Advanced Simulation - Lecture 11

Patrick Rebeschini

February 19th, 2018

Outline

- Often we have various possible models for the same dataset.
- Reversible jump enables joint parameter and model estimation, in one run.
- How to choose between models without resorting to reversible jump?
- Various Monte Carlo ways to estimate the evidence associated to each model.

Bayesian model choice

- Assume we have a collection of models $\mathcal{M}_k$ for $k \in \mathcal{K}$.
- With data we can learn parameters given each model $\mathcal{M}_k$, but we can also learn about the models.
- Put a prior on models $\mathcal{M}_k$. Within each model, prior $p(\theta_k \mid \mathcal{M}_k)$ on the parameters.
- Joint posterior distribution of interest:

$$\pi(\mathcal{M}_k, \theta_k \mid y) = \pi(\mathcal{M}_k \mid y)\pi(\theta_k \mid y, \mathcal{M}_k)$$

which is defined on

$$\cup_{k\in\mathcal{K}}\{\mathcal{M}_k\} \times \Theta_k.$$

Bayesian polynomial regression

- We select $k \in \{0, ..., M_{\max}\}$ and

$$\mathbb{P}\left(\mathcal{M}_k\right) = p_k = \frac{1}{M_{\max} + 1}$$

with $\Theta_k = \mathbb{R}^{k+1} \times \mathbb{R}^+$

$$p_k\left(\beta, \sigma^2\right) = \mathcal{N}\left(\beta; 0, \sigma^2 I_{k+1}\right) \mathcal{IG}\left(\sigma^2; 1, 1\right).$$

- In this case, we have analytic expression for

$$p_k\left(y_{1:n}\right) = \int_{\Theta_k} p_k\left(\beta, \sigma^2\right) \prod_{i=1}^{n} \mathcal{N}\left(y_i; f\left(x_i; \beta\right), \sigma^2\right) d\beta d\sigma^2.$$

- Bayesian model selection automatically prevents overfitting.

Bayesian polynomial regression

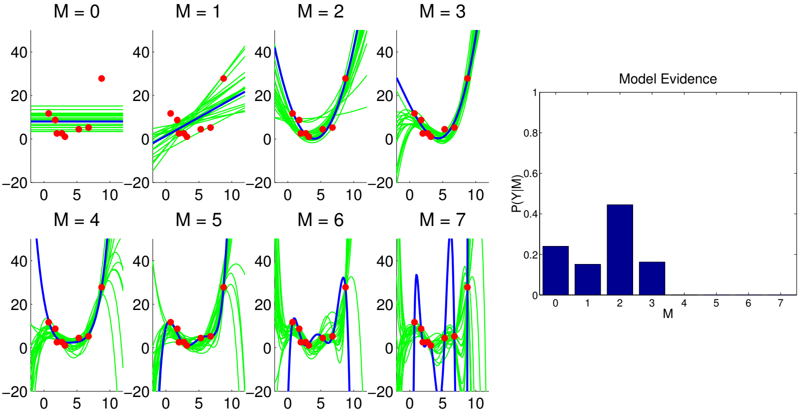

Figure: $f(x;\beta)$ for random draws from $p_M(\beta|\, y_{1:n})$ and evidence $p_M(y_{1:n})$.

Transdimensional samplers

- Reversible Jump aims at parameter estimation and model choice in one run.

- In general, hard to design auxiliary variables for dimension matching and deterministic mappings.

- Transdimensional samplers constitute an on-going research area.

Estimation of the evidence

- The model evidence, or normalizing constant, is $\pi(y \mid \mathcal{M}_k)$:

$$\pi(\theta_k \mid y, \mathcal{M}_k) = \frac{\pi(\theta_k \mid \mathcal{M}_k)\pi(y \mid \theta_k, \mathcal{M}_k)}{\pi(y \mid \mathcal{M}_k)}.$$

- Using some integral representation, for instance

$$\pi(y \mid \mathcal{M}_k) = \int \pi(\theta_k \mid \mathcal{M}_k)\pi(y \mid \theta_k, \mathcal{M}_k)d\theta_k,$$

  we can estimate the evidence using Monte Carlo methods.
- As a starter, we can consider

$$\pi(y \mid \mathcal{M}_k) \approx \frac{1}{N}\sum_{i=1}^{N} \pi(y \mid \theta_k^{(i)}, \mathcal{M}_k)$$

  where $\theta_k^{(i)}$ for $i \in \{1, \ldots, N\}$ are drawn i.i.d. from the prior $\pi(\theta_k \mid \mathcal{M}_k)$.

Estimation of the evidence

- How is this going to perform when the likelihood is peaky compared to the prior?

- We can design a proposal distribution q (e.g. using an approximate posterior sample), and consider

$$\pi(y \mid \mathcal{M}_k) \approx \frac{1}{N} \sum_{i=1}^{N} \frac{\pi(\theta_k^{(i)} \mid \mathcal{M}_k)\pi(y \mid \theta_k^{(i)}, \mathcal{M}_k)}{q(\theta_k^{(i)})}$$

  where $\theta_k^{(i)}$ for $i \in \{1, \ldots, N\}$ are drawn i.i.d. from q.

- This is an importance sampling strategy; the optimal distribution is proportional to the integrand, hence it is the posterior distribution itself.

Estimation of the evidence

- An approximate posterior sample, produced e.g. by MCMC, could thus be useful to estimate the evidence?
- Typically we cannot evaluate the corresponding "q".
- Can we write the normalizing constant as an integral with respect to the posterior?

$$\int \varphi(\theta)\pi(\theta \mid y)d\theta = \pi(y)$$

  for some choice φ? (dropping the index k for simplicity)
- Some people have proposed to use the following reasoning:

$$\int \varphi(\theta)\pi(\theta \mid y)d\theta = \pi(y)^{-1} \int \varphi(\theta)\pi(y \mid \theta)\pi(\theta)d\theta$$

  thus if $\varphi(\theta) = 1/\pi(y \mid \theta)$ we have

$$\int \varphi(\theta)\pi(\theta \mid y)d\theta = \pi(y)^{-1}.$$

Estimation of the evidence

- This leads to the monster

$$\pi(y)^{-1} \approx \frac{1}{N} \sum_{i=1}^{N} \pi(y \mid \theta^{(i)})^{-1}$$

  where $\theta^{(i)}$ for $i \in \{1, \ldots, N\}$ are approximating the posterior.

- By the law of large numbers, this is consistent when $N \to \infty$. Thus

$$\pi(y) \approx \left(\frac{1}{N} \sum_{i=1}^{N} \pi(y \mid \theta^{(i)})^{-1} \right)^{-1}$$

  is a consistent estimator too.

- What's wrong with it?

Toy example

- Consider a prior

$$\pi(\theta) = \mathcal{N}(\theta; 0, \sigma^2),$$

and a likelihood

$$\pi(y \mid \theta) = \mathcal{N}(\theta; 0, 1).$$

- For $\sigma^2 = 1$ and $\sigma^2 = 10^2$, we estimate Z using importance sampling from the prior and the harmonic mean estimator.

- We plot the obtained estimators as a function of the number of samples, to monitor convergence.

Numerical experiment

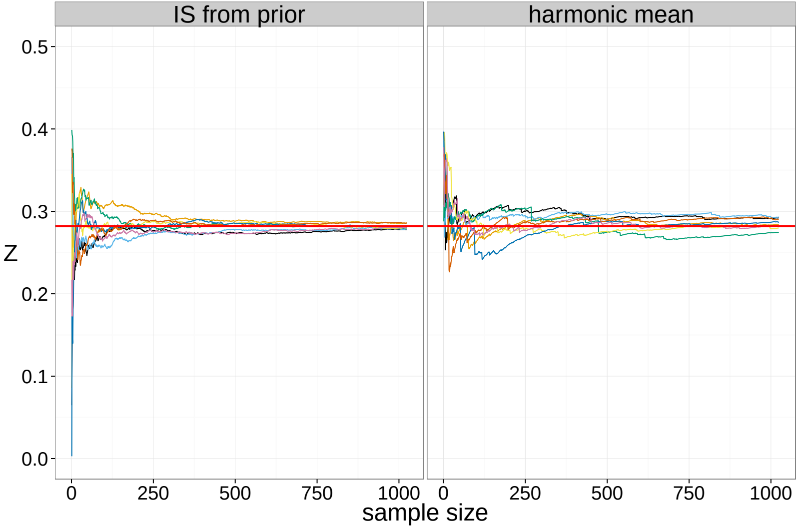

Figure: Normal model, prior variance = 1, likelihood variance = 1.

Numerical experiment

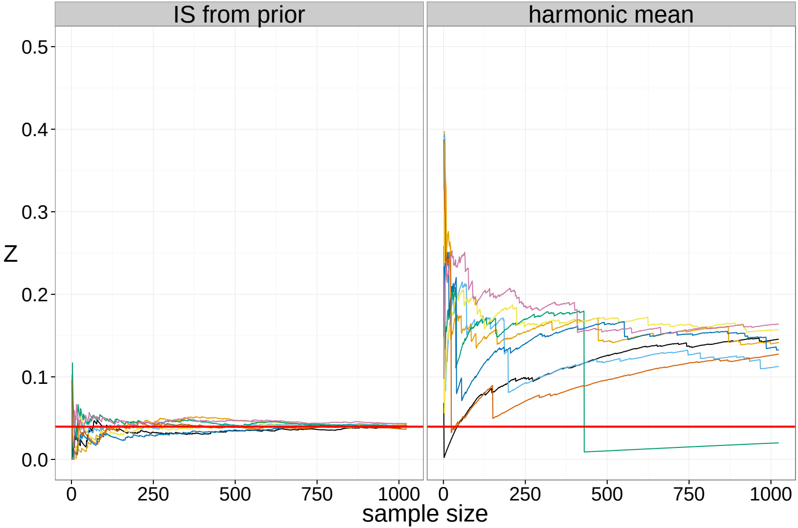

Figure: Normal model, prior variance $= 10^2$, likelihood variance $= 1$.

Estimation of the evidence

- We can also use rejection sampling to estimate the evidence.
- If we sample from q to target π, accept if

$$U_i \leq \frac{\widetilde{\pi}(X_i)}{\widetilde{M}\widetilde{q}(X_i)}$$

  where U_i is uniform and $X_i \sim q$.
- Then the probability of accepting a sample satisfies

$$\mathbb{P}\left(X \ \text{ accepted}\right) = \frac{1}{M} = \frac{Z_\pi}{Z_q \widetilde{M}}.$$

- On the toy example, sample from the prior and use $\widetilde{M} = 1$.

Numerical experiment

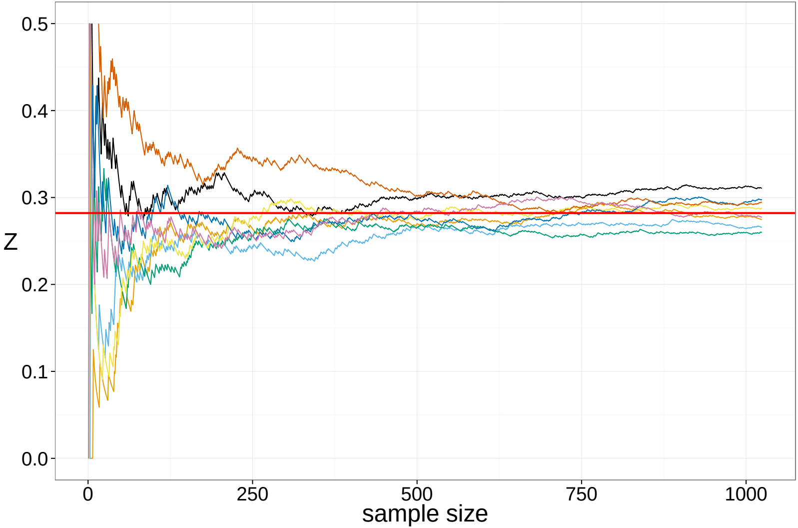

Figure: Normal model, prior variance = 1, likelihood variance = 1.

Numerical experiment

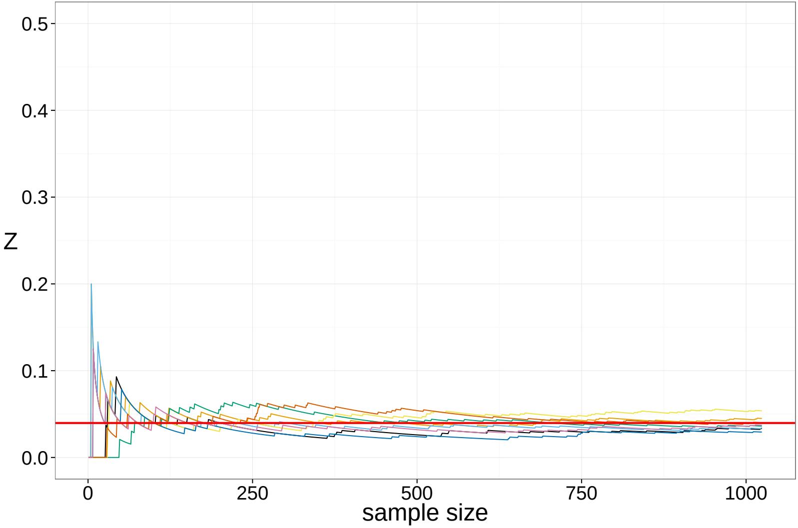

Figure: Normal model, prior variance $= 10^2$, likelihood variance $= 1$.

Estimation of the evidence

- Beyond those basic schemes, estimating the normalizing constant is an active area of research.

- Skilling. “Nested sampling.” Bayesian inference and maximum entropy methods in science and engineering 735 (2004): 395-405.

- Gelman and Meng. “Simulating normalizing constants: From importance sampling to bridge sampling to path sampling.” Statistical science (1998): 163-185.

- Del Moral, Doucet and Jasra. ”Sequential monte carlo samplers.” Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68.3 (2006): 411-436.

Slice Sampling!

Aside from classroom presentation, this is left as an opportunity for students to read well written paper:

Radford M. Neal "Slice sampling." The Annals of Statistics, Vol. 31, No. 3, 705-767 (2003).

1522 citations, and counting...