

Nested sampling with peers



Feishuang Wang

School of Computer Science
The University of Auckland

Supervisor: Dr Brendon James Brewer

A dissertation submitted in partial fulfillment of the requirements for the degree of Master of Professional Studies in Data Science OR Master of Data Science (delete one), The University of Auckland, 20XX.

Abstract

A nested sampling algorithm is a Bayesian approach to computing and comparing models and generating samples from posterior distributions. We introduce a general Monte Carlo method based on Nested Sampling, which name is Nested Sampling with peers, this method generates one particle above the threshold from the last iteration by querying the params from the server and updating it. We describe the new method over a test case and find that it has better accuracy than the original MCMC-based nested sampling with the same computational overhead. Put your abstract here. The abstract should contain a brief summary of the aim, methodologies, finding and conclusions of the dissertation. The abstract should normally be fewer than 350 words.

Contents

| | |
|--|-----------|
| Abstract | 1 |
| 1 Introduction | 5 |
| 1.1 Bayesian Analysis | 5 |
| 1.1.1 Marginal Likelihood | 6 |
| 1.2 Nested Sampling | 6 |
| 1.2.1 Model Selection with Nested Sampling | 6 |
| 1.2.2 Sorting | 7 |
| 1.2.3 A easy example | 8 |
| 1.2.4 Nested Sampling Algorithms | 8 |
| 1.3 Distributed System | 9 |
| 1.3.1 BONIC | 9 |
| 1.4 Spakslab Problem | 9 |
| 2 Implementation | 11 |
| 2.0.1 Itemized lists | 12 |
| 2.0.2 Inserting figures | 13 |
| 2.0.3 Tables | 13 |
| 2.0.4 How to Refer to Equations, Sections, etc | 14 |
| 3 Methodologies and analysis | 17 |
| 3.1 Methodologies | 17 |
| 3.2 Analysis | 17 |
| 4 Discussion | 19 |
| 4.1 Main results | 19 |
| 4.2 Discussion | 19 |
| 5 Conclusions | 21 |

| | |
|----------------------------|-----------|
| References | 23 |
| A Some extra things | 25 |

Chapter 1

Introduction

1.1 Bayesian Analysis

The Bayesian Analysis problem is in fact a parameter estimation problem, The term *parameter* means *unknown quantity*, however in real world, we don't have enough information to decide a quantity, so we need Bayesian Analysis to help us. The parameter you interested in are denoted by θ , to estimate them, first of all, you need to model a probability distribution on the *hypothesis space*, a *hypothesis space* is the collection of possibilities, this means that you are modeling the initial assumptions, and the probability distribution is called the *prior*, it is the distribution of θ . The data set is called D , after that we can use Bayes' Theorem to determine the *posterior distribution*:

$$p(\theta|D) = \frac{p(\theta)p(D|\theta)}{p(D)} \quad (1.1)$$

in the above equation,

1. The posterior distribution is $p(\theta|D)$, given the data set D with the conditional distribution of θ .
2. The prior distribution is $p(\theta)$.
3. The likelihood is $p(D|\theta)$, it means the conditional probability of observing the data.
4. The denominator $p(D)$ is called marginal likelihood or evidence, it doesn't depend on the parameter θ .

The posterior distribution is usually more constricted than the prior distribution, indicating that we have gained some insights from the data and that our level of uncertainty about the parameter values has been lowered. See the Figure 1.1 for how we often see the updating from the prior distribution to a posterior distribution.

1.1.1 Marginal Likelihood

The marginal likelihood shows how does the model be constructed. From probability theory we can learn that If you integrate over the posterior distribution, you must get 1 at the last. So we can write the marginal likelihood in the following form:

$$p(D|I) = \int p(\theta|I)p(D|\theta, I)d\theta \quad (1.2)$$

where I represents the underlying assumptions and information, and the integral is N-dimensional, go through all the continuous parameter space, the I is usually ignored and the marginal likelihood can be denoted as Z :

$$Z = \int p(\theta)p(D|\theta)d\theta \quad (1.3)$$

when sum the discrete parameter θ , it also has the following form:

$$Z = \sum p(\theta)p(D|\theta) \quad (1.4)$$

The ratio of evidence values is always called Bayes factors, calculating the value of Z is important for comparing different model assumptions, so the Z-value allows the results to be predicted into the future because the predicted model can be compared with the current model without the need for recalculation, in short, this parameter measures how well a model fits the data. However, the standard algorithm such as Markov chain Monte Carlo (MCMC) are considered for posterior distribution, it cannot provide the evidence like Bayesian Inference, for it only gives a collection of normalised posterior, it will be a challenge to get the evidence in Bayesian analysis.

1.2 Nested Sampling

1.2.1 Model Selection with Nested Sampling

Consider two mutually exclusive model M_1 and M_2 , M_1 has the parameter θ_1 and M_2 has the parameter θ_2 , from previous method we discussed, we can calculate the posterior distribution of θ_1 :

$$p(\theta_1|D, M_1) = \frac{p(\theta_1|M_1)p(D|\theta_1, M_1)}{p(D|M_1)} \quad (1.5)$$

Same, calculate the posterior distribution of θ_2

$$p(\theta_2|D, M_2) = \frac{p(\theta_2|M_2)p(D|\theta_2, M_2)}{p(D|M_2)} \quad (1.6)$$

For simplicity, we often calculate the ratio of posterior distributions to compare models:

$$\frac{p(M_1|D)}{p(M_2|D)} = \frac{p(M_1)}{p(M_2)} \times \frac{p(D|M_1)}{p(D|M_2)} \quad (1.7)$$

The result of this formula is called *posterior odds*, the posterior probability of M2 over M1 depends on the prior probability: is M2 more credible than M1 before considering the data? Another ratio is the ratio of likelihoods, sometimes called the Bayesian coefficient; how likely is the data for hypothetical M2 versus hypothetical M1? These possibilities are not the likelihood of a particular value of the parameter, but the likelihood of the entire model. We have to use the marginal likelihood to make a better progress.

1.2.2 Sorting

In multiple dimensions, the direct integration calculation for the edge likelihood becomes unrealistic, we define the X as the model's prior mass, it can be integrated in this form:

$$X(\lambda) = \int_{L(\theta) > \lambda} \pi(\theta) d\theta \quad (1.8)$$

we can write the inverse function $L(X)$, the marginal likelihood *i.e.* evidence will be in this form:

$$Z = \int_0^1 L(X) dX \quad (1.9)$$

To get the inverse function, we can divide the prior into little particles and then sort all the particles by likelihood. In geometrical terms, the function $L(X)$ is monotonically diminishing, and the likelihood decreases as the value of X increases from 0 to 1, as shown in Figure 1.1. From the figure 2.1 we can find that when approaching the maximum likelihood, X will keep decreasing.

The explanation of 1.5 is that the $L(X)$ is the probability corresponding to λ that makes the probability, $L(\theta)$, greater than λ for X . $L(0.4) = 0.01$ may be interpreted as a 40% probability that $\pi(\theta)$ will be greater than 0.01 for θ chosen from the prior. To evaluate the mapped integral in 1.5, Skilling proposed a simple trapezoidal rule by calculating weighted sums:

$$Z \approx \sum_{i=1}^j w_i L_i (w_i = \frac{1}{2}(X_{i-1} - X_{i+1})) \quad (1.10)$$

In every iteration i , the likelihood is $L(X_i)$, and j is total number of iterations.

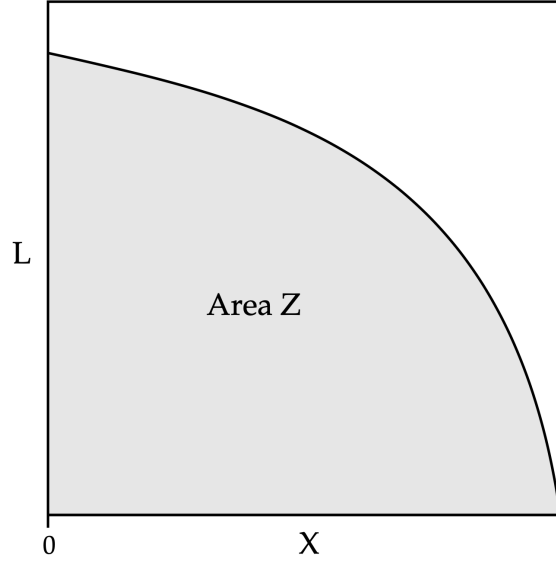


Figure 1.1: Likelihood function with area Z.

1.2.3 A easy example

Imagine we have a model whose parameter is a single X , and if the parameter X follows the uniform prior distribution $\pi(\theta)$ from 0 to 1, it also has a likelihood function $L(X)$ which is a decreasing function of X , the Figure 1.2. illustrate the graph of the example, It's easy to find that in this situation, the evidence will be $Z = \int_0^1 L(X)DX$

1.2.4 Nested Sampling Algorithms

This section will discuss the main idea of Nested Sampling, it is a Monte Carlo algorithm and not a typical MCMC algorithm in technical. John Skilling proposed the algorithm specifically for approximating these marginal integrals, *i.e.*, evidence, and it also generates samples from the posterior distribution. In the prior distribution $\pi(\theta)$, it needs to generate N *particles or live points* θ , just like an vector to initialize the starting state. To approximating the evidence, there's some important steps, if we find the points with the worst likelihood, always denoted by L_{worst} and this is often associated with the highest X value, If different particle N are taken, the highest X value will always be close to $X = 1$ and not close to $X = 0$. Skilling (2006) gives a simple treatment by estimating $X_{worst} = e^{(\frac{-1}{N})}$, and we also know the probability of this point.

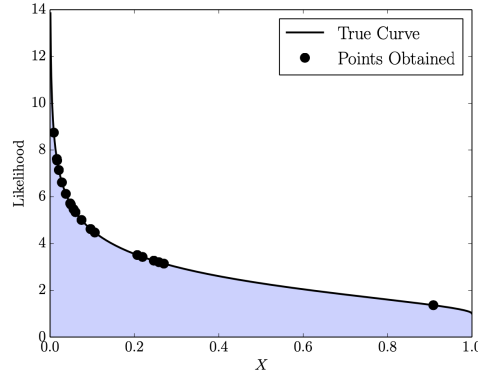


Figure 1.2: A simple model estimate problem with only one parameter

By generating a new point X_1 from the prior distribution $\pi(\theta)$ to replace the worst X_{worst} , more points can be obtained,

and the restriction is that the likelihood of the point must be higher than L_{worst} . Also, the prior distribution $\pi(\theta)$ is sampled from the conditional prior, $\pi(\theta|L(\theta > \lambda_{min,1}))$, we can denote the new set of particles as $\omega_1 = \{\theta|L(\theta > \lambda_{min,1})\}$, it's a subset of the whole parameter space ω . And in the second iteration, the estimation of the X value who has a worst particle is $e^{(-\frac{1}{N})} \times e^{(-\frac{1}{N})} = e^{(-\frac{2}{N})}$, because we already know the likelihood, so we can continue the iteration. So the second step to generate the new point X_2 and the X_{worst} , X will be repeated, we use ω_2 to denote them, and we should know that $\omega_1 \subset \omega_2$, so the X_{worst} will be replaced, too, from the prior distribution $\pi(\theta)$ from the conditional prior, $\pi(\theta|L(\theta > \lambda_{min,2}))$, so we can write when the iteration run i times and it will generate a sequence of $\lambda_{min,i}$ will be like following:

$$\lambda_{min,1} < \lambda_{min,2} < \dots < \lambda_{min,i-1} < \lambda_{min,i}$$

because every ω_i is subset of the whole space ω , so it will decrease continuously, so it can be viewed as a reduction or compressing of the parameter space ω and the subspaces are nested within each other, so we have the term *nested sampling*.

1.3 Distributed System

1.3.1 BONIC

1.4 Spakslab Problem

Chapter 2

Implementation

We first show some simple examples of mathematical formulae using latex typesetting.

1. The basic functions: $\cos(x), \sin(x), \ln(x)$, (`\cos (x) , \sin (x) , \ln (x)`).
2. Greek letters: $\alpha\beta\gamma\delta\epsilon\dots$ (`\alpha\beta\gamma\delta\epsilon\dots`).
3. Mathematical symbols: $\int \oint \sum \lim \bigcup \bigcap$ (`\int \oint \sum \lim \bigcup \bigcap`).
4. Fractions: $\frac{1}{2}, \frac{1}{2-x}$ (`\frac{1}{2} , \frac{1}{2-x}`).

The following matrix

$$\begin{bmatrix} U_r & r & W_r \\ 0 & 1 & V_x \\ 0 & 0 & W_x \end{bmatrix}, \quad (2.1)$$

is generated using the `equarray` environment:

```
\begin{equarray}\label{eqn:matrix}
\left[
\begin{array}{ccc}
U_{r}& r & W_{r}\\
0 & 1 & V_{x}\\
0 & 0 & W_{x}
\end{array}
\right],
\end{equarray}
```

The `\label{eqn:matrix}` command labels the equation with `{eqn:matrix}` which can be referred to somewhere else in the text by using `\ref{eqn:matrix}` or `\eqref{eqn:matrix}`.

The command `\notag` eliminates the numbering of the first equation,

$$\begin{aligned}\lambda^{(1)} &= \text{tr}[T^{(1)}P], \\ \lambda^{(2)} &= \text{tr}[T^{(2)}P - T^{(1)}ST^{(1)}P].\end{aligned}\tag{2.2}$$

```
\begin{eqnarray} \label{eqn:lambda_trace} \\ \lambda^{(1)}&=&\text{tr}[T^{(1)}P], \notag \\ \lambda^{(2)}&=&\text{tr}[T^{(2)}P - T^{(1)}ST^{(1)}P]. \\ \end{eqnarray}
```

2.0.1 Itemized lists

Example of an itemized list:

- muscle and fat cells remove glucose from the blood,
- cells use glucose for protein synthesis.

```
\begin{itemize} \\ \item muscle and fat cells remove glucose from the blood, \\ \item cells use glucose for protein synthesis. \\ \end{itemize}
```

This can be done by an enumerated list:

1. muscle and fat cells remove glucose from the blood,
2. cells use glucose for protein synthesis.

```
\begin{enumerate} \\ \item muscle and fat cells remove glucose from the blood, \\ \item cells use glucose for protein synthesis. \\ \end{enumerate}
```

2.0.2 Inserting figures

You may save your Matlab figures as jpg files. Figures should be stored in the same folder as the latex files. For the graphicx package to work you usually need to ask latex to create a pdf file (e.g., command `pdflatex` or `latexpdf`).

An example of an inserted image is given in Figure 2.1.

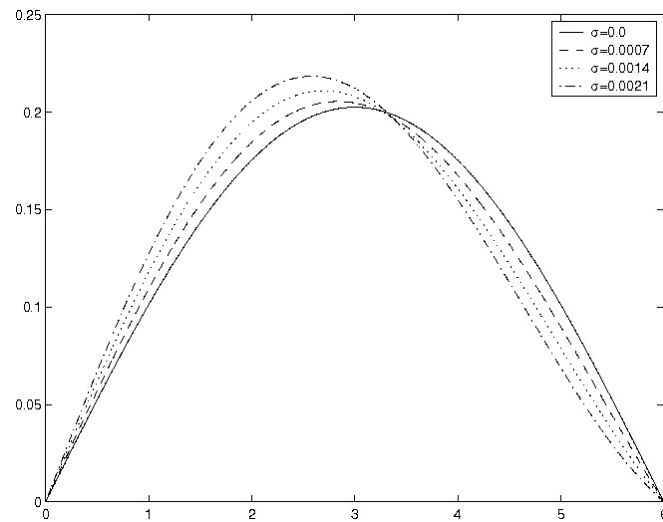


Figure 2.1: Mode shapes

2.0.3 Tables

Example of a table,

| Gene | GeneID | Length |
|---------------|--------|-----------|
| human latexin | 1234 | 14.9 kbps |
| mouse latexin | 2345 | 10.1 kbps |
| rat latexin | 3456 | 9.6 kbps |

Table 2.1: **title of table** - Overview of latexin genes.

```
\begin{table}[htp]
\centering
\begin{tabular}{ccc}
```

```
% ccc means 3 columns, all centered; alternatives are l, r
{\bf Gene} & {\bf GeneID} & {\bf Length} \\
\hline % draws a line under the column headers
human latexin & 1234 & 14.9 kbps \\
mouse latexin & 2345 & 10.1 kbps \\
rat latexin   & 3456 & 9.6 kbps \\
\end{tabular}
\caption[title of table]{\textbf{title of table} - Overview of latexin genes.}
\label{latexin_genes} % label for cross-links with \ref{latexin_genes}
\end{table}
```

See how to add two vertical lines in the table (Simply change {ccc} to {c|c|c})

| Gene | GeneID | Length |
|---------------|--------|-----------|
| human latexin | 1234 | 14.9 kbps |
| mouse latexin | 2345 | 10.1 kbps |
| rat latexin | 3456 | 9.6 kbps |

Table 2.2: **title of table** - Overview of latexin genes.

2.0.4 How to Refer to Equations, Sections, etc

- References can be linked to equations, figures, tables or sections using the command `\ref`: Equation (2.2), Figure 2.1, Table 2.2 and Section 2.0.3.
`Equation~(\ref{eqn:lambda_trace}), Figure~\ref{modes},`
`Table~\ref{latexin_genes2} and Section~\ref{table}.`
- Equations can be conveniently referred to using `\eqref`. See, for example, Equation (2.2).
`Equation \eqref{eqn:lambda_trace}`
 Note that `\eqref` includes the round brackets by itself.
- Citations are in a similar way but using the command `\cite`:
`[2], [3], and [4], or [2, 3, 4] .`

```
\cite{Salmond}, \cite{Stull}, and \cite{TandC},
or \cite{Salmond,Stull,TandC} .
```

There are many different styles for writing citations – you should follow the norms for your subject area.

A more advanced way to do citations is to use `bibtex`. This is a powerful tool and we encourage you to try it. There is plenty of information about it on the web.

Chapter 3

Methodologies and analysis

3.1 Methodologies

3.2 Analysis

Chapter 4

Discussion

4.1 Main results

4.2 Discussion

Chapter 5

Conclusions

You may add more chapters as needed in the file.

References

- [1] Farge Marie, *Wavelet Transforms and Their Applications to Turbulence*, Ann. Rev. Fluid Mech. volume 24, pages 395-457, 1992.
- [2] Salmond Jennifer, *Vertical Mixing of Ozone in the Very Stable Nocturnal Boundary Layer*, PhD Thesis, University of British Columbia, 2001.
- [3] Stull B. Ronald, *Introduction to Boundary Layer Meteorology*, Dordrecht; Boston: Kluwer Academic Publishers, 1988.
- [4] Torrence Christopher, Compo Gilbert P., *A Practical Guide to Wavelet Analysis*, Bulletin of the American Meteorological Society volume 79, pages 61-78, 1998.

Appendix A

Some extra things

This is an optional chapter for any additional material that does not fit conveniently into the body of the text (e.g., data, copies of computer programmes). Note that appendices won't necessarily be marked.