Nested Sampling and the Evaluation of the 'Evidence' for Bayesian Model Selection

Paul D. Baines, Nicholas Ulle
University of California, Davis

March 23, 2014

1 Introduction

A popular Bayesian model selection strategy is comparison of the "evidence" for each model. That is, the average likelihood over the prior probability space. This is quite intuitive—models with greater evidence are a better fit for the data. Here we will focus solely on methods for evaluating the evidence.

Formally, for likelihood L and prior π, the evidence is defined as

$$Z(y) = \int L(y;\theta)\pi(\theta)\,\mathrm{d}\theta\,.$$

This expression is deceptively simple-looking; in practice, direct evaluation is often impractical or impossible, because the prior parameter θ is multi-dimensional.

Nested sampling overcomes this difficulty by recasting the problem, to arrive at a one-dimensional integral which can the be evaluated using standard methods from numerical analysis. Consider the random variable $L(y;\theta) \geq 0$, which has survival function

$$S(\lambda) = \Pr_{\pi}\big[L(y;\theta) > \lambda\big].$$

A well-known result is that the integral of the survival function of a non-negative random variable is its expectation. In this case,

$$Z(y) = \mathbb{E}_{\pi}\big[L(y;\theta)\big] = \int_0^\infty S(\lambda)\,\mathrm{d}\lambda\,.$$

Since S is monotonic, this area is the same as

$$Z(y) = \int_0^1 S^{-1}(x)\,\mathrm{d}x,$$

where S^{-1} is the upper quantile function of $L(y;\theta)$. Explicit evaluation of S^{-1} may be quite challenging, but we can avoid this entirely. Instead, we sample n values $\theta_1,\ldots,\theta_n$ from the prior π, and define $L_i = L(y;\theta_i)$, for $i = 1,\ldots,n$. The smallest among these, $L_{(1)}$, is an estimate of the $(1-1/n)$ upper quantile. If we remove $L_{(1)}$ from the sample and replace it with a new point drawn in the same way, but constrained to be strictly greater, then the new $L_{(1)}$ is an estimate of the $(1-1/n)^2$ upper quantile. By proceeding in this fashion, we can estimate S^{-1} over its entire domain. It's then straightforward to approximate the integral numerically.

Here we have implemented the nested sampler and applied it to a toy example where the evidence can also be computed analytically. We also propose a mixture model, where the evidence cannot be computed analytically, for further study, including the use of alternative estimators of the evidence.

2 Toy Example

Let:

$$Y_i \sim N(\mu, \sigma^2), \qquad i = 1, \ldots, n,$$

with prior $p(\mu) \propto N(\mu_0, \tau_0^2)$ and σ^2 known.

Letting $C = (2\pi)^{-(n+1)/2}(\tau_0^2)^{-1/2}(\sigma^2)^{-n/2}$, the evidence, or marginal likelihood, is:

$$\begin{aligned}
p(y) = &\int p(y_1, \ldots, y_n|\mu)p(\mu)d\mu = \int p(\mu)\prod_{i=1}^{n} p(y_i|\mu)d\mu \\
&= \int C \times \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2 - \frac{1}{2\tau_0^2}(\mu-\mu_0)^2\right\}d\mu \\
&= C \times \int \exp\left\{-\frac{1}{2}\left(\frac{n}{\sigma^2}+\frac{1}{\tau_0^2}\right)\mu^2 + \frac{1}{2}\mu\left(\frac{\mu_0}{\tau_0^2}+\frac{\sum_{i=1}^{n} y_i}{\sigma^2}\right) - \frac{1}{2}\left(\frac{\mu_0^2}{\tau_0^2}+\frac{\sum_{i=1}^{n} y_i^2}{\sigma^2}\right)\right\}d\mu \\
&= C \times \exp\left\{-\frac{1}{2}\left(\frac{\mu_0^2}{\tau_0^2}+\frac{\sum_{i=1}^{n} y_i^2}{\sigma^2}\right) + \frac{1}{2}\left(\frac{n}{\sigma^2}+\frac{1}{\tau_0^2}\right)^{-1}\left[\frac{\mu_0}{\tau_0^2}+\frac{\sum_{i=1}^{n} y_i}{\sigma^2}\right]^2\right\} \\
&\times \int \exp\left\{-\frac{1}{2}\left(\frac{n}{\sigma^2}+\frac{1}{\tau_0^2}\right)\left(\mu - \left(\frac{n}{\sigma^2}+\frac{1}{\tau_0^2}\right)^{-1}\left[\frac{\mu_0}{\tau_0^2}+\frac{\sum_{i=1}^{n} y_i}{\sigma^2}\right]\right)^2\right\}d\mu \\
&= C \times \exp\left\{-\frac{1}{2}\left(\frac{\mu_0^2}{\tau_0^2}+\frac{\sum_{i=1}^{n} y_i^2}{\sigma^2}\right) + \frac{1}{2}\left(\frac{n}{\sigma^2}+\frac{1}{\tau_0^2}\right)^{-1}\left[\frac{\mu_0}{\tau_0^2}+\frac{\sum_{i=1}^{n} y_i}{\sigma^2}\right]^2\right\} \times (2\pi)^{1/2}\left(\frac{n}{\sigma^2}+\frac{1}{\tau_0^2}\right)^{-1/2} \\
&= C \times (2\pi)^{1/2}\left(\frac{n}{\sigma^2}+\frac{1}{\tau_0^2}\right)^{-1/2} \exp\left\{-\frac{1}{2}\left(\frac{\mu_0^2}{\tau_0^2}+\frac{\sum_{i=1}^{n} y_i^2}{\sigma^2}\right) + \frac{1}{2}\left(\frac{n}{\sigma^2}+\frac{1}{\tau_0^2}\right)^{-1}\left[\frac{\mu_0}{\tau_0^2}+\frac{\sum_{i=1}^{n} y_i}{\sigma^2}\right]^2\right\}.
\end{aligned}$$

This will allow us to verify the results obtained using nested sampling.

2.1 Evaluating the Evidence: Nested Sampling

We consider the simple case where $\mu_0 = 0$, $\tau_0^2 = 1$, $n = 1$, and $\sigma^2 = 1$. In this case, the analytic solution for the evidence yields

$$Z = (2\pi)^{-1/2}(2)^{-1/2}\exp\left\{-\frac{1}{2}y^2 + \frac{1}{2}(2)^{-1}y^2\right\} = \frac{1}{2\sqrt{\pi}}\exp\left\{-\frac{y^2}{4}\right\},$$

or equivalently, $\log Z \approx -7.5155$. Direct application of Gaussian quadrature to the integral confirms that this value is correct.

The nested sampler was applied three times, using 1500 iterations and a sample size of $n = 200$. The resulting log-estimates are shown in Table 1. All three fall reasonably close to the true value, particularly the third.

Trial	Estimate
1	-7.2957
2	-7.9335
3	-7.5609

Table 1: Log-estimates from the nested sampler.

3 Mixture Example

Here we take a look at the classic mixture of normals:

$$Y_i = \sum_{j=1}^{K} I_{ij} Z_{ij}, \qquad i = 1, \ldots, n,$$

where:

$$I_i = (I_{i1}, \ldots, I_{iK}) \sim \text{Multinomial}\,(1, p)\,,$$
$$Z_{ij} \overset{iid}{\sim} N\,(\mu_j, 1)\,.$$

The parameters in the model are the mixture proportions $p = (p_1, \ldots, p_K)$ (with $\sum_j p_j = 1$) and the mixture locations $\mu = (\mu_1, \ldots, \mu_k)$. The number of mixture components K will be fixed for a given model, and we will use the evidence to motivate a model selection procedure to select the appropriate K. For convenience we choose conditionally conjugate priors for μ and p:

$$\mu \sim N(\mu_0, \tau_0^2), \qquad p \sim \text{Dirichlet}(\alpha),$$

where μ_0, τ_0^2 and α are fixed hyperparameters chosen by the analyst.

3.1 Posterior Distributions

In a slight abuse of notation, let $\{I_i = j\}$ be the event that I_i has a one in the j^{th} position. The random variables $(Y_i|\mu, p)$ are independent, and each has density

$$\begin{aligned} f_{Y_i}(y_i|\mu, p) &= \sum_{j=1}^{K} f(y_i|\mu, p, I_i = j)\,\Pr(I_i = j|p) \\ &= (2\pi)^{-1/2} \sum_{j=1}^{K} \exp\left[-\frac{1}{2}(y_i - \mu_j)^2\right] p_j. \end{aligned}$$

Consequently, the posterior density is

$$\begin{aligned} f_{\mu,p}(\mu, p|y) &\propto \left\{ \prod_{i=1}^{n} f_{Y_i}(y|\mu, p) \right\} \pi_\mu(\mu)\pi_p(p) \\ &\propto \left\{ \prod_{i=1}^{n} \sum_{j=1}^{K} \exp\left[-\frac{1}{2}(y_i - \mu_j)^2\right] p_j \right\} \prod_{j=1}^{K} \exp\left[-\frac{1}{2\tau_0^2}(\mu_j - \mu_0)^2\right] p_j^{\alpha_j - 1}. \end{aligned}$$

Computing the evidence for this distribution directly is intractable.

Since we may want to sample from the posterior distribution, the conditional posteriors are also of interest. The probability mass of $I_i|\mu, p, Y$ is

$$\begin{aligned}\Pr(I_i = j|\mu, p, Y) &\propto f_{Y_i}(y_i|\mu, p, I_i = j)\Pr(I_i = j|p)\\ &\propto \exp\left[-\frac{1}{2}(y_i - \mu_j)^2\right]p_j.\end{aligned}$$

These probabilities can be normalized easily upon computation. Next, define

$$n_j = \sum_{i\,:\,I_i=j} 1 \qquad \text{and} \qquad \bar{y}_j = \frac{1}{n_j}\sum_{i\,:\,I_i=j} y_i.$$

Then the density of $(\mu_j|I, Y)$ is

$$\begin{aligned}f_{\mu_j}(\mu_j|I, Y) &\propto \left\{\prod_{i\,:\,I_i=j} f_{Y_i}(y_i|\mu_j, I_i)\right\}\pi_{\mu_j}(\mu_j)\\ &\propto \exp\left[-\frac{1}{2}\sum_{i\,:\,I_i=j}(y_i - \mu_j)^2\right]\exp\left[-\frac{1}{2\tau_0^2}(\mu_j - \mu_0)^2\right]\\ &\propto \exp\left[-\frac{1}{2}\sum_{i\,:\,I_i=j} y_i^2 - 2y_i\mu_j + \mu_j^2\right]\exp\left[-\frac{1}{2\tau_0^2}(\mu_j - \mu_0)^2\right]\\ &\propto \exp\left[-\frac{1}{2}(-2n_j\bar{y}_j\mu_j + n_j\mu_j^2) - \frac{1}{2\tau_0^2}(\mu_j^2 - 2\mu_0\mu_j)\right]\\ &\propto \exp\left[-\frac{1}{2}\Big\{(n_j + 1/\tau_0^2)\mu_j^2 - 2(n_j\bar{y}_j + \mu_0/\tau_0^2)\mu_j\Big\}\right]\\ &\propto \exp\left[-\frac{1}{2}(n_j + 1/\tau_0^2)\Big\{\mu_j^2 - 2\,\frac{n_j\bar{y}_j + \mu_0/\tau_0^2}{n_j + 1/\tau_0^2}\,\mu_j\Big\}\right],\end{aligned}$$

from which we can infer that

$$(\mu_j|I, Y) \sim \mathrm{N}\left(\frac{n_j\bar{y}_j + \mu_0/\tau_0^2}{n_j + 1/\tau_0^2},\ \frac{1}{n_j + 1/\tau_0^2}\right), \qquad j = 1, \ldots, K.$$

Finally, the probability density of $(p|I, Y)$ is

$$\begin{aligned}f_p(p|I, Y) &\propto f_I(I|p)\pi_p(p)\\ &\propto \left\{\prod_{j=1}^{K} p_j^{n_j}\right\}\prod_{j=1}^{K} p_j^{\alpha_j-1}\\ &\propto \prod_{j=1}^{K} p_j^{\alpha_j+n_j-1}.\end{aligned}$$

Thus $(p|I, Y) \sim \text{Dirichlet}(\alpha + \vec{n})$, for $\vec{n} = (n_1, \ldots, n_K)$.

3.2 Evaluating the Evidence: Nested Sampling

A sample of $n = 1000$ observations was generated from the model, with parameters

$$\begin{aligned}K &= 3,\\ p &= (0.3439, 0.0537, 0.6024),\\ \mu &= (-1.5, 0, 1.5).\end{aligned}$$

This will be the subject of further study.

3.3 Evaluating the Evidence: Other Methods

Several alternatives are available for evaluating the evidence. One of these is the harmonic mean estimator

$$\hat{Z}_1 = \left[\frac{1}{m} \sum_{i=1}^{m} f_Y \left(y | \mu^{(i)}, p^{(i)} \right)^{-1} \right]^{-1}$$

first proposed by Newton and Raftery. Another is

$$\hat{Z}_2 = \frac{\delta m + (1-\delta) \sum_{i=1}^{m} \frac{f_Y(y|\mu^{(i)}, p^{(i)})}{\delta \hat{Z}_2 + (1-\delta) f_Y(y|\mu^{(i)}, p^{(i)})}}{\delta m \hat{Z}_2 + (1-\delta) \sum_{i=1}^{m} \{\delta \hat{Z}_2 + (1-\delta) f_Y(y|\mu^{(i)}, p^{(i)})\}^{-1}},$$

which must be evaluated using an iterative method.

These estimators require sampling from the posterior distribution $(\mu, p|Y)$, which we implement as a 2-stage Gibbs sampler:

1. Sample from $(I_i|\mu, p, Y)$, for $i = 1, \ldots, n$;
2. Sample from $(\mu|I, Y)$ and $(p|I, Y)$.

The necessary distributions were derived in Section 3.1. The implementation was used to draw a sample of 15000 observations. Trace plots are shown in Figure 1. They indicate that the chain managed to find all three modes of the posterior likelihood.

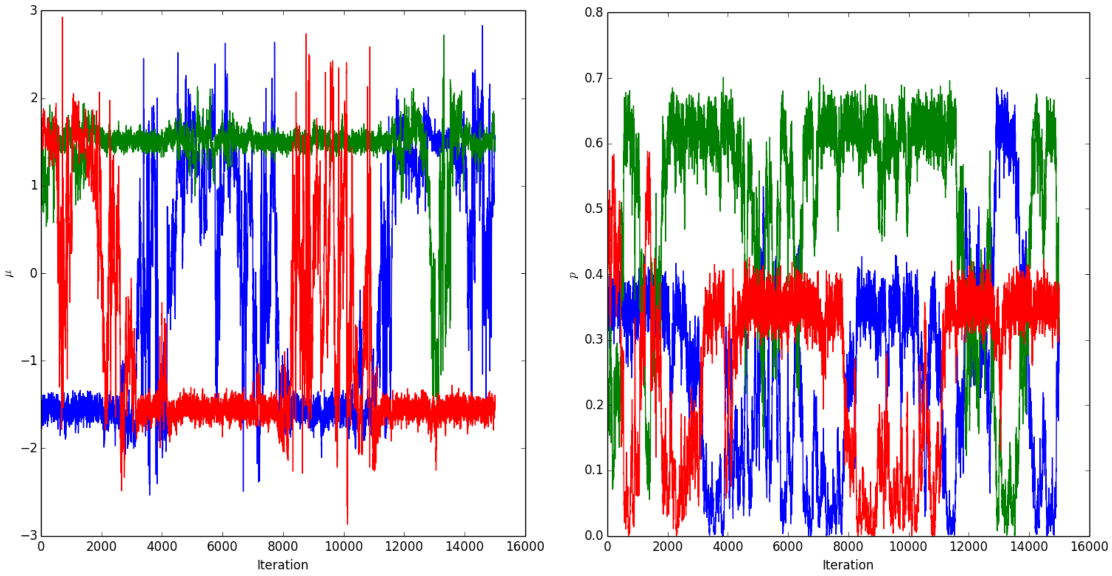

Figure 1: Trace plots for the sampled values of μ and p. Each component is shown in a different color.