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