text{unlabelled}} - \sigma^2_{\text{labelled}} = \frac{1}{n}\int f(\boldsymbol{\Theta})^2\left[\sum_{k=1}^{n}\frac{1}{h_{k,n}(\boldsymbol{\Theta})} - \frac{n}{(\sum_{k=1}^n h_{k,n}(\boldsymbol{\Theta}))}\right]d\boldsymbol{\Theta} < 0, \tag{31}$$

(recalling that all densities here are strictly positive, of course); thus, we observe the ordering,

$$\sigma^{2\ [g(\cdot),\text{pseudo}]}_{\text{unlabelled}} < \sigma^{2\ [g(\cdot),\text{true}]}_{\text{unlabelled}} < \sigma^2_{\text{labelled}}.$$

This is, as has been remarked in the past, the paradox of the 'losing the labels' idea; that *by throwing away information about our sampling process we appear to* ***gain*** *information about our target*! In fact, however, all we are really doing by choosing to estimate $\mathcal{Z}$ with $\hat{\mathcal{Z}}_{\text{unlabelled}}$ rather than $\hat{\mathcal{Z}}_{\text{labelled}}$ is to use the information we have extracted from $f(\cdot)$ in a more efficient manner, as understood (from the above error analysis) *a priori* to our actual importance sampling. The strict ordering shown here explains why we have selected a pseudo-importance sampling strategy for combining the ellipsoidal draws in MULTINEST as opposed to, e.g., modifying our sampling from the $E_n(\cdot)$ to match (defensively) the support of $\pi(\boldsymbol{\Theta})$ and compiling estimators $\hat{\mathcal{Z}}_k$ separately—though the latter would simplify our convergence analysis the former should be (asymptotically) much more efficient.

B.2. *Consistency of INS*

To establish a heuristic argument for consistency of the INS scheme we must first consider the nature of the limiting distribution for the sequence of ellipsoidal decompositions, $\{E_i(\cdot)\}$, as $N_{\text{live}} \to \infty$. To do so we introduce the following strong assumption: that for the constrained-likelihood contour corresponding to each $X_i \in [0,1]$ there exists a unique, limiting ellipsoidal decomposition, $E^*_{X_i}(\cdot)$, to which the MULTINEST algorithm's $E_i(\cdot)$ will converge (if not stopped early) *almost surely* for all N_{live} for which $X_i = \exp(-i/N_{\text{live}})$ for some $i \in \{1, 2, \ldots, c \times N_{\text{live}}\}$. In particular, we suppose that both the design of the EM plus k-means code for constructing our $E_i(\cdot)$ and the nature of the likelihood function, $\mathcal{L}(\boldsymbol{\Theta})$, are such for any given $\epsilon > 0$ there is an N_{live} large enough that thereafter

$$\sup_i \left(\frac{\mathrm{P}_{\pi(\boldsymbol{\Theta})}\{\boldsymbol{\Theta} \in E_i(\cdot) \cap E^*_{X_i=\exp(-i/N_{\text{live}})}(\cdot)\}}{\mathrm{P}_{\pi(\boldsymbol{\Theta})}\{\boldsymbol{\Theta} \in E_i(\cdot) \cup E^*_{X_i=\exp(-i/N_{\text{live}})}(\cdot)\}} \right) > 1 - \epsilon.$$

Another supposition we make is that the limiting family of ellipsoidal decompositions, $\{E^*_{X_i}(\cdot)\}$, is at least left or right 'continuous' in the same sense at every point of its rational baseline; i.e., for each X_i and any $\epsilon > 0$ there exists a $\delta > 0$ such that $X_i - X_j < \delta$ and/or $X_j - X_i < \delta$ implies

$$\frac{\mathrm{P}_{\pi(\boldsymbol{\Theta})}\{\boldsymbol{\Theta} \in E^*_{X_i}(\cdot) \cap E^*_{X_j}(\cdot)\}}{\mathrm{P}_{\pi(\boldsymbol{\Theta})}\{\boldsymbol{\Theta} \in E^*_{X_i}(\cdot) \cup E^*_{X_j}(\cdot)\}} > 1 - \epsilon.$$

Various conditions for almost sure convergence of EM (Wu 1983) and k-means (Pollard 1981) algorithms have been demonstrated in the past, but we have an intractable dependence structure operating on the $E_k(\cdot)$ for INS and it is not at all obvious how to clearly formulate such conditions here. The complexity of this task can perhaps most easily be appreciated by considering the limited availability of results for the convergence in volume of random convex hulls from uniform sampling within regular polygons in high-dimensions (e.g. Schneider and Wieacker 1980; Schneider 2008). On the other hand, we may suspect that the necessary conditions for the above are similar to those required in any case for almost sure 'coverage' of each constrained-likelihood surface by its corresponding ellipsoidal decomposition; the latter being an often ignored assumption of rejection NS. That is, even for a generous dilation of the simple proposal ellipsoids, as suggested by Mukherjee et al. (2006), one can easily identify some family of (typically non-convex) $\mathcal{L}(\boldsymbol{\Theta})$ for which the given dilation factor will be demonstrably insufficient; though whether such a 'pathological' $\mathcal{L}(\boldsymbol{\Theta})$ would be likely to arise in standard statistical settings is perhaps another matter entirely!

The necessity of these assumptions, and in particular our second regarding the 'continuity' of the limiting $\{E^*_{X_i}(\cdot)\}$, is two-fold: to ensure that a limiting distribution exists (this echoes the requirement for there to exist an optimal proposal in the equivalent AMIS analysis of Marin et al. 2012), *and* to ensure that its form is such as to render irrelevant the inevitable stochastic variation and bias in our negative binomial sampling of n_i points from $E_i(\cdot)$. Important to acknowledge is that not only is the number of points drawn from each ellipsoidal decomposition, n_i, a random variable, but the collection of $n_i - 1$ draws in $E_i(\cdot) \cap \{\mathcal{L}(\boldsymbol{\Theta}) < \mathcal{L}_{i-1}\}$ plus one in $E_i(\cdot) \cap \mathcal{L}(\boldsymbol{\Theta}) > \mathcal{L}_{i-1}$ from a single realization cannot strictly be considered a uniform draw from $E_i(\cdot)$, though we treat it as such in our summation for $g(\boldsymbol{\Theta})$. Indeed the expected proportion of these draws represented by the single desired $\mathcal{L}(\boldsymbol{\Theta}) > \mathcal{L}_{i-1}$ point, namely $\mathrm{E}[\frac{1}{n_i}] = \frac{-p_i \log p_i}{1-p_i}$, does not even match its fraction of $\pi(\boldsymbol{\Theta})$ by 'volume', here p_i. Our argument must therefore be that (asymptotically) with more and more near-identical ellipsoids converging in the vicinity of each $E^*_i(\cdot)$ as $N_{\text{live}} \to \infty$ the *pool* of all such biased draws from our constrained-likelihood sampling within each of these nearby, near-identical ellipsoids ultimately approximates an unbiased sample, *and* that the mean number of draws from these will approach its long-run average, say n^*_i.

With such convergence towards a limiting distribution, $F^*(\boldsymbol{\Theta})$, defined by the set of pairings, $\{E^*_i(\cdot), n^*_i\}$, supposed it is then straightforward to confirm that via the Strong Law of Large Numbers the empirical distribution function of the samples $\boldsymbol{\Theta}_k$ drawn under INS, $F^{\text{INS}}(\boldsymbol{\Theta})$, will converge in distribution to this asymptotic form; the convergence-determining classes here being simply the (lower open, upper closed) hyper-cubes in the compact subset, $[0,1]^N$, of $\mathbb{R}^N$. From $F^{\text{INS}}(\boldsymbol{\Theta}) \overset{\text{d}}{\Rightarrow} F^*(\boldsymbol{\Theta})$ we have

$$g^{[\text{biased}]}(\boldsymbol{\Theta}), g^{[\text{unbiased}]}(\boldsymbol{\Theta}) \to \frac{\partial^N}{\partial\Theta_1, \ldots, \partial\Theta_N} F^*(\boldsymbol{\Theta}), \tag{32}$$

and thus

$$\mathrm{E}[\hat{\mathcal{Z}}^{\text{INS}}] = \int \frac{f(\boldsymbol{\Theta}) g^{[\text{biased}]}(\boldsymbol{\Theta})}{g^{[\text{unbiased}]}(\boldsymbol{\Theta})} d\boldsymbol{\Theta} \to \int f(\boldsymbol{\Theta}) d\boldsymbol{\Theta} = \mathcal{Z}, \tag{33}$$

and hence (with the corresponding $\mathrm{Var}[\hat{\mathcal{Z}}^{\mathrm{INS}}] \to 0$) the consistency of $\hat{\mathcal{Z}}^{\mathrm{INS}}$.

B.3. *Variance breakdown of INS*

Given the evident dependence of the INS variance on three distinct sources—(I) the stochasticity of the live point set, and its decompositions, $\{E_i(\cdot)\}$; (II) the negative binomial sampling of the n_i; and (III) the importance sampling variance of the drawn $f(\boldsymbol{\Theta}_k)/g(\boldsymbol{\Theta}_k)$—it makes good sense to break these components down into their contributory terms via the Law of Total Variance as follows:

$$\begin{aligned}
&\mathrm{Var}_{\pi(\{E_i(\cdot)\},\{n_i\},\{\boldsymbol{\Theta}_k\})}[\hat{\mathcal{Z}}^{\mathrm{INS}}] \\
&\quad = \mathrm{Var}_{\pi(\{E_i(\cdot)\})}[\mathrm{E}_{\pi(\{n_i\},\{\boldsymbol{\Theta}_k\}|\{E_i(\cdot)\})}[\hat{\mathcal{Z}}^{\mathrm{INS}}]] \\
&\quad +\mathrm{E}_{\pi(\{E_i(\cdot)\})}[\mathrm{E}_{\pi(\{\boldsymbol{\Theta}_k\}|\{E_i(\cdot)\})}[\mathrm{Var}_{\pi(\{n_i\}|\{\boldsymbol{\Theta}_k\},\{E_i(\cdot)\})}[\hat{\mathcal{Z}}^{\mathrm{INS}}]]] \\
&\quad +\mathrm{E}_{\pi(\{E_i(\cdot)\})}[\mathrm{Var}_{\pi(\{\boldsymbol{\Theta}_k\}|\{E_i(\cdot)\})}[\mathrm{E}_{\pi(\{n_i\}|\{\boldsymbol{\Theta}_k\},\{E_i(\cdot)\})}[\hat{\mathcal{Z}}^{\mathrm{INS}}]]]. \qquad (34)
\end{aligned}$$

Now the first term represents explicitly the variance contribution from the inherent randomness of the ellipsoidal decomposition sequence, $\{E_i(\cdot)\}$, which we might suppose negligble provided the geometric exploration of the posterior has been 'sufficiently thorough', meaning that the N_{live} point set has evolved through all the significant posterior modes. The second and third terms represent the negative binomial sampling and 'ordinary' importance sampling variance contributions expected under the distribution of $\{E_i(\cdot)\}$. With the realised $\{E_i(\cdot)\}$ being, of course, an unbiased draw from its parent distribution any unbiased estimator of these two additional variance components applied to our realised $\{n_i\}$ and $\{\boldsymbol{\Theta}_k\}$ could be considered likewise unbiased. However, no such estimators are readily available, hence we pragmatically suppose the second term also negligble and make do with the 'ordinary' importance sampling estimator, given by Eq. (20), for the third term.

Acknowledging the possibility for under-estimation of the INS variance in this way it becomes prudent to consider strategies for minimising our unaccounted variance contributions. The first, suggested by our preceeding discussion of asymptotic consistency for the INS, is to maximise the size of the live point set used. Of course, whether for INS or vanilla NS with MULTINEST we have no prescription for the requisite N_{live}, and the range of feasible N_{live} will often be strongly limited by the available computational resources. Hence we can give here only the general advice of cautious application; in particular it may be best to confirm a reasonable similarly between the estimated variance from Eq. (20) above and that observed from repeat simulation at an N_{live} of manageable runtime prior to launching MULTINEST at a more expensive N_{live}. The second means of reducing the variance in our two unaccounted terms is to stick with the default mode of MULTINEST, rather than opt for 'constant efficiency mode', since by bounding the maximum rate at which the ellipsoidal decompositions may shrink towards the posterior mode we automatically reduce the variance in the random variable, $\{E_i(\cdot)\}$, *and* that of $\{n_i\}$ and $\hat{\mathcal{Z}}$ conditional upon it!

C. SOME MEASURE-THEORETIC CONSIDERATIONS

When outlining in Sec. 3 the transformation from integration of $\mathcal{L}(\boldsymbol{\Theta})$ over $\pi(\boldsymbol{\Theta})d\boldsymbol{\Theta}$ to integration of $\mathcal{L}(X)$ over dX (the prior mass cumulant) at the heart of the NS algorithm, we elected, in the interest of simplicity, to omit a number of underlying measure-theoretic details. The significance of these are perhaps only of particular relevance to understanding the use of the NS posterior weights, $\mathcal{L}_i w_i/\hat{\mathcal{Z}}$ from Eq. (10), for inference regarding functions of $\boldsymbol{\Theta}$ (e.g. its first, second, and higher-order moments) with respect to the posterior density. However, as this issue has been raised by Chopin and Robert (2010) and we feel that their Lemma 1 deserves clarification we give here a brief measure-theoretic formulation of NS with this in mind.

As with many Bayesian inference problems we begin by supposing the existence of two well-defined probability spaces: (I) that of the prior with (normalised) probability measure, P_π, defined for events in the σ-algebra, $\Sigma_{\boldsymbol{\Theta}}$, of its sample space, $\Omega_{\boldsymbol{\Theta}}$, i.e., $(P_\pi, \Omega_{\boldsymbol{\Theta}}, \Sigma_{\boldsymbol{\Theta}})$, and (II) that of the posterior with measure $P_{\pi'}$ defined on the same space, i.e., $(P_{\pi'}, \Omega_{\boldsymbol{\Theta}}, \Sigma_{\boldsymbol{\Theta}})$. Moreover, we suppose that each may be characterised by its Radon–Nikodym derivative with respect to a common σ-finite baseline measure on a complete, separable metric space; that is, $P_\pi(A \in \Sigma_{\boldsymbol{\Theta}}) = \int_A \pi(\boldsymbol{\Theta})\{d\boldsymbol{\Theta}\}$ and $P_{\pi'}(A \in \Sigma_{\boldsymbol{\Theta}}) = \int_A \pi(\boldsymbol{\Theta})\mathcal{L}(\boldsymbol{\Theta})/\mathcal{Z}\{d\boldsymbol{\Theta}\}$. NS then proposes to construct the induced measure, P_X, on the σ-algebra, Σ_X, generated by the Borel sets of the sample space, $\Omega_X = [0,1] \in \mathbb{R}$, and defined by the transformation, $X : \Omega_{\boldsymbol{\Theta}} \mapsto \Omega_X$ with $X(\boldsymbol{\Theta}') = \int_{\{\boldsymbol{\Theta}:\mathcal{L}(\boldsymbol{\Theta})>\mathcal{L}(\boldsymbol{\Theta}')\}} \pi(\boldsymbol{\Theta})\{d\boldsymbol{\Theta}\}$. The validity of which requires only the measurability of this transformation (i.e., $X^{-1}B \in \Sigma_{\boldsymbol{\Theta}}$ for all $B \in \Sigma_X$); e.g. in the metric space $\mathbb{R}^k$ with reference Lebesgue measure, the almost everywhere continuity of $\mathcal{L}(\boldsymbol{\Theta})$. However, for the proposed Riemann integration of vanilla NS to be valid we will also need the induced P_X to be absolutely continuous with respect to the Lebesgue reference measure on $[0,1]$, such that we can write $P_X(B \in \Sigma_X) = \int_B \mathcal{L}(X)/\mathcal{Z}\{d\boldsymbol{X}\}$. The additional condition for this given by Chopin and Robert (2010) is that $\mathcal{L}(\boldsymbol{\Theta})$ has connected support. To state the objective of vanilla NS in a single line: if we can compute $\mathcal{L}(X)$ we can find $\mathcal{Z}$ simply by solving for $\int_0^1 \mathcal{L}(X)/\mathcal{Z}\{dX\} = 1$.

In their Sec. 2.2 Chopin and Robert (2010) examine the NS importance sampling estimator proposed for the posterior expectation of a given function $f(\boldsymbol{\Theta})$, namely

$$\mathrm{E}_{\pi'}[f(\boldsymbol{\Theta})] = \int_{\Omega_{\boldsymbol{\Theta}}} f(\boldsymbol{\Theta})\pi(\boldsymbol{\Theta})\mathcal{L}(\boldsymbol{\Theta})/\mathcal{Z}\{d\boldsymbol{\Theta}\}, \qquad (35)$$

which one may approximate with $\hat{\mathrm{E}}[f(\boldsymbol{\Theta})] = \sum f(\boldsymbol{\Theta}_i)\mathcal{L}_i w_i$ from NS Monte Carlo sampling. They note that $f(\boldsymbol{\Theta})$ is in this

context a noisy estimator of $\tilde{f}(X) = \mathrm{E}_\pi[f(\mathbf{\Theta})|\mathcal{L}(\mathbf{\Theta}) = \mathcal{L}(X)]$, and propose in their Lemma 1 that the equality,

$$\int_0^1 \tilde{f}(X)\mathcal{L}(X)\{dX\} = \int_{\Omega_{\mathbf{\Theta}}} f(\mathbf{\Theta})\pi(\mathbf{\Theta})\mathcal{L}(\mathbf{\Theta})\{d\mathbf{\Theta}\}, \tag{36}$$

holds *when $\tilde{f}(X)$ is absolutely continuous*. We agree that this is true and Chopin and Robert (2010) give a valid proof in their Appendix based on the Monotone Convergence Theorem. However, we suggest that given the already supposed validity of the measure P_X, and its Radon–Nikodym derivative with respect to the reference Lebesgue measure, $\{dX\}$, upon which NS is based, the equality of the above relation is already true without absolute continuity via the change of variables theorem (Halmos 1950), in the sense that wherever one side exists the other exists and will be equal to it. One trivial example of a discontinuous $\tilde{f}(X)$ for which both sides of 36 exist and are equal is that induced by the indicator function for $\mathcal{L}(\mathbf{\Theta}) > X^*$. To see that the $\tilde{f}(X)$ corresponding to a given, measurable $f(\mathbf{\Theta})$ has the stated interpretation as a conditional expectation we observe that $\mathrm{E}_\pi[f(\mathbf{\Theta})|\mathcal{L}(\mathbf{\Theta}) = \mathcal{L}(X)]$ may be written as

$$\int_{\Omega_{\mathbf{\Theta}}} f(\mathbf{\Theta})\pi(\mathbf{\Theta})\mathrm{I}_{\mathcal{L}(\mathbf{\Theta})=\mathcal{L}(X)}\{d\mathbf{\Theta}\}, \tag{37}$$

a function of X, using the interpretation of conditional distributions as derivatives (Pfanzagl 1979). Thus,

$$\int_{\Omega_{\mathbf{\Theta}}} f(\mathbf{\Theta})\pi(\mathbf{\Theta})\mathcal{L}(\mathbf{\Theta})\{d\mathbf{\Theta}\} = \int_0^1 \left(\int_{\Omega_{\mathbf{\Theta}}} f(\mathbf{\Theta})\pi(\mathbf{\Theta})\mathrm{I}_{\mathcal{L}(\mathbf{\Theta})=\mathcal{L}(X)}\{d\mathbf{\Theta}\}\right)\mathcal{L}(X)\{dX\}. \tag{38}$$

REFERENCES

Allanach, B. C., and Lester, C. G. (2008), "Sampling using a bank of clues," *Computer Physics Communications*, 179, 256–266.

Beaujean, F., and Caldwell, A. (2013), "Initializing adaptive importance sampling with Markov chains," *ArXiv e-prints [arXiv:1304.7808]*, .

Billingsley, P. (1995), "Probability and measure.," *John Wiley & Sons, New York*, .

Bridges, M., Feroz, F., Hobson, M. P., and Lasenby, A. N. (2009), "Bayesian optimal reconstruction of the primordial power spectrum," MNRAS, 400, 1075–1084.

Cameron, E., and Pettitt, A. (2013), "Recursive Pathways to Marginal Likelihood Estimation with Prior-Sensitivity Analysis," *ArXiv e-prints [arXiv:1301.6450]*, .

Chen, M.-H., and Shao, Q.-M. (1997), "On Monte Carlo methods for estimating ratios of normalizing constants," *The Annals of Statistics*, 25(4), 1563–1594.

Chopin, N., and Robert, C. P. (2010), "Properties of nested sampling," *Biometrika*, 97(3), 741–755.

Clyde, M. A., Berger, J. O., Bullard, F., Ford, E. B., Jefferys, W. H., Luo, R., Paulo, R., and Loredo, T. (2007), Current Challenges in Bayesian Model Choice,, in *Statistical Challenges in Modern Astronomy IV*, eds. G. J. Babu, and E. D. Feigelson, Vol. 371 of *Astronomical Society of the Pacific Conference Series*, p. 224.

Cornuet, J., Marin, J.-M., Mira, A., and Robert, C. P. (2012), "Adaptive multiple importance sampling," *Scandinavian Journal of Statistics*, 39(4), 798–812.

Corsaro, E., and De Ridder, J. (2014), "DIAMONDS: A new Bayesian nested sampling tool," A&A, 571, A71.

Feroz, F., Balan, S. T., and Hobson, M. P. (2011*a*), "Bayesian evidence for two companions orbiting HIP 5158," MNRAS, 416, L104–L108.

Feroz, F., Balan, S. T., and Hobson, M. P. (2011*b*), "Detecting extrasolar planets from stellar radial velocities using Bayesian evidence," MNRAS, 415, 3462–3472.

Feroz, F., Gair, J. R., Graff, P., Hobson, M. P., and Lasenby, A. (2010), "Classifying LISA gravitational wave burst signals using Bayesian evidence," *Classical and Quantum Gravity*, 27(7), 075010.

Feroz, F., Gair, J. R., Hobson, M. P., and Porter, E. K. (2009), "Use of the MultiNest algorithm for gravitational wave data analysis," *Classical and Quantum Gravity*, 26(21), 215003–+.

Feroz, F., and Hobson, M. P. (2008), "Multimodal nested sampling: an efficient and robust alternative to Markov Chain Monte Carlo methods for astronomical data analyses," MNRAS, 384, 449–463.

Feroz, F., Hobson, M. P., and Bridges, M. (2009), "MultiNest: an efficient and robust Bayesian inference tool for cosmology and particle physics," MNRAS, 398, 1601–1614.

Feroz, F., Hobson, M. P., Zwart, J. T. L., Saunders, R. D. E., and Grainge, K. J. B. (2009), "Bayesian modelling of clusters of galaxies from multifrequency-pointed Sunyaev-Zel'dovich observations," MNRAS, 398, 2049–2060.

Feroz, F., Marshall, P. J., and Hobson, M. P. (2008), "Cluster detection in weak lensing surveys," *ArXiv e-prints [arXiv:0810.0781]*, .

Geyer, C. J. (1994), "Estimating Normalizing Constants and Reweighting Mixtures in Markov Chain Monte Carlo,".

Gill, R. D., Vardi, Y., and Wellner, J. A. (1988), "Large sample theory of empirical distributions in biased sampling models," *The Annals of Statistics*, pp. 1069–1112.

Graff, P., Feroz, F., Hobson, M. P., and Lasenby, A. (2012), "BAMBI: blind accelerated multimodal Bayesian inference," MNRAS, 421, 169–180.

Habeck, M. (2012), "Bayesian Estimation of Free Energies From Equilibrium Simulations," *Physical Review Letters*, 109(10), 100601.

Halmos, P. R. (1950), *Measure theory*, Vol. 2 van Nostrand New York.

Handley, W.J., Hobson, M. P., and Lasenby, A.N. (2015a), "POLYCHORD: nested sampling for cosmology," MNRAS*Letters*, 450, L61–L65.

Handley, W.J., Hobson, M. P., and Lasenby, A.N. (2015b), "POLYCHORD: next-generation nested sampling," MNRAS, 453, 4384–4398.

Hesterberg, T. (1995), "Weighted average importance sampling and defensive mixture distributions," *Technometrics*, 37(2), 185–194.

Higson, E., Handley, W.J., Hobson, M. P., and Lasenby, A.N. (2015), "Dynamic nested sampling: an improved algorithm for parameter estimation and evidence calculation," *Statistics and Computing*,.

Karpenka, N. V., Feroz, F., and Hobson, M. P. (2013), "A simple and robust method for automated photometric classification of supernovae using neural networks," MNRAS, 429, 1278–1285.

Karpenka, N. V., March, M. C., Feroz, F., and Hobson, M. P. (2012), "Bayesian constraints on dark matter halo properties using gravitationally-lensed supernovae," *ArXiv e-prints [arXiv:1207.3708]*, .

Keeton, C. R. (2011), "On statistical uncertainty in nested sampling," *Monthly Notices of the Royal Astronomical Society*, 414(2), 1418–1426.

Kipping, D. M., Bakos, G. Á., Buchhave, L., Nesvorný, D., and Schmitt, A. (2012), "The Hunt for Exomoons with Kepler (HEK). I. Description of a New Observational project," ApJ, 750, 115.

Kong, A., McCullagh, P., Meng, X.-L., Nicolae, D., and Tan, Z. (2003), "A theory of statistical models for Monte Carlo integration," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(3), 585–604.

Liu, J. (2008), *Monte Carlo Strategies in Scientific Computing*, Springer Series in Statistics Series.

Mackay, D. J. C. (2003), *Information Theory, Inference and Learning Algorithms* Cambridge University Press, Cambridge, UK.

Marin, J.-M., Pudlo, P., and Sedki, M. (2012), "Consistency of the Adaptive Multiple Importance Sampling," *ArXiv e-prints [arXiv:1211.2548]*, .

Mukherjee, P., Parkinson, D., and Liddle, A. R. (2006), "A nested sampling algorithm for cosmological model selection," *The Astrophysical Journal Letters*, 638(2), L51.

Owen, A., and Zhou, Y. (2000), "Safe and effective importance sampling," *Journal of the American Statistical Association*, 95(449), 135–143.

Pfanzagl, P. (1979), "Conditional distributions as derivatives," *The Annals of Probability*, pp. 1046–1050.

Pollard, D. (1981), "Strong Consistency of K-Means Clustering," *The Annals of Statistics*, 9(1), 135–140.

Rubinstein, R. Y., and Kroese, D. P. (2004), *The cross-entropy method: a unified approach to combinatorial optimization, Monte-Carlo simulation and machine learning* Springer Verlag.

Schneider, R. (2008), "Recent results on random polytopes," *Boll. Unione Mat. Ital.(9)*, 1, 17–39.

Schneider, R., and Wieacker, J. A. (1980), "Random polytopes in a convex body," *Probability Theory and Related Fields*, 52(1), 69–73.

Shaw, J. R., Bridges, M., and Hobson, M. P. (2007), "Efficient Bayesian inference for multimodal problems in cosmology," MNRAS, 378, 1365–1370.

Sivia, D., and Skilling, J. (2006), *Data Analysis A Bayesian Tutorial* Oxford University Press.

Skilling, J. (2004), Nested Sampling,, in *American Institute of Physics Conference Series*, eds. R. Fischer, R. Preuss, and U. V. Toussaint, Vol. 119, pp. 1211–1232.

Skilling, J. (2006), "Nested sampling for general Bayesian computation," *Bayesian Analysis*, 1(4), 833–859.

Speagle, J.S. (201(), "dynesty: A Dynamic Nested Sampling Package for Estimating Bayesian Posteriors and Evidences," in *ArXiv e-prints [arXiv:1904.02180]*, .

Strege, C., Bertone, G., Feroz, F., Fornasa, M., Ruiz de Austri, R., and Trotta, R. (2013), "Global fits of the cMSSM and NUHM including the LHC Higgs discovery and new XENON100 constraints," JCAP, 4, 13.

Tan, Z., Gallicchio, E., Lapelosa, M., and Levy, R. M. (2012), "Theory of binless multi-state free energy estimation with applications to protein-ligand binding," *The Journal of Chemical Physics*, 136, 144102.

Teachey, A., and Kipping, D. (2018), "Evidence for a large exomoon orbiting Kepler-1625b," *Science Advances*, 4, 10.

Trotta, R. (2008), "Bayes in the sky: Bayesian inference and model selection in cosmology," *Contemporary Physics*, 49, 71–104.

Vardi, Y. (1985), "Empirical distributions in selection bias models," *The Annals of Statistics*, pp. 178–203.

Veach, E., and Guibas, L. (1995), "Bidirectional estimators for light transport," in *Photorealistic Rendering Techniques* Springer, pp. 145–167.

Veitch, J., and Vecchio, A. (2010), "Bayesian coherent analysis of in-spiral gravitational wave signals with a detector network," in *Physical Review D*, 81, 6.

Weinberg, M. D., Yoon, I., and Katz, N. (2013), "A remarkably simple and accurate method for computing the Bayes Factor from a Markov chain Monte Carlo Simulation of the Posterior Distribution in high dimension," *ArXiv e-prints [arXiv:1301.3156]*, .

White, M. J., and Feroz, F. (2010), "MSSM dark matter measurements at the LHC without squarks and sleptons," *Journal of High Energy Physics*, 7, 64.

Wu, C. (1983), "On the convergence properties of the EM algorithm," *The Annals of Statistics*, 11(1), 95–103.

This paper was built using the Open Journal of Astrophysics LaTeX template. The OJA is a journal which provides fast and easy peer review for new papers in the `astro-ph` section of the arXiv, making the reviewing process simpler for authors and referees alike. Learn more at `http://astro.theoj.org`.