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

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

March 19, 2014

1 Introduction

- Explanation of Nested Sampling
- Intuitive explanation of the algorithm
- Basics of computing the evidence and Bayesian model selection

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. In the simple case where $\mu_0 = 0$, $\tau_0^2 = 1$, $n = 1$, $\sigma^2 = 1$ we obtain:

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

3 Mixture Example

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

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

where:

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

The parameters in the model are the mixture proportions $p = (p_1, \dots, p_K)$ (with $\sum_j p_j = 1$) and the mixture locations $\mu = (\mu_1, \dots, \mu_k)$. 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

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

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

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

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.