Digital Signal Processing 70 (2017) 84–93

Contents lists available at ScienceDirect

Digital Signal Processing

www.elsevier.com/locate/dsp

Combined-chain nested sampling for efficient Bayesian model comparison

R. Wesley Henderson *, Paul M. Goggans, Lei Cao

Department of Electrical Engineering, University of Mississippi, University, MS 38677, United States

ARTICLE INFO

Article history:
Available online 3 August 2017

Keywords:
Bayesian inference
Model comparison
MCMC
Nested sampling
Parallel computing

ABSTRACT

Model comparison problems arise in many fields of science and engineering, including signal processing. In these problems, we wish to quantify how well each of a set of possible models describes a set of observations. Many numerical techniques exist to perform model comparison, but this paper focuses on nested sampling, which is a numerical integration algorithm for evaluating probabilities of models. The original formulation of nested sampling is a strictly sequential algorithm. Most modern advances in computing are via parallel processing, however, and we therefore present a novel method for parallelizing nested sampling. This paper sets out the mathematical foundation for this parallelization, as well as ideas for implementing it. Three examples demonstrate the effectiveness of the present parallel technique in realistic scientific and engineering data analysis problems.

1. Introduction

Bayesian inference allows scientists and engineers to draw conclusions from data in the presence of uncertainty. Inference problems confront practitioners in many contexts, including acoustics, astronomy, and digital signal processing in general. Within inference, model comparison is an important class of problems, which involves quantifying the plausibility of mathematical models given a set of observations, so as to compare and rank various models quantitatively.

Knuth et al. [1] give an overview of model comparison problems in various domains, with an emphasis on signal processing. Design-as-inference refers to the application of Bayesian model comparison and parameter estimation to design problems; it has been successfully applied to the design of finite impulse response (FIR) filters by Chan and Goggans [2] and to the design of infinite impulse response (IIR) filters by Botts et al. [3]. In acoustic signal processing, Bayesian model comparison has been applied to the analysis of multiple decay slopes in coupled volumes [4–7], the analysis of room modes [8,9], and the design and analysis of multilayer sound absorbers [10].

These and other examples motivate the development of an efficient and effective method for performing Bayesian model selection. Nested sampling [11–13] [14, Chapter 9] provides a good starting point. Nested sampling is a robust method for numerically evaluating model probability integrals. In its original form it is a serial algorithm. Modern advances in computing have mostly been directed to increasing the amount of parallel processing power available to users, rather than simply increasing serial processing speed. Algorithms designed to be implemented on parallel computing architectures are therefore able to solve problems more quickly on modern hardware.

The present paper develops a method for 'parallelizing' nested sampling in a way that is simple and easy to implement. Other researchers have published methods for parallelizing nested sampling. Burkoff et al. [15] describe an alternate way to parallelize nested sampling, which works by discarding and replacing multiple live samples for each likelihood constraint. (See Section 2.1.) Martiniani et al. [16] describe another application of the Burkoff method, and we have previously described [17] a method for implementing this method while maintaining a given level of precision in the log-evidence estimate. The method described below implements the parallelization of nested sampling in a different and more effective way.

Other ways to improve the performance of nested sampling without necessarily involving parallelization have been presented by Brewer et al. (diffusive nested sampling) [18], Feroz et al. (MultiNest) [19], and Handley et al. (PolyChord) [20].

This paper is organized as follows. Section 2 provides a brief overview of Bayesian inference and nested sampling, and motivates the development of a parallelized nested sampling algorithm. Section 3 describes the specific method by which the samples produced by multiple independent runs of nested sampling can be

* Corresponding author.
E-mail addresses: rwhender@go.olemiss.edu (R.W. Henderson), goggans@olemiss.edu (P.M. Goggans), lcao@olemiss.edu (L. Cao).

http://dx.doi.org/10.1016/j.dsp.2017.07.021
1051-2004/

combined so as to generate a single estimate of the model probability. Section 4 details several strategies for implementing this method of combining independent nested sampling chains. Section 5 provides three illustrative examples of this method and demonstrates the method's effectiveness. Section 6 concludes the paper.

2. Bayesian inference and nested sampling

Bayesian inference provides a uniquely consistent way to learn from observations in the presence of uncertainty. Model-based Bayesian inference can be broken into two levels: parameter estimation and model comparison. In traditional statistics, probability can be used to describe only random variables, but in the Bayesian view of probability any proposition can be assigned a probability conditioned upon any other. We may use Bayes' theorem to write the probability for a set of model parameters $\boldsymbol{\Theta}$ given observed data $\mathbf{D}$, model M, and prior information I, as

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

Below, the following abbreviations will often be used: $\Pr(\boldsymbol{\Theta}|\mathbf{D}, M, I) \equiv \mathcal{P}(\boldsymbol{\Theta})$ for the posterior, $\Pr(\mathbf{D}|\boldsymbol{\Theta}, M, I) \equiv \mathcal{L}(\boldsymbol{\Theta})$ for the likelihood, $\Pr(\boldsymbol{\Theta}|M, I) \equiv \pi(\boldsymbol{\Theta})$ for the prior, and $\Pr(\mathbf{D}|M, I) \equiv \mathcal{Z}$ for the evidence. Equation (1) provides a framework for performing Bayesian parameter estimation. Various analytical and numerical techniques exist for implementing (1) in this way. Gregory [21] provides a fine explanation of Bayesian parameter estimation in general and techniques for implementing it in practice.

This paper is less concerned with parameter estimation and more concerned with the second layer of inference, model comparison. Once again using Bayes' theorem, we can express the probability for a model M, given data $\mathbf{D}$ and prior information I, as

$$\Pr(M|\mathbf{D}, I) \propto \Pr(\mathbf{D}|M, I)\Pr(M|I). \tag{2}$$

The normalizing constant $\Pr(\mathbf{D}|I)$ is omitted here because proper posterior probabilities for models are rarely necessary (or, in fact, available). To calculate this constant an exhaustive set of models must be specified, which is usually impossible. Model selection mostly calls for the pairwise comparison of different models in the light of the same data. It is then useful to write the ratio of probabilities of model M_i and model M_j as

$$\frac{\Pr(M_i|\mathbf{D}, I)}{\Pr(M_j|\mathbf{D}, I)} = \frac{\Pr(\mathbf{D}|M_i, I)}{\Pr(\mathbf{D}|M_j, I)} \frac{\Pr(M_i|I)}{\Pr(M_j|I)}. \tag{3}$$

This equation shows that the posterior ratio of probabilities is given by multiplying the prior ratio by the ratio of likelihoods appearing in the first term on the right-hand side, which imports the data.

The model priors (in the right-most fraction in (3)) are set according to the user's prior knowledge, and are known in advance of any observations. To find the pairwise ratios of model posteriors (the left-hand side in (3)), we need to specify the model likelihood values. In fact the model likelihood in (3) is the same expression as the evidence (the denominator) in the associated parameter estimation problem (1).

The evidence in (1) acts as a normalizing constant for the posterior distribution over the set of parameters $\boldsymbol{\Theta}$. It can therefore be found by integrating the product of the prior and the likelihood for a given model over the parameter space:

$$\Pr(\mathbf{D}|M, I) = \int_{\boldsymbol{\Theta}} \Pr(\mathbf{D}|\boldsymbol{\Theta}, M, I)\Pr(\boldsymbol{\Theta}|M, I)\, \mathrm{d}\boldsymbol{\Theta}. \tag{4}$$

The integrand in (4) is often close to zero over much of $\boldsymbol{\Theta}$, with large values concentrated in a small portion of the parameter space. Models may also have many parameters, so that the integration is in a multi-dimensional space. As a result, numerical integration over the parameter space using any reasonable discretization of the variables gives rise to unacceptably large errors.

Alternative techniques for computing the evidence have been adapted from statistical mechanics. The evidence for a model given a set of data is analogous to the free energy in a given thermodynamic state. Based on this analogy, thermodynamic integration [22] computes the Bayesian evidence by integrating the expectation of the log-likelihood over an inverse temperature parameter.

Unfortunately, thermodynamic integration suffers from several serious limitations. It typically takes a long time to run for problems with large numbers of data or a large number of parameters. Also, likelihood functions with discontinuities—comparable to phase transitions in statistical mechanics—tend to confound thermodynamic integration. Nested sampling [12] is another technique for computing Bayesian evidence and was developed partly to overcome these limitations.

2.1. Nested sampling

Nested sampling is similar to thermodynamic integration in that it side-steps the multi-dimensional integral in (4) by using a one-dimensional reparameterization to find the evidence. Instead of cooling a temperature parameter to gradually introduce the likelihood, nested sampling integrates the likelihood over the prior mass. The prior mass, X, is defined as the proportion of the prior distribution contained within a likelihood threshold L,

$$X(L) = \int_{\{\boldsymbol{\Theta}:\mathcal{L}(\boldsymbol{\Theta})>L\}} \pi(\boldsymbol{\Theta})\, \mathrm{d}\boldsymbol{\Theta}. \tag{5}$$

As the prior mass is a 1-to-1 function of the likelihood, the likelihood threshold can be expressed as a function of the prior mass. Ultimately the evidence can be expressed as

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

The detailed derivation is given in Skilling's original paper on nested sampling [12].

At first sight it appears that we are no better off with this representation, because the integral in (5) is no easier to evaluate than the integral in (4). Nested sampling does not require exact computation of the prior mass, however; an estimate is sufficient. Nested sampling simultaneously generates estimates for the prior mass and incorporates them into a numerical evaluation of (6). This process will now be described.

Nested sampling proceeds by setting the initial likelihood threshold to 0 and drawing N samples from $\pi(\boldsymbol{\Theta})$. These N samples are known as "live" samples. The likelihood values for each of these live samples are computed exactly. The live sample with the least likelihood is then discarded from the set and recorded for later use. The likelihood of the discarded sample is set as the new likelihood threshold, within which a new sample will be generated in the next round of operations.

The shrinkage t_i in the prior mass X_i at the ith step in the process is distributed as [12]

$$t_i \sim \mathrm{Beta}(N, 1). \tag{7}$$

The prior mass is estimated at each likelihood threshold using the log-geometric mean of the shrinkage ($\mathrm{E}(\log t_i) = -1/N$). Remembering that t_i is the shrinkage in the prior mass at step i, the prior

mass at each step can be computed using the cumulative product of the shrinkage estimates,

$$X_i = \prod_{k=1}^{i} t_k. \tag{8}$$

Once these prior mass estimates are computed, the evidence is estimated using a quadrature approximation of (6). Upon assuming that we have collected m discarded samples, and setting $X_0 = 1$,

$$Z \approx \sum_{i=1}^{m} (X_{i-1} - X_i) L_i. \tag{9}$$

The bulk of the posterior distribution usually occupies a tiny fraction of the prior, approximately $\exp(-H)$, where H is the negative relative entropy of the posterior ($\mathrm{d}P$) with respect to the prior ($\mathrm{d}X$),

$$H = \int \log(\mathrm{d}P/\mathrm{d}X)\, \mathrm{d}P. \tag{10}$$

Each discarded sample shrinks the prior mass by about $\exp(-1/N)$; it therefore takes about $N \times H$ steps to reach the bulk of the posterior. To ensure this region is reached, $2 \times N \times H$ steps are usually allowed for before a nested sampling run is terminated. The error in the evidence estimate is dominated by two factors: the error in the shrinkage estimate, and the error in the process used to replace the discarded sample at each step. The latter is difficult to estimate ahead of time, but the uncertainty in the estimate due to the shrinkage is bounded by [12]

$$\exp\left(\pm\sqrt{H/N}\right).$$

3. Parallelizing nested sampling

As in other numerical integration techniques, there is an antagonistic relationship between precision and speed in nested sampling. Precision in the log-evidence estimate increases with the square root of N, whereas the number of necessary likelihood thresholds increases with N. In this section we discuss our new method for parallelizing nested sampling (subsection 3.1) by combining the results (i.e., the discarded samples) of several nested sampling runs. We also contrast this new technique with the above-mentioned method for discarding and replacing multiple live samples at once (subsection 3.2), originally developed by Burkoff [15] and refined in our 2014 paper [17].

3.1. Combining independent chains

Our new method aims to side-step the antagonism between precision and speed by combining the results (discarded samples) of several nested sampling runs, sorting the combined samples by likelihood, estimating the new shrinkage between consecutive pairs of samples, and computing a new evidence estimate. To do this it is necessary to know the distribution of the shrinkage in a combined and re-sorted set of discarded samples. Lemmas 1 and 2 and Theorem 1 establish this distribution.

Lemma 1. *The negative log of the prior mass X_i for the ith sample discarded from a nested sampling run using N live samples has an Erlang density,*

$$f_{-\log X_i}(x) = \frac{N^i x^{i-1} \exp(-Nx)}{(i-1)!}, \tag{11}$$

with shape parameter i and rate parameter N.

Proof. We prove Lemma 1 using the method of induction. Begin with the density of the shrinkage between the $(i+1)$th and ith prior mass values (7) and write it in functional form,

$$f_{t_i}(t) = N t^{N-1}. \tag{12}$$

From (8), we can write the prior mass of the second sample as

$$X_2 = t_1 t_2 \tag{13}$$

so that, by taking the negative logarithm of both sides,

$$-\log X_2 = (-\log t_1) + (-\log t_2). \tag{14}$$

The density of the 1-to-1 function $Y = g(X)$ of a continuous numerical proposition X is given in probability texts such as [23] as

$$f_Y(y) = \begin{cases} f_X(h(y)) \left|\frac{dh}{dy}\right|, & h(y) \in \tilde{X} \\ 0, & \text{otherwise} \end{cases}, \tag{15}$$

where $h(Y)$ is the inverse of $g(X)$, and $\tilde{X}$ is the support of $f_X(x)$. Here, $Y = g(t_i) = -\log t_i$ and $h(Y) = \exp(-Y)$. By applying (15), we have

$$f_{-\log t_i}(y) = f_{t_i}(\exp(-y)) \left|\frac{d}{dy}\exp(-y)\right| = N\exp(-Ny). \tag{16}$$

Equation (16) shows that $-\log t_i$ has an exponential density with mean $1/N$. The sum of two continuous numerical propositions has a density given by the convolution of the densities of the two propositions [23]:

$$\begin{aligned} f_{-\log X_2}(x) &= f_{-\log t_1 - \log t_2}(x) \\ &= \int N\exp(-Nk) N\exp[-(x-k)N]\, dk. \end{aligned} \tag{17}$$

An exponential density has support on the interval $[0, \infty)$, so we have the following restrictions: $0 \le k < \infty$ and $0 \le x - k < \infty$, or in other words

$$0 \le k \le x. \tag{18}$$

The convolution integral becomes

$$f_{-\log X_2}(x) = N^2 \exp(-Nx) \int_0^x db = N^2 x \exp(-Nx). \tag{19}$$

Equation (19) establishes the base case ($k = 1$) for the method of induction.

Assume that the expression for $k = i - 1$ is given by

$$f_{-\log X_{i-1}}(x) = \frac{N^{i-1} x^{i-2} \exp(-Nx)}{(i-2)!}. \tag{20}$$

By using (17) and the previously mentioned limits on distribution support, we have

$$f_{-\log X_i}(x) = \int_0^x \frac{N^{i-1} k^{i-2} \exp(-Nk)}{(i-2)!} N\exp[-(x-k)N]\, dk, \tag{21}$$

$$= \frac{N^i x^{i-1} \exp(-Nx)}{(i-1)!}. \tag{22}$$

Equation (22) specifies a gamma density with shape parameter i and rate parameter N. As i is an integer, (22) is specifically an Erlang density [24]. □

Lemma 2. *Let $\{M_1(t);\ 0 < t < 1\}$ and $\{M_2(t);\ 0 < t < 1\}$ be two independent processes, $0 < t < 1$, and let the following conditions hold:*

$$M_i(t) \geq 0 \tag{23}$$

$$M_i(t) \in \mathbb{Z} \tag{24}$$

$$t_2 \geq t_1 \rightarrow M_i(t_2) \leq M_i(t_1). \tag{25}$$

Let $\mathcal{H}\{\cdot\}$ be a transformation that executes the substitution $t = \exp(-s)$, such that $N_i(s) = \mathcal{H}\{M_i(t)\}$, and $N_i(s)$ is a Poisson process. Let the sum of two processes, e.g. $M_\Sigma(t) = M_1(t) + M_2(t)$, represent a single process with events that consist of the combination of events from the individual processes. Then the transformation $\mathcal{H}\{\cdot\}$ is linear in addition, i.e.

$$\mathcal{H}\{M_1(t) + M_2(t)\} = \mathcal{H}\{M_1(t)\} + \mathcal{H}\{M_2(t)\} = N_1(s) + N_2(s). \tag{26}$$

Proof. See Appendix A. □

Theorem 1. *The shrinkage s_i between consecutive prior mass values Y_{i+1} and Y_i in a set of discarded samples combined from M independent nested sampling runs, each using N_j, $j = 1, \cdots, M$ live samples, has a beta density*

$$f_{s_i}(s) = \left(\sum_{j=1}^{M} N_j\right) s^{\left(\sum_{j=1}^{M} N_j\right)-1} \tag{27}$$

with shape parameters $\sum_{j=1}^{M} N_j$ and 1.

Proof. The wait time for the ith event in a Poisson process is distributed as an Erlang density (which is a special case of the gamma density with an integer shape parameter), [25, p. 291] as shown in (22). This fact and Lemma 1 show that the variable $-\log X_i$ for a set of samples discarded from a single run of nested sampling follows a Poisson process. We can therefore apply the properties of Poisson processes to see what happens when we combine sets of discarded samples from multiple independent runs of nested sampling.

Let Y_i be the prior mass associated with the ith sample in a set of discarded samples combined from M independent nested sampling runs, each of which uses N_j, $j = 1, \cdots, M$ live samples. Let $Q = \sum_{j=1}^{M} N_j$. If M separate and independent Poisson processes with intensities N_j, $j = 1, \cdots, M$ are combined, the result is a single Poisson process with intensity $\sum_{j=1}^{M} N_j$ [26, p. 6]. It follows that $-\log Y_i$ is distributed as

$$f_{-\log Y_i}(y) = \frac{Q^i y^{i-1} \exp(-Q y)}{(i-1)!}, \tag{28}$$

which is an Erlang density with shape parameter i and rate parameter Q.

It follows from the definition of the shrinkage that $-\log Y_i = \sum_{j=1}^{i} -\log s_j$. The sum of propositions with identical densities has a gamma density if and only if the propositions being added also have gamma densities. In this case where the shape parameter is 1, the densities of $-\log s_i$ are also exponential,

$$f_{-\log s_i}(s) = Q \exp(-Q s). \tag{29}$$

The transform in (16) leads to

$$f_{s_i}(s) = Q s^{Q-1}, \tag{30}$$

which is the beta distribution with parameters $Q = \sum_{j=1}^{M} N_j$ and 1. Finally, Lemma 2 shows that the transformation from a process with wait time $-\log s_i$ to a process with wait time s_i is a linear transformation, and our result follows. □

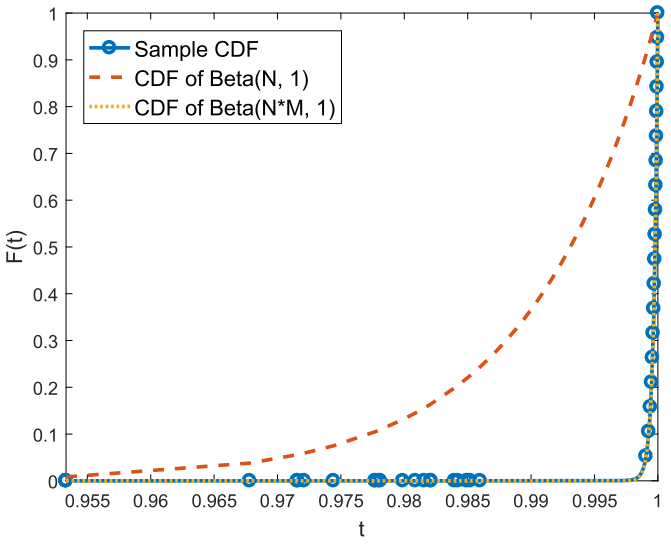


Fig. 1. CDFs for the combined and sampled shrinkage, Beta$(N, 1)$, and Beta$(N \times M, 1)$.

Theorem 1 demonstrates that using a combined set of discarded samples from M independent nested sampling processes which use N live samples each, the evidence estimate can be computed as if one nested sampling process had produced the samples, using $M \times N$ live samples.

3.1.1. Numerical example

We now give a numerical example that demonstrates the validity of the result in Theorem 1.

Example. Let $M = 32$. Generate M sets of 10,000 shrinkage samples from Beta$(100, 1)$ (i.e., $N = 100$). Use (8) to compute the associated prior mass for each sample in each set. Combine the independent sets of samples, then sort the combined sample by prior mass.[1] Compute the actual shrinkage between each pair of consecutive samples. Compare the cumulative distribution function (CDF) of the combined shrinkage samples with the CDFs of Beta$(N, 1)$ and Beta$(N \times M, 1)$. The CDF of the samples should closely match the CDF of Beta$(N \times M, 1)$.

Result. Execution of this procedure yielded the empirical CDF shown in Fig. 1. The CDFs for Beta$(N, 1)$ and Beta$(N \times M, 1)$ are also shown in Fig. 1. The empirical CDF closely matches the CDF for Beta$(N \times M, 1)$, confirming the result in Theorem 1.

3.2. Discarding and replacing several samples at once

Here we describe the parallel nested sampling technique developed by Burkoff et al. [15] and subsequently refined by us [17]. The main idea behind this method for parallelizing nested sampling is that, for a single nested sampling process with N_1 live samples, by discarding and replacing $R \ll N_1$ samples at once we need only HN_1/R steps to reach the desired convergence in the log-evidence estimate. The variance of the shrinkage distribution also increases linearly with R [27], so if we wish to maintain the same precision in the log-evidence estimate then we must also scale the number of live samples as $N_R = \sqrt{R} N_1$. This scaling implies that $HN_1/\sqrt{R}$ steps are required to reach convergence, so that our speed-up factor is approximately $\sqrt{R}$.

[1] Prior mass is a monotonic function of the likelihood constraint, so that sorting by prior mass is equivalent to sorting by likelihood.

4. Implementation

Implementation of the method described in section 3 is flexible, and depends on the user's specific needs. The main idea is that two or more independent runs of nested sampling are conducted with identical numbers of live samples, and the results are then combined and used to generate an estimate of the evidence.

The combination of results for multiple runs proceeds as follows. Let each sample be represented by a data structure with three fields: parameter array, log-likelihood, and log-weight. The samples from each run are combined into one large array, the associated log-weights are discarded, and the samples are sorted by log-likelihood. The log-weight is then re-estimated for each sample using the new ordering and the combined number of live samples, according to the shrinkage distribution given in Theorem 1. The log-evidence is then estimated according to (9) using the new log-weights.

The nested sampling runs can be implemented concurrently or sequentially. Here are some situations that might plausibly arise:

- If you know the value of N necessary to give the precision you require, then instead of performing one run with N live samples, perform M runs concurrently on multiple processor cores or supercomputer nodes, each run using $\lceil N/M \rceil$ live samples.
- If you have an evidence estimate but it is not precise enough, perform several more nested sampling runs incorporating the data, then combine the old and new results to obtain a more precise estimate.
- If you suspect that the distribution you are exploring is highly multi-modal, perform many independent nested sampling runs using a relatively small N so as to increase the likelihood that each mode is well-explored.

4.1. Speed-up

In the first of these three cases the speed-up may be substantial, theoretically by up to a factor of M, but this is limited by several factors. Each nested sampling run involves no communication with the other runs. The speed-up is therefore limited by the total number of live samples used, by the time needed to start each independent process, and by the time needed to combine and process the final results.

Each process must use at least one live sample. If N total live samples are used then not more than N processing units can be used to divide the work.

Whatever the specific architecture used to parallelize the computation, a certain amount of time and computational effort will be needed to copy the necessary data and instructions to the individual processing units before beginning each run. Further time is involved when each processing unit is finished and communicates its results back to the central controlling implementer. Also, the combining of the sets of discarded samples, recomputing of the sample weights, and computing the final evidence takes further time.

Relative to the older method of parallelizing nested sampling by discarding and replacing multiples samples at once, the theoretical speed-up, given M workers, is by a factor of $\sqrt{M}$. Thus, our new method requires $1/\sqrt{M}$ as many likelihood evaluations per worker compared with the older method. The newer method should therefore be preferred over the older method in all cases, unless an external factor (such as the MCMC method used to replace samples) points to use of the older method.

5. Examples

In this section we demonstrate the performance of combined-chain nested sampling using three problems as examples. These include an artificial, highly multi-modal likelihood function, a simulated spectrum analysis problem, and a twin Gaussian shell problem in 20 and 30 dimensions. For these examples, we use nested sampling as described in sections 2.1 and 3. Each example uses a different number of live samples N and a different number of independent runs M.

The method used to replace the discarded sample at each likelihood constraint is slightly different in each example. In both examples we use some form of Metropolis MCMC, which takes a sample at random from the surviving collection of live samples and moves it randomly around the likelihood-constrained parameter space. The final location is used as the new live sample. Both examples use uniform priors, and the Metropolis acceptance criterion is then simple: if at any time the sample lands outside of the likelihood constraint, it is rejected; if it lands inside the likelihood constraint, it is accepted.

In the following examples, constrained prior exploration differs in how the samples are moved throughout the parameter space. For the eggcrate example and the twin Gaussian shell example (section 5.1 and 5.3), the moves are drawn from a standard normal distribution scaled by a factor that is adjusted at each step. Both parameters are changed simultaneously, and, at the end of each step, the step size scaling factor is increased or decreased so as to maintain a roughly even acceptance ratio.

For the multiple stationary frequency example (section 5.2), a Gaussian step is also used, but each parameter is varied individually. At each step of the exploration procedure a random ordering of the parameters is chosen, and the sample is moved in one dimension at a time. The acceptance rate is monitored separately for each parameter, and a scaling factor corresponding to each parameter is updated at the end of each MCMC step in order to maintain a roughly even acceptance ratio.

5.1. Eggcrate likelihood

The first example is the "eggcrate" toy problem from [19]. The posterior density in this problem is highly multimodal, so that a large number of live samples is necessary to estimate the evidence accurately and to sample the entire density. Our results for this example show that combining many independent nested sampling chains with few live samples gives a result equivalent to using one chain with many live samples in the presence of many modes.

The eggcrate function has two independent parameters, with joint prior

$$\pi(\boldsymbol{\Theta}) = \left(\frac{1}{10\pi}\right)^2 \mathbb{1}_{[0,10\pi]}(\Theta_1)\mathbb{1}_{[0,10\pi]}(\Theta_2). \tag{31}$$

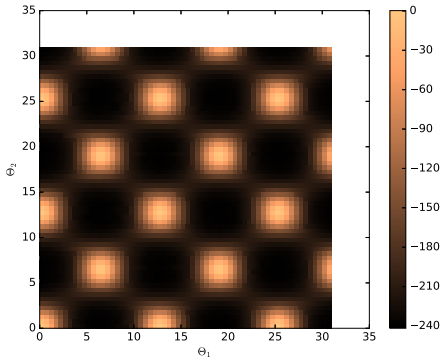
The likelihood function is

$$\mathcal{L}(\boldsymbol{\Theta}) = \exp\left\{\left[2 + \cos\left(\frac{\Theta_1}{2}\right)\cos\left(\frac{\Theta_2}{2}\right)\right]^5\right\}. \tag{32}$$

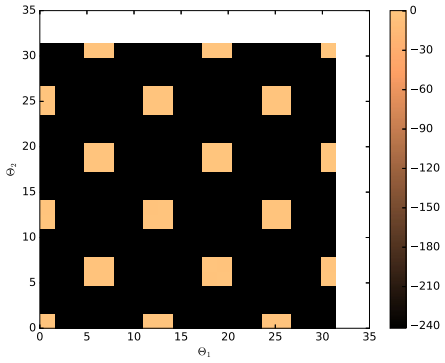
In this case the log-likelihood is more useful for visualization and the numerical dynamic range:

$$\log\mathcal{L}(\boldsymbol{\Theta}) = \left[2 + \cos\left(\frac{\Theta_1}{2}\right)\cos\left(\frac{\Theta_2}{2}\right)\right]^5. \tag{33}$$

Feroz et al. [19] provide an evidence value of $\log\mathcal{Z} = 235.88$ using numerical integration over a fine grid. Upon applying Bayes' theorem, the posterior distribution for the parameters $\boldsymbol{\Theta}$ is



(a) Log-posterior for the eggcrate problem (log of (34))



(b) Log pseudo-color plot of normalized histogram of re-sampled nested sampling results, 20x20 bin grid

Fig. 2. Eggcrate figures. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)

$$\mathcal{P}(\boldsymbol{\Theta}) = \frac{\exp\left\{\left[2+\cos\left(\frac{\Theta_1}{2}\right)\cos\left(\frac{\Theta_2}{2}\right)\right]^5\right\}\left(\frac{1}{10\pi}\right)^2 \mathbb{1}_{[0,10\pi]}(\Theta_1)\mathbb{1}_{[0,10\pi]}(\Theta_2)}{\exp(235.88)}. \tag{34}$$

5.1.1. Results

The results shown here were achieved using $N = 16$ live samples in each of $M = 20$ independent nested sampling runs. The uncertainty in the log-evidence estimate was estimated by performing 20 separate runs of the configuration with $N = 16$, $M = 20$. The estimated log-evidence value is $\log \mathcal{Z} = 235.84 \pm 0.1616$. This estimate is well within a single standard deviation of the log-evidence estimate of 235.88 given in [19]. Fig. 2a shows the log-posterior for the eggcrate function, and Fig. 2b shows a normalized log-histogram of the samples obtained using nested sampling, re-sampled using importance sampling as described by Goggans and Chi [22]. The normalized log-histogram in Fig. 2b shows that each mode of the posterior was sampled well.

5.2. Detection of multiple stationary frequencies

In the second example, we want to estimate the number of stationary frequencies present in a signal as well as the value of each frequency. This problem is similar to the problem of multiple stationary frequency estimation in [28, Chapter 6], with the additional task of determining the number of stationary frequencies present. This example demonstrates the value of the present parallel nested sampling method. Differences among log-evidence values for models containing not less than the most probable number of frequencies tend to be small, meaning that a precise estimate of these log-evidence values is essential to the task of determining the most probable model.

Each stationary frequency (j) in the model is determined by three parameters: the in-phase amplitude (A_j), the quadrature amplitude (B_j), and the frequency (f_j). Given J stationary frequencies, the model at time step t_i takes the following form:

$$g(t_i;\boldsymbol{\Theta}) = \sum_{j=1}^{J} A_j \cos(2\pi f_j t_i) + B_j \sin(2\pi f_j t_i), \tag{35}$$

where $\boldsymbol{\Theta}$ is the parameter vector

$$\boldsymbol{\Theta} = \left[A_1\, B_1\, f_1 \cdots A_J\, B_J\, f_J\right]^T.$$

For the purposes of this example the noise variance used to generate the simulated data is known, and we consequently use a Gaussian likelihood function,

Table 1
Prior bounds for multiple stationary frequency model parameters.

	Lower bound	Upper bound
A_j	-2	2
B_j	-2	2
f_j	0 Hz	6.4 Hz

Table 2
Parameters used to generate simulated signal.

j	A_j	B_j	f_j (Hz)
1	1.0	0.0	3.1
2	1.0	0.0	5.9

$$\mathcal{L}(\boldsymbol{\Theta}) = \prod_{i=1}^{I} \exp\left\{-\frac{[g(t_i;\boldsymbol{\Theta}) - d_i]^2}{2\sigma^2}\right\}, \tag{36}$$

for I simulated data d_i and noise variance σ^2. The log-likelihood function is then

$$\log \mathcal{L}(\boldsymbol{\Theta}) = -\sum_{i=1}^{I} \frac{[g(t_i;\boldsymbol{\Theta}) - d_i]^2}{2\sigma^2}. \tag{37}$$

Each model parameter is assigned a uniform prior distribution with limits as shown in Table 1.

Our test signal is a sum of two sinusoids, and zero-mean Gaussian noise with variance $\sigma^2 = 0.01$. This signal is sampled at randomly-spaced instants of time, in order to demonstrate that this time-domain method does not require uniform sampling to perform spectrum estimation. Bretthorst [29] demonstrates that the Nyquist critical frequency in the case of nonuniform sampling is $1/2\Delta T'$, where $\Delta T'$ is the dwell time. The dwell time is not defined for arbitrary-precision time values as used in this example, so we must choose another limiting value. A more conservative limit is given by $1/10\Delta T_{\text{avg}}$, where ΔT_{avg} is the average spacing between time steps, 1/64 s. This formulation yields a prior maximum limit of 6.4 Hz, as shown in Table 1. The parameters used to generate the simulated data are shown in Table 2.

The samples from the signal with noise are shown in Fig. 3.

5.2.1. Results

Using the data shown in Fig. 3, we estimate log-evidence values under the assumption that 1, 2, 3, and 4 sinusoids are present. The log-evidence is estimated using nested sampling with three sets of algorithm parameters: $N = 20$ and $M = 1$; $N = 200$ and

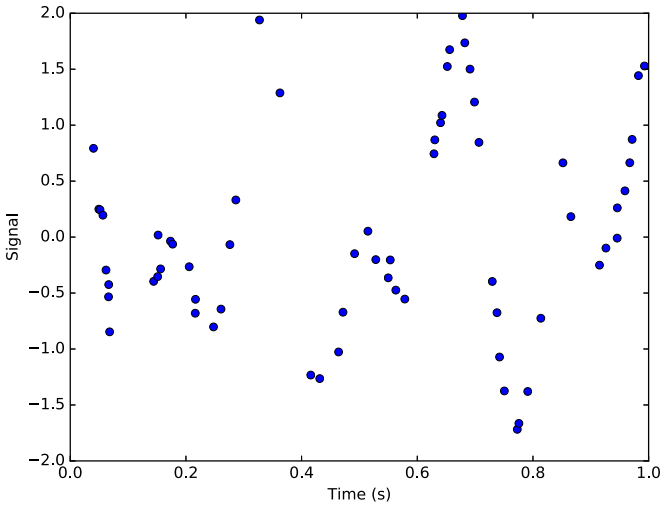


Fig. 3. The simulated signal. The points represent the non-uniformly sampled points from the original signal corrupted by Gaussian noise.

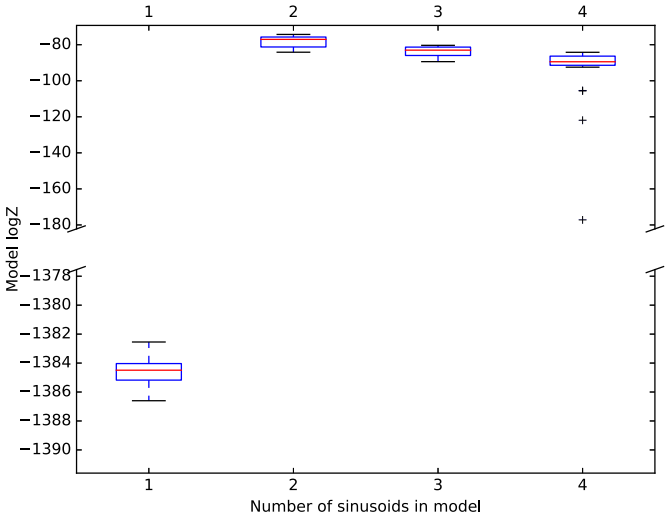


Fig. 4. Model log-evidence, $N = 20$, $M = 1$, 20 tests each.

$M = 1$; and $N = 50$ and $M = 4$. For comparison, we also include results from the old method of parallelizing nested sampling, using multiple replacement of live samples with $N = 400$ and $R = 4$. For each of the proposed models and algorithm parameter choices we conduct 20 separate nested sampling runs in order to quantify empirically the precision of the log-evidence estimate associated with each set of algorithm parameters. Box and whisker plots are shown in Figs. 4, 5, and 6 so as to summarize the log-evidence results from each of these nested sampling runs. A box and whisker plot is also shown in Fig. 7 for the log-evidence results for 20 runs each of nested sampling using the old parallel algorithm. In these box and whisker plots, the center line represents the median, the top and bottom of the box represents the third and first quartiles, the ends of the whiskers represent the maxima and minima of the observed values that fall within 1.5 times the interquartile range, and any values that fall outside that range are plotted as plus signs. The ordinate axis in each of these plots is truncated, so that the log-evidence values for each proposed model can be clearly displayed. For the 2-sinusoid model, one outlier is excluded from the plot in Fig. 4 because it is some 10^2 standard deviations away from the mean.

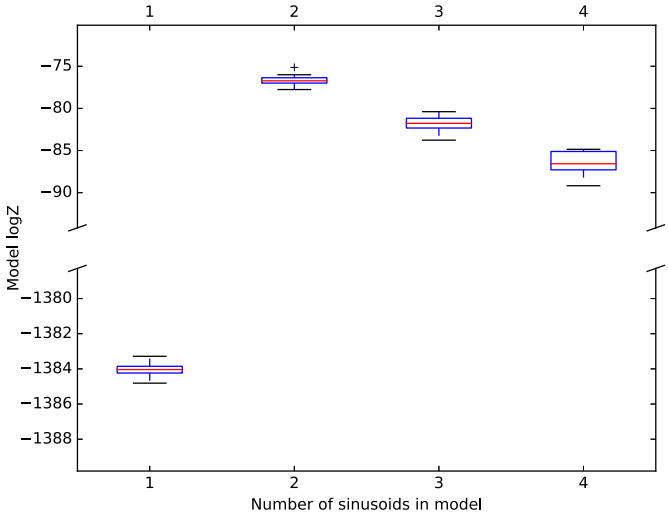


Fig. 5. Model log-evidence, $N = 200$, $M = 1$, 20 tests each.

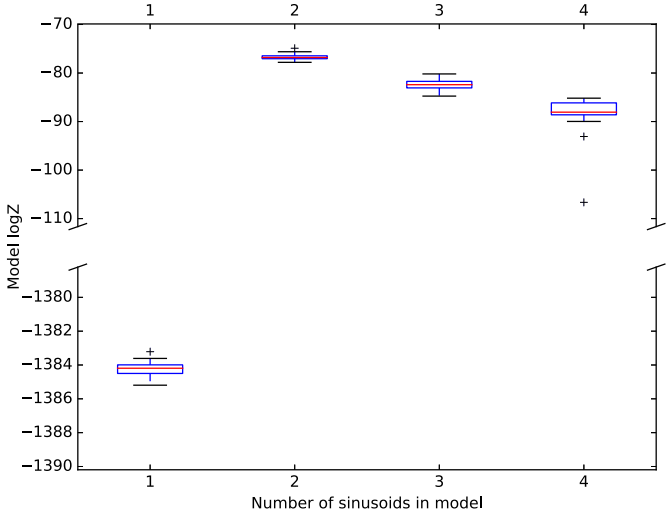


Fig. 6. Model log-evidence, $N = 50$, $M = 4$, 20 tests each.

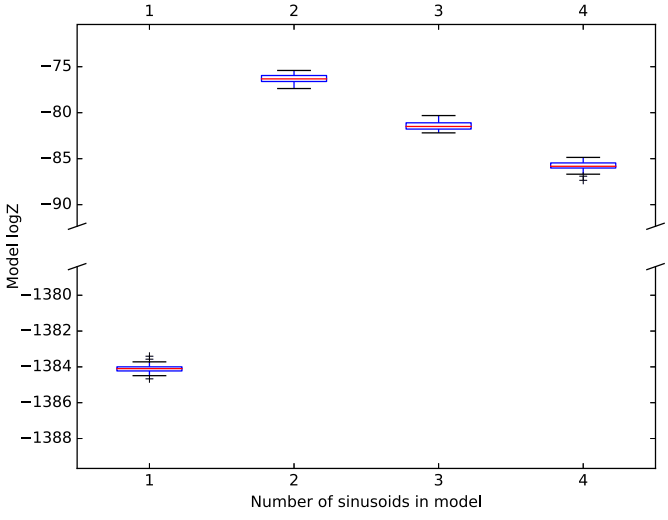


Fig. 7. Model log-evidence, $N = 400$, $R = 4$, 20 tests each, using the old parallel nested sampling algorithm.

Table 3
Multiple stationary frequency log-evidence for the model with 2 sinusoids, for each set of algorithm parameters. One outlier is excluded from the mean and standard deviation values for $N = 20$, $M = 1$.

Algorithm	N	M	R	Mean logZ	StDev logZ
New Parallel NS	20	1	–	−78.01	3.135
”	200	1	–	−76.68	0.5466
”	50	4	–	−76.72	0.6665
Old Parallel NS	400	–	4	−76.37	0.5512

Table 3 shows the mean and standard deviation of the log-evidence results for the 2-sinusoid model according to the trials using various parameter settings.

In the results for one run of nested sampling with $N = 20$ live samples (Fig. 4), the log-evidence estimates for the models with 2, 3, and 4 sinusoidal components overlap to some extent. In usual practice we would only wish to perform one run of nested sampling to obtain a log-evidence estimate, but these results demonstrate that $N = 20$ does not yield sufficient precision that would let us determine the maximum a-posteriori (MAP) model from a single run. Fig. 5 demonstrates that, by raising the number of live samples to $N = 200$, a level of precision is obtained that picks out model 2 as the clear MAP model, with no overlap with model 3 or model 4. Fig. 6 shows that upon combining the results of $M = 4$ independent runs of nested sampling, each using $N = 50$ live samples, we obtain similar precision to that in the case with $M = 1$ and $N = 200$.

Table 3 summarizes the mean and standard deviation for the log-evidence for the 2-sinusoid model in the MSF problem. The results show that the standard deviation in the log-evidence estimates for the correct model are similar for: the $N = 200$, $M = 1$ case; the $N = 50$, $M = 4$ case; and the $N = 400$, $R = 4$ case. The standard deviation is slightly greater for the $N = 50$, $M = 4$ case, possibly indicating a less efficient likelihood-constrained prior exploration process with fewer live samples per worker. The difference is slight, however, and the speed-up apparent in the $N = 50$, $M = 4$ case supports the use of the new parallel approach in this application.

As mentioned in section 3, the run time of nested sampling increases linearly with the number of live samples. So, because we can obtain similar precision from 4 separate runs with $N = 50$ and 1 run with $N = 200$, we obtain a nearly four-fold speed increase by using the combined-chain nested sampling technique.

5.3. Twin Gaussian shells

The final example is the twin Gaussian shell problem, also from [19]. In [19], the authors present results for this problem in up to 30 dimensions. Handley et al. [20] also use this problem in 100 dimensions to test their algorithm. This problem presents a few interesting challenges to our nested sampling implementation. Because the likelihood takes the form of a thin, curved density whose mass centers on a hyper-spherical shell, exploration of the constrained prior at high likelihood values is difficult. The bimodal nature of the problem also challenges the constrained prior exploration process. Finally, the examples we explore are high-dimensional to the point that standard numerical integration techniques would be useless.

The likelihood function in the twin Gaussian shells problem takes the form,

$$\mathcal{L}(\boldsymbol{\Theta}) = \frac{1}{\sqrt{2\pi}\, w_1} \exp\left[-\frac{(|\boldsymbol{\Theta} - \mathbf{c}_1| - r_1)^2}{2w_1^2} \right] + \frac{1}{\sqrt{2\pi}\, w_2} \exp\left[-\frac{(|\boldsymbol{\Theta} - \mathbf{c}_2| - r_2)^2}{2w_2^2} \right]. \tag{38}$$

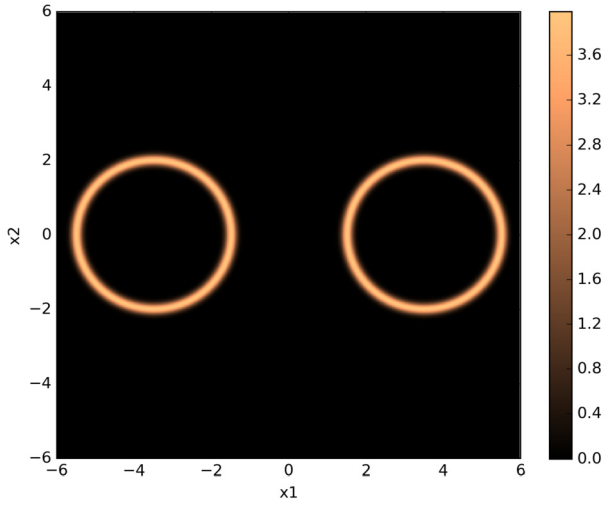


Fig. 8. Pseudo-color plot of a 2-dimensional twin Gaussian shell with $w_1 = w_2 = 0.1$, $r_1 = r_2 = 2$, $\mathbf{c}_1 = [-3.5, 0]^T$, and $\mathbf{c}_2 = [3.5, 0]^T$. The color values correspond to likelihood values. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 4
20- and 30-dimension twin Gaussian shell log-evidence results for each algorithm setting, 20 runs each. The mean, standard deviation, and root mean squared error (RMSE) with respect to the analytic value given in [19] are shown. The analytic log-evidence value is −36.09 for 20-D and −60.13 for 30-D.

ndim	N	M	Mean logZ	StDev logZ	RMSE
20	200	1	−36.23	0.4130	0.4349
20	50	4	−36.09	0.3341	0.3341
30	300	1	−60.22	0.4984	0.5067
30	75	4	−60.09	0.4100	0.4120

Following [19], we set the parameters as follows: $w_1 = w_2 = 0.1$, $r_1 = r_2 = 2$, $\mathbf{c}_1 = [-3.5, 0, \cdots, 0]^T$, and $\mathbf{c}_2 = [3.5, 0, \cdots, 0]^T$. We use a uniform prior over the hypercube that spans $[-6, 6]$ in each dimension. Fig. 8 shows a pseudo-color plot of a 2-dimensional twin Gaussian shell with parameters and prior range as described previously.

5.3.1. Results

We tested our nested sampling algorithm using this problem in 20 and 30 dimensions. The results are presented in this section.

We performed 20 tests each using $N = 200, M = 1$ and $N = 50, M = 4$ for the 20-dimensional case and $N = 300, M = 1$ and $N = 75, M = 4$ for the 30-dimensional case. The results from these tests are shown in Table 4. The analytic log-evidence value given in [19] for the 20-D and 30-D twin Gaussian shell problems are −36.09 and −60.13. The results in Table 4 show that both algorithm settings produce results within one standard deviation of the true value for both problems. Additionally, the results for $ndim = 20$, $N = 50$, $M = 4$ have a mean value that is exactly correct up to four significant digits and have standard deviation and root mean square error values that are noticeably less than those for the $N = 200$, $M = 1$ setting. While the results for $ndim = 30$, $N = 75$, $M = 4$ do not have a mean that is exactly correct, the mean is much closer to the analytic value than that in the serial configuration, and the RMSE is again significantly less in the parallel configuration than in the serial configuration. The reason for the decrease in error for the parallel configuration versus the serial configuration could be that the parallel configuration somehow compensates for imperfect sampling of the prior distribution within each likelihood contour in the serial configuration. These results demonstrate that our algorithm can handle distributions

with curving degeneracies, multi-modal distributions, and high-dimensional problems at least as well as standard (serial) nested sampling, all while providing a significant speed-up on parallel hardware.

6. Conclusion

The nested sampling algorithm, as conceived originally, is implemented in a strictly serial fashion. By combining the samples produced by multiple independent runs of nested sampling, we have developed a method for efficiently performing Bayesian model selection in a way that takes advantage of modern parallel computing architectures. We have given the mathematical foundation for estimating the evidence from these combined chains, as well as several ideas for practical implementation of the method. Three examples have demonstrated the utility and effectiveness of the technique for a variety of problem types, including a problem with a highly multi-modal distribution, a data analysis problem, and a problem with a distribution that is high-dimensional, curving, and multi-modal.

Combined-chain nested sampling is a generalized technique that is capable of providing the foundation for other, potentially more specialized methods. Future work could adapt current techniques that involve nested sampling and make use of the present combined-chain method.

Acknowledgment

The authors thank Dr. John Skilling for inspiring this project in a conversation at MaxEnt 2013 in Canberra, Australia. We are also grateful to Dr. John Daigle of the University of Mississippi for his help in developing the mathematical justification for our technique in section 3. This work was supported in part by NASA EPSCoR program under grant NNX14N38A.

Appendix A. Proof of Lemma 2

Proof. $M_1(t)$ and $M_2(t)$ are discrete random variables parameterized by t. A necessary condition for (26) to be satisfied is therefore that the p.m.f. of the RV on the left is equivalent to the p.m.f. of the RV on the right.

Let $\{N_1(s);\, s>0\}$ and $\{N_2(s);\, s>0\}$ be Poisson counting processes with rates λ_1 and λ_2. Let $N(s)=N_1(s)+N_2(s)$. The probability mass functions for $N_1(s)$ and $N_2(s)$ are given as

$$p_{N_1(s)}(n_1)=\frac{(\lambda_1 s)^{n_1}\exp(-\lambda_1 s)}{n_1!} \tag{A.1}$$

$$p_{N_2(s)}(n_2)=\frac{(\lambda_2 s)^{n_2}\exp(-\lambda_2 s)}{n_2!}. \tag{A.2}$$

The distribution of $N(s)$, which comprises the sum of $N_1(s)$ and $N_2(s)$, is given as a convolution of the distributions for $N_1(s)$ and $N_2(s)$:

$$p_{N(s)}(n)=\sum_{k=-\infty}^{\infty} p_{N_1(s)}(k)\,p_{N_2(s)}(n-k). \tag{A.3}$$

These are counting processes, which are zero-valued unless $k\geq 0$ and $n-k\geq 0$. Thus, $0\leq k\leq n$, and the convolution sum (A.3) becomes

$$p_{N(s)}(n)=\sum_{k=0}^{n} p_{N_1(s)}(k)\,p_{N_2(s)}(n-k) \tag{A.4}$$

$$=\sum_{k=0}^{n}\frac{(\lambda_1 s)^k\exp(-\lambda_1 s)}{k!}\,\frac{(\lambda_2 s)^{n-k}\exp(-\lambda_2 s)}{(n-k)!} \tag{A.5}$$

$$=\sum_{k=0}^{n}\frac{(\lambda_1/\lambda_2)^k(\lambda_2 s)^n\exp[-(\lambda_1+\lambda_2)s]}{k!(n-k)!} \tag{A.6}$$

$$=(\lambda_2 s)^n\exp[-(\lambda_1+\lambda_2)s]\sum_{k=0}^{n}\frac{(\lambda_1/\lambda_2)^k}{k!(n-k)!} \tag{A.7}$$

$$=(\lambda_2 s)^n\exp[-(\lambda_1+\lambda_2)s]\,\frac{[(\lambda_1+\lambda_2)/\lambda_2]^n}{n!} \tag{A.8}$$

$$=\frac{[s(\lambda_1+\lambda_2)]^n\exp[-(\lambda_1+\lambda_2)s]}{n!}. \tag{A.9}$$

This result establishes the right-hand side of (26).

Now carry out the inverse transform $\mathcal{H}^{-1}\{\cdot\}$ on (A.1) and (A.2):

$$p_{M_1(t)}(m_1)=\frac{(-\lambda_1\log(t))^{m_1}\exp(\lambda_1\log(t))}{m_1!}=\frac{(-\lambda_1\log(t))^{m_1}\,t^{\lambda_1}}{m_1!} \tag{A.10}$$

$$p_{M_2(t)}(m_2)=\frac{(-\lambda_2\log(t))^{m_2}\exp(\lambda_2\log(t))}{m_2!}=\frac{(-\lambda_2\log(t))^{m_2}\,t^{\lambda_2}}{m_2!}. \tag{A.11}$$

Let $M(t)=M_1(t)+M_2(t)$, and find the p.m.f. for $M(t)$ as in (A.3). The result is

$$p_{M(t)}(m)=\frac{(-1)^m(\lambda_1+\lambda_2)^m(\log t)^m t^{\lambda_1+\lambda_2}}{m!}. \tag{A.12}$$

Carry out the transform

$$p_{\mathcal{H}\{M(t)\}}(m)=p_{N(s)}(n)=\frac{[s(\lambda_1+\lambda_2)]^n\exp[-(\lambda_1+\lambda_2)s]}{n!} \tag{A.13}$$

(A.9) and (A.13) are identical, so we have shown that

$$p_{\mathcal{H}\{M_1(t)+M_2(t)\}}(m)=p_{\mathcal{H}\{M_1(t)\}+\mathcal{H}\{M_2(t)\}}(m). \tag{A.14}$$

As $M(t)$ and $N(s)$ represent random processes, their p.m.f.s are actually probability mass *functionals*, not functions. Let $f(t)=t^{\lambda_1+\lambda_2}$ and $g(s)=(\lambda_1+\lambda_2)s$. We can express these functionals as

$$p_{M(t)}(m)=F_1[f(t)]=\frac{(-1)^m(\log f(t))^m f(t)}{m!} \tag{A.15}$$

$$p_{N(s)}(n)=F_2[g(t)]=\frac{[g(s)]^n\exp[-g(s)]}{n!}. \tag{A.16}$$

The mean and autocorrelation functions of these mass functionals can depend *only* on the functions $f(t)$ and $g(s)$. Since $f(t)$ is the same on both sides of (26), as we have shown in (A.9) and (A.13), it follows that the mean and autocorrelation functions are also equivalent, and that:

$$\mathcal{H}\{M_1(t)+M_2(t)\}=\mathcal{H}\{M_1(t)\}+\mathcal{H}\{M_2(t)\}. \tag{A.17}$$

We can therefore say that $\mathcal{H}\{\cdot\}$ is linear in addition. □

References

[1] K.H. Knuth, M. Habeck, N.K. Malakar, A.M. Mubeen, B. Placek, Bayesian evidence and model selection, Digit. Signal Process. 47 (2015) 50–67, http://dx.doi.org/10.1016/j.dsp.2015.06.012, arXiv:1411.3013v1.

[2] C.-Y. Chan, P.M. Goggans, Using Bayesian inference for the design of FIR filters with signed power-of-two coefficients, Signal Process. 92 (12) (2012) 2866–2873, http://dx.doi.org/10.1016/j.sigpro.2012.05.009.

[3] J. Botts, J. Escolano, N. Xiang, Design of IIR filters with Bayesian model selection and parameter estimation, IEEE Trans. Audio Speech Lang. Process. 21 (3) (2013) 669–674, http://dx.doi.org/10.1109/tasl.2012.2226159.

[4] N. Xiang, P.M. Goggans, Evaluation of decay times in coupled spaces: Bayesian parameter estimation, J. Acoust. Soc. Am. 110 (3) (2001) 1415–1424.

[5] N. Xiang, P.M. Goggans, Evaluation of decay times in coupled spaces: Bayesian decay model selection, J. Acoust. Soc. Am. 113 (5) (2003) 2685–2697.
[6] N. Xiang, P.M. Goggans, T. Jasa, M. Kleiner, Evaluation of decay times in coupled spaces: reliability analysis of Bayesian decay time estimation, J. Acoust. Soc. Am. 117 (6) (2005) 3707–3715.
[7] N. Xiang, P. Goggans, T. Jasa, P. Robinson, Bayesian characterization of multiple-slope sound energy decay in coupled-volume systems, J. Acoust. Soc. Am. 129 (2) (2011) 741–752.
[8] W. Henderson, P. Goggans, N. Xiang, J. Botts, Bayesian inference approach to room-acoustic modal analysis, AIP Conf. Proc. 1553 (1) (2013) 38–45, http://dx.doi.org/10.1063/1.4819981.
[9] P.M. Goggans, R.W. Henderson, N. Xiang, Using nested sampling with Galilean Monte Carlo for model comparison problems in acoustics, Proc. Meet. Acoust. 19 (1) (2013) 055089, http://dx.doi.org/10.1121/1.4800876.
[10] C.J. Fackler, Bayesian Model Selection for Analysis and Design of Multilayer Sound Absorbers, Ph.D. dissertation, Rensselaer Polytechnic Institute, Troy, New York, October 2014.
[11] J. Skilling, Nested sampling, AIP Conf. Proc. 735 (1) (2004) 395–405, http://dx.doi.org/10.1063/1.1835238.
[12] J. Skilling, Nested sampling for general Bayesian computation, Bayesian Anal. 1 (4) (2006) 833–859.
[13] J. Skilling, Nested sampling's convergence, AIP Conf. Proc. 1193 (1) (2009) 277–291.
[14] D.S. Sivia, J. Skilling, Data Analysis: A Bayesian Tutorial, 2nd edition, Oxford University Press, 2006.
[15] N. Burkoff, C. Vàrnai, S. Wells, D. Wild, Exploring the energy landscapes of protein folding simulations with Bayesian computation, Biophys. J. 102 (4) (2012) 878–886, http://dx.doi.org/10.1016/j.bpj.2011.12.053.
[16] S. Martiniani, J.D. Stevenson, D.J. Wales, D. Frenkel, Superposition enhanced nested sampling, Phys. Rev. X 4 (3) (2014) 031034, http://dx.doi.org/10.1103/PhysRevX.4.031034.
[17] R.W. Henderson, P.M. Goggans, Parallelized nested sampling, AIP Conf. Proc. 1636 (1) (2014) 100–105, http://dx.doi.org/10.1063/1.4903717.
[18] B.J. Brewer, L.B. Pártay, G. Csányi, Diffusive nested sampling, Stat. Comput. 21 (4) (2011) 649–656, http://dx.doi.org/10.1007/s11222-010-9198-8.
[19] F. Feroz, M.P. Hobson, M. Bridges, Multinest: an efficient and robust Bayesian inference tool for cosmology and particle physics, Mon. Not. R. Astron. Soc. 398 (4) (2009) 1601–1614, http://dx.doi.org/10.1111/j.1365-2966.2009.14548.x.
[20] W. Handley, M. Hobson, A. Lasenby, Polychord: next-generation nested sampling, Mon. Not. R. Astron. Soc. 453 (4) (2015) 4384–4398, http://dx.doi.org/10.1093/mnras/stv1911.
[21] P. Gregory, Bayesian Logical Data Analysis for the Physical Sciences, paperback edition, Cambridge University Press, 2010.
[22] P.M. Goggans, Y. Chi, Using thermodynamic integration to calculate the posterior probability in Bayesian model selection problems, AIP Conf. Proc. 707 (1) (2004) 59–66, http://dx.doi.org/10.1063/1.1751356.
[23] S.M. Ross, Introduction to Probability Models, tenth edition, Academic Press, 2010.
[24] E.W. Weisstein, Erlang distribution, http://mathworld.wolfram.com/ErlangDistribution.html. (Accessed 22 February 2016), from Mathworld – a Wolfram web resource.
[25] H.M. Taylor, S. Karlin, An Introduction to Stochastic Modeling, 3rd edition, Academic Press, 1998.
[26] H.C. Tijms, A First Course in Stochastic Models, John Wiley & Sons, 2003.
[27] J. Skilling, Bayesian computation in big spaces – nested sampling and Galilean Monte Carlo, AIP Conf. Proc. 1443 (1) (2012) 145–156.
[28] G.L. Bretthorst, Bayesian Spectrum Analysis and Parameter Estimation, Springer-Verlag, Berlin, Heidelberg, 1988, http://bayes.wustl.edu/glb/book.pdf.
[29] G.L. Bretthorst, Nonuniform sampling: bandwidth and aliasing, AIP Conf. Proc. 567 (1) (2001) 1–28, http://dx.doi.org/10.1063/1.1381847.

R. Wesley Henderson is a Ph.D. student in the electrical engineering program at the University of Mississippi, working with adviser Prof. Paul Goggans. His research interests include the development and application of numerical Bayesian inference techniques. He graduated from Rensselaer Polytechnic Institute in 2012 with a Master of Science degree in Architectural Sciences, with concentration on acoustics, under the guidance of adviser Prof. Ning Xiang. He graduated from Louisiana Tech University in 2011 with a Bachelor of Science degree in Civil Engineering.

Paul M. Goggans received the Ph.D. degree from Auburn University, Auburn, AL, USA, in 1990. In that same year he was appointed Assistant Professor in the Department of Electrical Engineering at the University of Mississippi, Oxford, MS, USA, where he is currently a Professor. His research interests include the application of Bayesian inference to engineering model-comparison, parameter-estimation, and design problems, Markov chain Monte Carlo methods, and computational electromagnetics.

Lei Cao received his B.E. degree in Electrical Engineering from Hefei University of Technology in 1990, his M.S. degree in Computer Science from University of Science and Technology of China in 1993, and his Ph.D. degree in Electrical Engineering from University of Missouri-Columbia in 2002. He was a lecturer of Electronic Engineering at University of Science and Technology of China from 1993 to 1998. He has been with The University of Mississippi since 2002 and now he is a professor of Electrical Engineering. His research interests include source and channel coding, wireless communications and applied probability modeling.