Inference for Trans-dimensional Bayesian Models with Diffusive Nested Sampling

Brendon J. Brewer

November 12, 2014

Abstract

Many inference problems involve inferring the number N of objects in some region, along with their properties $\{\mathbf{x}_i\}_{i=1}^N$, from a dataset $\mathcal{D}$. A common statistical example is finite mixture modelling. In the Bayesian framework, these problems are typically solved using one of the following two methods: i) by executing a Monte Carlo algorithm (such as Nested Sampling) once for each possible value of N, and calculating the marginal likelihood or evidence as a function of N; or ii) by doing a single run that allows the model dimension N to change (such as Markov Chain Monte Carlo with birth/death moves), and obtaining the posterior for N directly. In this paper we present a general approach to this problem that uses trans-dimensional MCMC embedded *within* a Nested Sampling algorithm, allowing us to explore the posterior distribution and calculate the marginal likelihood (summed over N) even if the problem contains a phase transition or other difficult features such as multimodality. We present two example problems, finding sinusoidal signals in noisy data, and finding and measuring galaxies in a noisy astronomical image. Both of the examples demonstrate phase transitions in the relationship between the likelihood and the cumulative prior mass.

1 Introduction

Consider the following class of inference problems. There is some unknown number, N, of objects in a (physical or metaphorical) region. Each of the N objects has a property $\mathbf{x}$, which may be a single scalar value (for example, a mass), or a set of values (e.g. a two-dimensional position and a mass). These problems are often challenging to solve due to the unknown (and potentially large) dimensionality. In addition, they also have the so-called "label-switching degeneracy" issue (Jasra et al, 2005), where the meaning of the model is invariant to switching the identities of the objects. Finite mixture modelling is a common statistical example.

To carry out the inference, we must assign a prior to the objects' properties $\{\mathbf{x}_i\}_{i=1}^N$. Since N may be large, it is usually easier to assign an "interim prior" conditional on some hyperparameters $\boldsymbol{\alpha}$, and then assign a prior to $\boldsymbol{\alpha}$. This kind of model is usually called *hierarchical*. The prior for N, $\boldsymbol{\alpha}$, and $\{\mathbf{x}_i\}$ is usually factorised in the following way:

$$p(N, \boldsymbol{\alpha}, \{\mathbf{x}_i\}) = p(N)p(\boldsymbol{\alpha}|N)p(\{\mathbf{x}_i\}|\boldsymbol{\alpha}, N) \quad (1)$$

$$= p(N)p(\boldsymbol{\alpha})\prod_{i=1}^{N} p(\mathbf{x}_i|\boldsymbol{\alpha}). \quad (2)$$

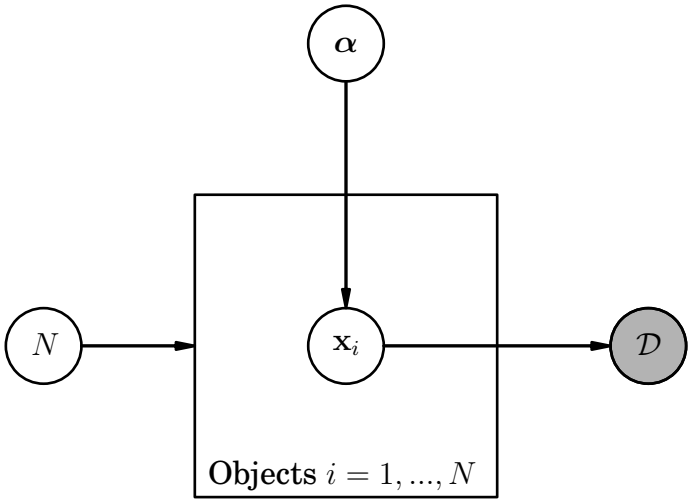

Figure 1: *A probabilistic graphical model (PGM) depicting the kind of model discussed in this paper. Produced using daft (*`daft-pgm.org`*). The number, N, of objects in the model is an unknown parameter, as are the properties of the objects. The data depends on the properties of the objects. Note that it is possible (and common) to have other parameters in the model that point directly to the data. However, we have omitted such parameters for simplicity.*

Here we have assumed the priors for N and $\boldsymbol{\alpha}$ are independent, and the interim prior for $\{\mathbf{x}_i\}$ is iid and does not depend on N. Given data $\mathcal{D}$, we usually want to calculate properties of the posterior distribution, given by:

$$p(N, \boldsymbol{\alpha}, \{\mathbf{x}_i\}|\mathcal{D}) \propto p(N, \boldsymbol{\alpha}, \{\mathbf{x}_i\})p(\mathcal{D}|N, \boldsymbol{\alpha}, \{\mathbf{x}_i\}) \tag{3}$$

The structure of the model is depicted graphically in Figure 1.

The computational method used to calculate properties of the posterior distribution is to generate samples from it using Markov Chain Monte Carlo (MCMC). Since N is unknown, an MCMC sampler needs to be able to jump between candidate solutions with different numbers of objects, using either birth-death MCMC (Stephens, 2000) or the reversible jump framework of Green (1995).

The marginal likelihood, or evidence, is also an important quantity, and is the normalisation constant of the posterior given in Equation 3:

$$\mathcal{Z} = p(\mathcal{D}) \tag{4}$$

$$= \sum_{i=0}^{N_{\max}} \int p(N, \boldsymbol{\alpha}, \{\mathbf{x}_i\})p(\mathcal{D}|N, \boldsymbol{\alpha}, \{\mathbf{x}_i\})\, d^N\mathbf{x}_i\, d\boldsymbol{\alpha} \tag{5}$$

Some authors have addressed these kinds of problems by doing separate runs of Nested Sampling (or another method), with N fixed at various trial values (e.g. Hou et al., 2014; Feroz and Hobson, 2014). Then, the posterior for N can be calculated based on the estimates of the evidence

obtained as a function of N. This approach does not generalise well to large N, as it would require a large number of parallel runs.

In this paper, we introduce an approach that uses trans-dimensional MCMC moves to infer N, but embeds this process in the Nested Sampling framework to overcome difficulties that would be encountered by sampling the posterior distribution directly. Our motivation for using Nested Sampling is not primarily to calculate the marginal likelihood $\mathcal{Z}$ (including summing over N, as in Equation 5), although $\mathcal{Z}$ is readily available from the output. The motivation for using Nested Sampling, rather than just sampling the posterior using trans-dimensional MCMC, is that the posterior often contains difficult features, such as strong dependencies or multiple modes, that cause problems with mixing. Diffusive Nested Sampling (DNS Brewer, Pártay, & Csányi, 2011b) replaces the posterior distribution with an alternative target distribution composed of a mixture of the prior distribution with other, more constrained versions of the prior, facilitating mixing between the multiple modes and along degeneracy curves. This is similar to the idea of annealed importance sampling (Neal, 2001) or parallel tempering (Hansmann, 1997), where the target distribution is modified from the posterior to something "easier". However, the sequence of distributions used in Nested Sampling avoids some of the problems with annealing-based methods. In particular, there is no need to choose a temperature schedule (a potentially large set of tuning parameters), and the method does not fail when phase transitions are present.

Since DNS is effectively the Metropolis-Hastings algorithm applied to a distribution other than the posterior, proposal distributions are needed which define how a walker moves around the hypothesis space. In Section 2, we outline a set of proposal distributions used to explore the hypothesis space of possible values for N, $\boldsymbol{\alpha}$, and $\{\mathbf{x}_i\}$.

A C++ implementation of the ideas described in this paper is available online (under the terms of the GNU General Public Licence) at `https://github.com/eggplantbren/RJObject`.

2 Metropolis proposals for the general problem

We will now define a set of proposal distributions that can be used to sample the prior distribution $p(N, \boldsymbol{\alpha}, \{\mathbf{x}_i\}) = p(N)p(\boldsymbol{\alpha})\prod_{i=1}^{N} p(\mathbf{x}_i|\boldsymbol{\alpha})$. These proposals will need to satisfy detailed balance in order to be valid when used inside DNS. When the proposal satisfies detailed balance with respect to the prior, the DNS algorithm can incorporate hard likelihood constraints by rejecting any proposal whose likelihood does not satisfy the constraint. The overall proposal is a mixture of all of the proposals listed here.

When using DNS, the proposals typically should be very heavy tailed. When DNS is exploring a distribution close to the prior, large proposals will generally be needed to explore the prior efficiently. When DNS is exploring a distribution constrained to very high values of the likelihood function, much smaller proposals will be appropriate. Rather than attempting expensive tuning (since the number of distributions involved may number in the hundreds or even thousands), we will apply the heavy tailed proposal distributions and simply accept that there will be some waste. If a parameter u has a uniform prior between 0 and 1, then a particular choice of heavy-tailed proposal is defined by:

$$u' \quad := \quad \mathrm{mod}\left(u + 10^{1.5-6a}b, 1\right) \tag{6}$$

where $a \sim \mathrm{Uniform}(0, 1)$ and $b \sim \mathcal{N}(0, 1)$. This is like a standard gaussian/normal random walk proposal, but with a non-fixed step size, resulting in a scale mixture of normals. When a is close to zero, the size of the proposal is of order 10, which (because of the mod function) effectively randomises u' from the prior. If a is close to 1, the scale of the proposal is roughly 10^{-4} times

the prior width. The coefficient of a was set to 6 because smaller proposals are usually not necessary. The value 1.5 in the exponent was found to optimize the performance of this proposal when sampling from $U(0,1)$ and $\mathcal{N}(0,1)$ priors.

Most of the proposals discussed in this paper involve changing one (or a subset) of the parameters and/or hyperparameters, while keeping the others fixed. Whenever a heavy-tailed proposal distribution is required, we use the proposal from Equation 6, or an analogous one, e.g. a discrete version when the proposal is to change N.

2.1 Proposals that modify N

The first kind of proposal we consider are proposals that change the dimension of the model, i.e. proposals that change the value of N. By necessity, these will also change the parameters $\{\mathbf{x}_i\}_{i=1}^{N}$ because the number of parameters will be changed. The proposals used here are traditionally called *birth and death* proposals.

We will assume that the prior for N is a discrete uniform distribution between 0 and some $N_{\max}$, inclusive. The proposal starts by choosing a new value N' according to

$$N' = \mod(N + \delta_N, N_{\max} + 1) \tag{7}$$

where δ_N is drawn from a heavy tailed distribution which is a discrete analogue of Equation 6. The most probable values of δ_N are ± 1, but values of order $N_{\max}$ also have some probability, to allow fast exploration when the target distribution is similar to the prior. If $N' = N$, N' is regenerated as $N + 1$ with probability 0.5, and $N - 1$ with probability 0.5, and then wrapped back into the range $\{0, ..., N_{\max}\}$ if necessary.

If $N' > N$, i.e. the proposal is to add objects to the model, the extra parameters needed, $\{\mathbf{x}_i\}_{i=N+1}^{N'}$, are drawn from their prior conditional on the current value of the hyperparameters, i.e. the interim prior $p(\mathbf{x}|\boldsymbol{\alpha})$. If $N' < N$, i.e. the proposal is to remove objects from the model, then $N - N'$ objects must be selected for removal. All of the objects have the same probability of being selected for removal.

2.2 Proposals that modify objects

We now consider a proposal distribution that modifies one or more of the objects $\{\mathbf{x}_i\}$, while keeping the number of objects N, as well as the hyperparameters $\boldsymbol{\alpha}$, fixed.

Let $F(x; \boldsymbol{\alpha})$ be a function that takes an object x and transforms it to a value u that has a uniform distribution between 0 and 1, given $\boldsymbol{\alpha}$. If the object x consists of a single scalar value, F is the cumulative distribution of the interim prior. Denote the inverse of F by G.

A proposal for an object involves transforming its parameters to $[0,1]$ using F, making the proposal in terms of u, and then transforming back. Specifically, the proposal chooses a new value $\mathbf{x}_i'$ from the current value $\mathbf{x}_i$ as follows:

$$u_i := F(\mathbf{x}_i; \boldsymbol{\alpha}) \tag{8}$$

$$u_i' := \mod(u_i + \delta_u, 1) \tag{9}$$

$$\mathbf{x}_i' := G(u_i'; \boldsymbol{\alpha}). \tag{10}$$

where δ_u is drawn from a heavy-tailed distribution as in Equation 6.

Choosing just one object to change is most appropriate when the DNS distribution is very constrained. When it is close to the prior, bolder proposals that change more than one object at a time are possible. Hence, we choose the number of objects to change from a heavy tailed distribution wherer the most probable value is 1 but there is also a nontrivial probability of proposing to change $\sim N$ objects at once.

2.3 Proposals that change the hyperparameters, keeping the objects fixed

Another kind of proposal that we will consider is a proposal that keeps all of the objects fixed in place (i.e. leaves $\{\mathbf{x}_i\}$) unchanged, but changes the hyperparameter(s) from their current value $\boldsymbol{\alpha}$ to a new value $\boldsymbol{\alpha}'$. The proposal for the hyperparameter(s) is chosen to satisfy detailed balance with respect to $p(\boldsymbol{\alpha})$ and should be heavy-tailed.

The overall Metropolis acceptance probability for this kind of move, if sampling the prior (or the constrained prior of DNS) must include the following factor:

$$\frac{\prod_{i=1}^{N} p(\mathbf{x}_i|\boldsymbol{\alpha}')}{\prod_{i=1}^{N} p(\mathbf{x}_i|\boldsymbol{\alpha})} \tag{11}$$

Since this proposal leaves the objects $\{\mathbf{x}_i\}$ fixed, it will usually not affect the value of the likelihood, and therefore the likelihood will not need to be recomputed. Therefore this proposal is most useful for mixing the values of the hyperparameters $\boldsymbol{\alpha}$ when the target distribution is highly constrained.

2.4 Proposals that change the hyperparameters and all of the objects

The above proposals allow for changes to N, $\boldsymbol{\alpha}$, and $\{\mathbf{x}_i\}$, and are therefore sufficient to allow for "correct" exploration of the prior distribution, and the constrained prior distributions used in DNS. However, they do not necessarily allow for *efficient* exploration, even of the prior itself. The main reason for this is the inability for large changes to be made to the hyperparameters $\boldsymbol{\alpha}$. If the proposed change from $\boldsymbol{\alpha}$ to $\boldsymbol{\alpha}'$ is large, the ratio in Equation 11 is likely to be very small, and the move will probably be rejected.

Therefore, we allow an additional move that changes $\boldsymbol{\alpha}$ to a new value $\boldsymbol{\alpha}'$, but rather than leaving the objects $\{\mathbf{x}_i\}$ fixed, they are "dragged" so they represent the distribution $p(\mathbf{x}|\boldsymbol{\alpha}')$ rather than $p(\mathbf{x}|\boldsymbol{\alpha})$. We do this by making use of the transformation functions $F(x;\boldsymbol{\alpha})$ and $G(x;\boldsymbol{\alpha})$ defined in Section 2.2.

The "dragging" process works as follows, and must be carried out on each object:

$$u_i := F(\mathbf{x}_i;\boldsymbol{\alpha}) \tag{12}$$

$$\mathbf{x}'_i := G(\mathbf{x}_i;\boldsymbol{\alpha}') \tag{13}$$

In other words, the objects' parameters are transformed to $[0,1]$ using the current value of the hyperparameters, and then transformed back using the proposed value of the hyperparameters so they represent the new interim prior rather than the old one.

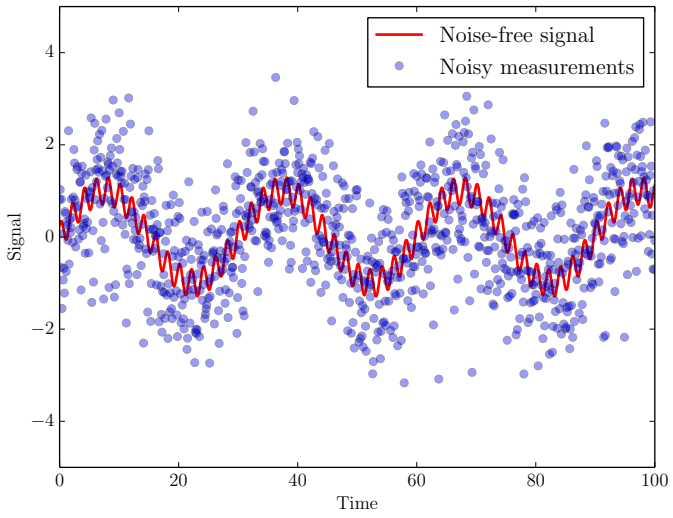

Figure 2: The simulated data for the sinusoidal example. The solid line shows the true signal and the points are the measurements, simulated based on a noise standard deviation of $\sigma = 1$.

3 Sinusoidal example

In this section we demonstrate a seemingly simple example which exhibits a phase transition, making a standard MCMC approach difficult. Suppose a signal is composed of N sinusoids, of different periods, amplitudes, and phases. The signal is observed at various times $\{t_i\}_{i=1}^N$ with noise, and we want to use the resulting data to infer the number of sinusoids N, along with the periods $\{T_i\}_{i=1}^N$, amplitudes $\{A_i\}_{i=1}^N$, and phases $\{\phi_i\}_{i=1}^N$. This kind of model has many applications, and has been solved analytically under certain sets of assumptions (see e.g. Bretthorst, 1988). Similar models have been used in many different fields (e.g. Bretthorst, 2003; Umstätter et al., 2005; Brewer et al., 2007; Brewer and Stello, 2009). Note that this problem is also very closely related to the problem of detecting exoplanet signals in radial velocity data, which has attracted a lot of research attention in recent years (e.g. Brewer et al., 2007; Gregory, 2011; Hou et al., 2014; Feroz et al., 2011).

3.1 Simulated Dataset

To demonstrate the techniques, we simulated some data based on the assumption $N = 2$, i.e. the true signal was composed of two sinusoids. The simulated data is shown in Figure 2, and shows a large, low period oscillation with a smaller, much faster oscillation superimposed. The noise level is such that the fast oscillation is difficult to detect. The true values of the parameters were $N = 2$, $\mathbf{A} = \{1, 0.3\}$, $\mathbf{T} = \{30, 2\}$, and $\boldsymbol{\phi} = \{0, 1\}$. The signal was observed at $n = 1001$ points equally spaced between $t = 0$ and $t = 100$. The true form of the signal is:

$$y(t) = \sin\left(\frac{2\pi t}{30}\right) + 0.3\sin\left(\frac{2\pi t}{2} + 1\right) \tag{14}$$

and the noise standard deviation was $\sigma = 1$.

3.2 Prior Distributions

To infer N, the number of sinusoids, as well as all the periods, amplitudes, and phases, we need to define prior distributions. In the notation of Section 1, the parameters of the each object are

$$\mathbf{x}_i \quad = \quad \{A_i, T_i, \phi_i\}. \tag{15}$$

For the interim prior $p(\mathbf{x}_i|\boldsymbol{\alpha})$, we introduced a single hyperparameter μ, such that A_i given μ has an exponential prior with mean μ. For the periods, we assigned a log-uniform distribution for the periods between fixed boundaries, and a uniform distribution for the phases between 0 and 2π. In other words, our prior is only hierarchical for the amplitudes $\{A_i\}$. The prior for μ was a log-uniform prior, with probability density $\propto 1/\mu$, where $\mu \in [10^{-3}, 10^3]$.

The model for the shape of the (noise-free) signal is

$$y(t) \quad = \quad \sum_{i=1}^{N} A_i \sin\left(\frac{2\pi t}{T_i} + \phi_i\right) \tag{16}$$

where there are N sinusoids in the signal, the amplitudes are $\{A_i\}$, the periods are $\{T_i\}$, and the phases are $\{\phi_i\}$. The sampling distribution (probability distribution for the data given the parameters) was a normal (gaussian) distribution with mean zero and standard deviation σ, applied independently to each data point:

$$Y_i|N, \{A_i\}, \{T_i\}, \{\phi_i\}, \sigma \sim \mathcal{N}\left(\mu(t_i), \sigma^2\right). \tag{17}$$

This also depends on an additional noise standard deviation parameter σ which will also be inferred. We used a log-uniform prior for σ between 10^{-3} and 10^3.

3.3 Results

We ran DNS on the simulated data shown in Figure 2. The shape of the likelihood function with respect to the prior mass (the usual likelihood curve associated with Nested Sampling) is shown in Figure 3. Two phase transitions can be seen, where this curve is concave up. The first phase transition, at $\log(X) \approx -10$ nats, separates "noise-only" models (where the entire dataset is accounted for by the noise term) from "one-sinusoid" models where $N = 1$. At $\log(X) \approx -28$ nats, a second phase transition separates $N = 1$ models from $N = 2$ models. This phase transition causes a double-peaked feature in the posterior weights (lower panel of Figure 3).

If we were to do this analysis by trying to explore the posterior directly, rather than by using Nested Sampling, it would be difficult to jump between the $N = 1$ solution and the $N = 2$ solution, because the posterior distribution would be composed of a mixture of a small-volume but high-likelihood spike and a large-volume but (relatively) low likelihood slab. Note that this phase transition still exists if we condition on $N = 2$, so is not caused by the trans-dimensional model.

Although DNS explores a distribution other than the posterior (a mixture of constrained priors is used instead), the saved samples can be assigned importance weights so we can represent the posterior distribution. The marginal posterior distribution for N is shown in Figure 4. The most probable value is $N = 2$ which is also the true value, and there is a small probability for $N = 1$. The decreasing probabilities for $N > 2$ are due to the well-known natural "Occam's Razor" effect that can occur in Bayesian Inference (e.g. MacKay, 2003, Chapter 28).

The marginal likelihood estimate returned by DNS for this data was $\log(\mathcal{Z}) = -1410.8$, and the information, or Kullback-Leibler divergence from the prior to the posterior, was estimated to be 29.7 nats.

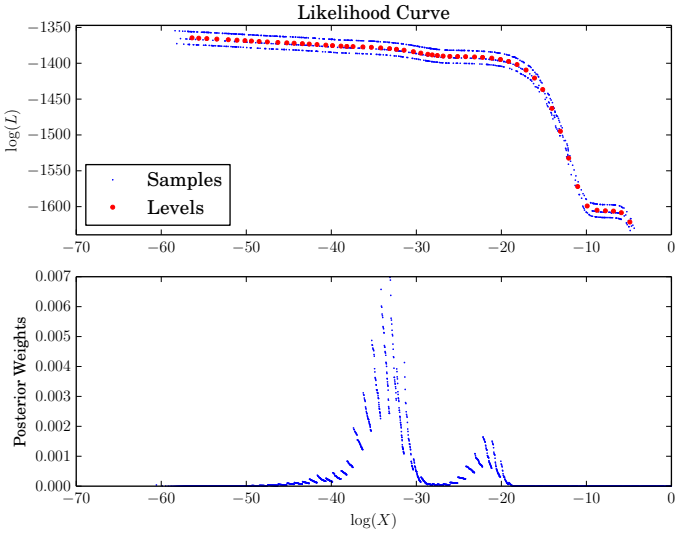

Figure 3: **Top panel:** The shape of the likelihood function with respect to prior mass. **Bottom panel:** The posterior weights of the saved samples. The existence of a phase transition causes this to display two separate peaks, which would be difficult to mix between if we simply attempted to sample the posterior distribution.

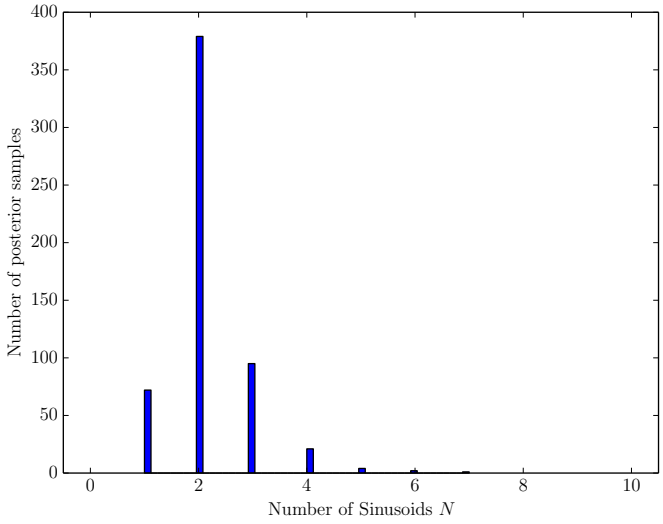

Figure 4: The inference for N, the number of sinusoids, based on the sinusoid data. The true value was $N = 2$, which is also the most probable value given the data. Of course, this is not always the case.

4 "Galaxy Field" Example

Source detection is an important problem in astrophysics. Given some data, usually one or more noisy, blurred, images of the sky, we would like to know how many objects N (such as stars or galaxies) are in the image, and their properties (such as flux, size, orientation). Various approaches exist, covering a wide spectrum between ad-hoc approaches and principled inference approaches (e.g. Irwin, 1985; Bertin and Arnouts, 1996; Dolphin, 2000; Hobson & McLachlan, 2003; Brewer et al., 2013). The more principled techniques tend to be much more computationally intensive, so should only be applied to interesting or especially challenging subsets of astronomical imaging. In fact, the approach used by Brewer et al. (2013) was essentially an early version of the methodology presented in this paper.

In this section we apply the DNS approach to a toy version of the problem of detecting and quantifying extended objects such as galaxies in a noisy image. We model each galaxy as a mixture of two concentric elliptical gaussian components, with the following eight parameters. Firstly, two parameters x_c and y_c describe the central position of the galaxy within the image. The flux f describes the total integral of the galaxy's intensity profile. The axis ratio q describes the ellipticity of the galaxy and θ its orientation angle with respect to horizontal. The radius of the bigger gaussian is a parameter w. Finally we include a "radius ratio" u describing the ratio of the radius of the smaller gaussian with respect to that of the bigger gaussian, and a parameter v describing the fraction of the total flux in the smaller gaussian (therefore the fraction of the light in the bigger gaussian is $1 - v$). In a coordinate system aligned with the major axis of the ellipse, the surface brightness profile of a "galaxy" is given by

$$\rho(x,y) = \frac{f(1-v)}{2\pi w^2}\exp\left[-\frac{1}{2w^2}\left(qx^2 + y^2/q\right)\right] \quad (18)$$

$$+\frac{fv}{2\pi(uw)^2}\exp\left[-\frac{1}{2(uw)^2}\left(qx^2 + y^2/q\right)\right]. \quad (19)$$

We generated a simulated 200×200 pixel noisy image (Figure 5) containing 47 "galaxies", in order to test the algorithm.

4.1 Priors

For the interim prior for the total fluxes of the galaxies, and for the widths of the galaxies, we used the Pareto distribution. The probability density function for a variable x given hyperparameters $x_{\min}$ and a is

$$p(x|x_{\min}, a) = \begin{cases} \frac{ax_{\min}^a}{x^{a+1}}, & x \geq x_{\min} \\ 0, & \text{otherwise.} \end{cases} \quad (20)$$

Since we are using this for both the fluxes and the widths of the galaxies, there will be four hyperparameters in total: two lower limits (one for the fluxes and one for the widths) and two slopes. The lower limits were assigned log-uniform priors between 10^{-3} and 10^3, and uniform priors between 0 and 1 were assigned to the *inverses* of the slopes. This latter choice was based on the fact that a Pareto distribution for x is an exponential distribution for $\ln(x)$, with scale length $1/a$.

The interim priors for the radius ratios u and the flux ratios w were both uniform distributions. Since $u \in [0,1]$ and $v \in [0,1]$, the upper limits of the uniform interim priors must be between 0 and 1. Therefore the prior for the upper limit was a uniform distribution between 0 and 1, and

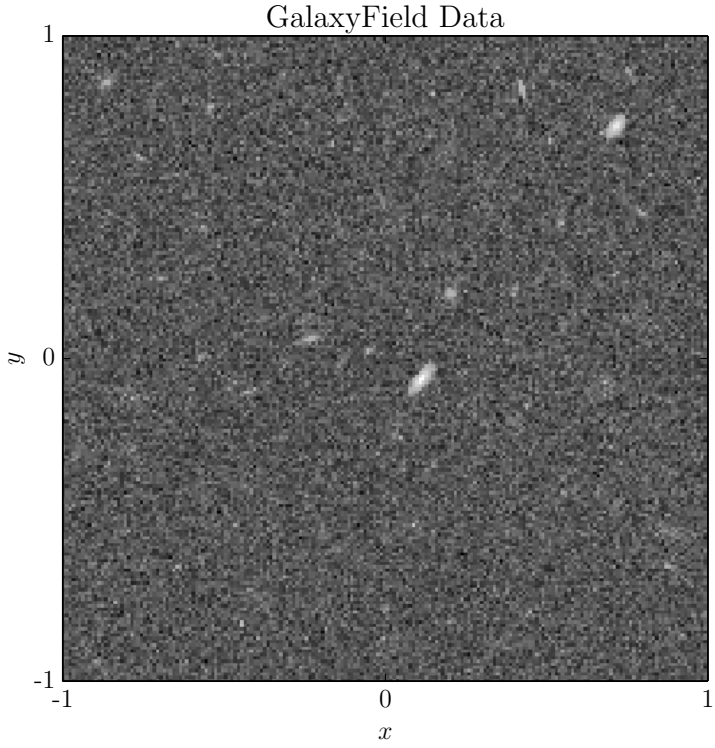

Figure 5: The simulated image for the GalaxyField example. The image is 200×200 pixels in size and contains $N = 47$ "galaxies".

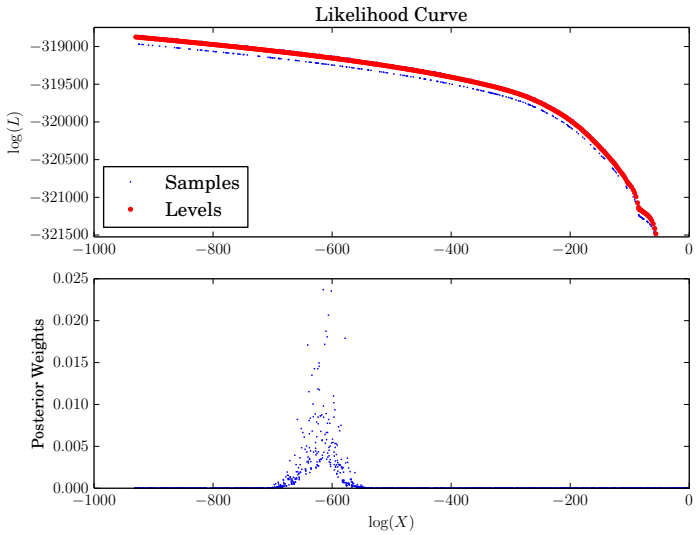

Figure 6: **Top panel:** The shape of the likelihood function with respect to prior mass for the GalaxyField data. Note the phase transition at $\log(X) \approx -100$ nats. This separates models with just the few brightest galaxies from models with many faint galaxies as well. While this doesn't affect the posterior distribution (unlike the sinewave example), it would affect calculation of the marginal likelihood if annealing were used. **Bottom panel:** The posterior weights of the saved samples.

the prior for the lower limit given the upper limit was a uniform distribution between 0 and the upper limit.

As in Section 3 we allowed the noise standard deviation σ to be a free parameter with a log-uniform prior between 10^{-3} and 10^{3}.

4.2 Results

The results from running DNS on the simulated galaxy data are shown in Figures 6 and 7. This problem, like the sinusoidal problem, exhibits a phase transition, which can be seen in Figure 6. The phase transition separates models which only fit the brightest galaxies from models that also fit the faint galaxies close to the noise level. This phase transition does not affect the posterior distribution, as the low-N solutions are eventually found to have very low importance. However it would cause difficulty for a thermal approach to calculating the marginal likelihood.

The marginal likelihood estimate returned by DNS for this data was $\log(\mathcal{Z}) = -319707.6$, and the information, or Kullback-Leibler divergence from the prior to the posterior, was estimated to be 550.2 nats.

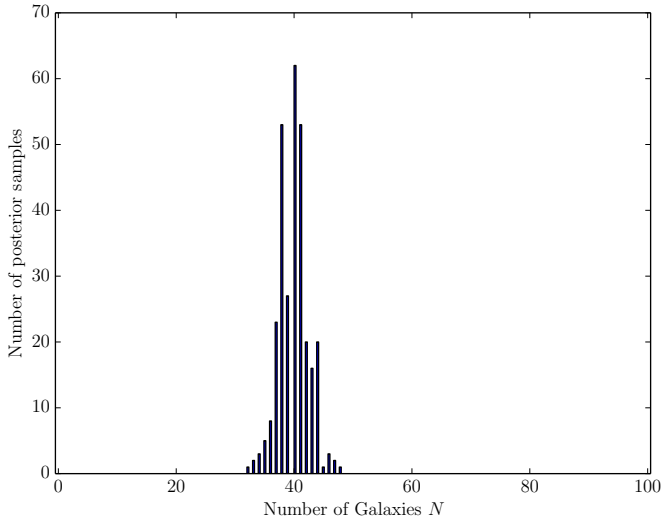

Figure 7: The inference for N, the number of galaxies, based on the GalaxyField data. The true value was $N = 47$.

5 Optimisation Techniques

In both of the examples discussed in this paper, the likelihood evaluation involved computing a mock noise-free signal from the current value of the model parameters. The mathematical form of the mock signal was a sum over all N components present. Computing the mock signal is often the most expensive step in the evaluation of the likelihood.

However, many of the proposal distributions used involve changing a subset of the parameters, while keeping others fixed. Considerable speedups can be achieved by, when appropriate, updating the mock signal to reflect the proposed change to the model, rather than computing the entire mock signal from scratch. For example, if the proposal is to add two new objects to the model, the mock signal can simply be updated by adding the effect of the two new objects to the model.

In general, it is possible to compute the number of components that have been affected by the proposal. If this is greater than or equal to N, we compute the mock signal from scratch, otherwise we subtract the effect of those components that have been removed and add in the effect of those that have been added. In the C++ implementation of this work, when a proposal takes place, the difference between the old and new model is cached and is available to the user, so that the likelihood can be updated rather than computed from scratch. Using this technique offers a speed advantage of a factor of ~ 2 for both of the example problems. The sinusoidal example could be solved in a few minutes on a desktop computer, whereas the galaxy example took much longer, about a day, to execute.

6 Conclusions

In this paper, we presented a general approach to trans-dimensional Bayesian inference problems. To compute the posterior distribution in these contexts, birth and death MCMC moves are

commonly used. However, the approach presented here replaces the posterior distribution with the alternative target distribution used in Diffusive Nested Sampling. This is done to facilitate mixing between multiple modes, along strong dependencies, and between different phases. Additionally, the marginal likelihood can be computed including the sum over the unknown number of components. The method was demonstrated on two illustrative examples inspired by astronomical data analysis, but should be applicable in other contexts such as mixture modelling. Software (in C++) implementing the general Metropolis proposals for these models, as well as the specific examples, is available at `https://github.com/eggplantbren/RJObject`.

Acknowledgements

This work is supported by a Marsden Fast-Start grant from the Royal Society of New Zealand. I would like to thank the following people for valuable conversations and inspiration: Anna Pancoast (UCSB), David Hogg (NYU), Daniel Foreman-Mackey (NYU), Courtney Donovan (Auckland), Tom Loredo (Cornell), Iain Murray (Edinburgh), John Skilling (MaxEnt Data Consultants), and Daniela Huppenkothen (Amsterdam, NYU).

References

Bertin, E., Arnouts, S. 1996. SExtractor: Software for source extraction. Astronomy and Astrophysics Supplement Series 117, 393-404.

Bretthorst, G. Larry. 1988. Bayesian Spectrum Analysis and Parameter Estimation. In Lecture Notes in Statistics, 48, Springer-Verlag, New York, New York.

Bretthorst, G. L. 2003. Frequency Estimation, Multiple Stationary Nonsinusoidal Resonances With Trend. Bayesian Inference and Maximum Entropy Methods in Science and Engineering 659, 3-22.

Brewer, B. J., Bedding, T. R., Kjeldsen, H., Stello, D. 2007. Bayesian Inference from Observations of Solar-like Oscillations. The Astrophysical Journal 654, 551-557.

Brewer, B. J., Stello, D. 2009. Gaussian process modelling of asteroseismic data. Monthly Notices of the Royal Astronomical Society 395, 2226-2233.

Brewer, B. J., Foreman-Mackey, D., Hogg, D. W. 2013. Probabilistic Catalogs for Crowded Stellar Fields. The Astronomical Journal 146, 7.

Brewer B. J., Pártay L. B., Csányi G., 2011, Statistics and Computing, 21, 4, 649-656. arXiv:0912.2380

Brewer, B. J., Lewis, G. F., Belokurov, V., Irwin, M. J., Bridges, T. J., Evans, N. W. 2011. Modelling of the complex CASSOWARY/SLUGS gravitational lenses. Monthly Notices of the Royal Astronomical Society 412, 2521-2529

Dolphin, A. E. 2000, PASP, 112, 1383

Feroz, F., Balan, S. T., Hobson, M. P. 2011. Detecting extrasolar planets from stellar radial velocities using Bayesian evidence. Monthly Notices of the Royal Astronomical Society 415, 3462-3472.

Feroz, F., Hobson, M. P. 2014. Bayesian analysis of radial velocity data of GJ667C with correlated noise: evidence for only two planets. Monthly Notices of the Royal Astronomical Society 437, 3540-3549.

Green, P. J., 1995, Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination, Biometrika 82 (4): 711732.

Gregory, P. C. 2011. Bayesian exoplanet tests of a new method for MCMC sampling in highly correlated model parameter spaces. Monthly Notices of the Royal Astronomical Society 410, 94-110.

Hansmann, Ulrich HE., 1997, "Parallel tempering algorithm for conformational studies of biological molecules.", Chemical Physics Letters 281, no. 1 (1997): 140-150.

Hobson, M. P., & McLachlan, C. 2003, MNRAS, 338, 765

Hou, F., Goodman, J., Hogg, D. W. 2014. The Probabilities of Orbital-Companion Models for Stellar Radial Velocity Data. ArXiv e-prints arXiv:1401.6128.

Irwin, M. J. 1985, MNRAS, 214, 575

Jasra, A., Holmes, C. C, and Stephens, D. A., Markov Chain Monte Carlo Methods and the Label Switching Problem in Bayesian Mixture Modeling, Statistical Science Vol. 20, No. 1, pp. 50-67

MacKay, David J. C., 2003, "Information theory, inference, and learning algorithms". Vol. 7. Cambridge: Cambridge university press, 2003.

Neal, R. M., 2001, Annealed importance sampling, Statistics and Computing, vol. 11, pp. 125-139.

Skilling J., 1998, Massive Inference and Maximum Entropy, in Maximum Entropy and Bayesian Methods, Kluwer Academic Publishers, Dordrecht/Boston/London p.14

Skilling, J., 2006, Nested Sampling for General Bayesian Computation, Bayesian Analysis 4, pp. 833-860.

Stephens, M., 2000, "Bayesian analysis of mixture models with an unknown number of components-an alternative to reversible jump methods.", Annals of Statistics (2000), 40-74.

Umstätter, R., Christensen, N., Hendry, M., Meyer, R., Simha, V., Veitch, J., Vigeland, S., Woan, G. 2005. Bayesian modeling of source confusion in LISA data. Physical Review D 72, 022001.