MT4531/5731: (Advanced) Bayesian Inference

Importance Sampling

Nicolò Margaritella

School of Mathematics and Statistics, University of St Andrews

–

Outline

Outline

Importance sampling (1)

- We again wish to obtain a sample from the posterior distribution $\pi(\theta|\boldsymbol{x})$, which we assume is difficult to do directly.
- However, suppose that we can easily sample from some other distribution $g(\theta)$ (where g has the same support as π).
- We initially consider the case where the constants of proporitonality are known for both π and g.
- Suppose further that we are interested in estimating of $\mathbb{E}_\pi(f(\theta))$.
- We would normally estimate $\mathbb{E}_\pi(f(\theta))$ by the usual MC estimate,

$$\hat{E}_\pi(f(\theta)) = \frac{1}{n}\sum_{i=1}^{n} f(\theta^i),$$

where θ^i would be samples from $\pi(\theta|\boldsymbol{x})$. But sampling from $\pi(\theta|\boldsymbol{x})$ is not easy!

Importance sampling (2)

- Note that,

$$\mathbb{E}_\pi(f(\theta)) = \int f(\theta)\pi(\theta|\boldsymbol{x})d\theta = \int \frac{f(\theta)\pi(\theta|\boldsymbol{x})}{g(\theta)} g(\theta)d\theta.$$

So, the expectation with respect to $\pi(\theta|\boldsymbol{x})$ can also be seen as an expectation with respect to $g(\theta)$, which is easy to sample from.

- Let $\theta^1, \theta^2, \ldots, \theta^n$ be a sample from $g(\theta)$.
- We estimate $\mathbb{E}_\pi(f(\theta))$ by

$$\hat{E}_\pi(f(\theta)) = \hat{E}_g(\frac{f(\theta)\pi(\theta|\boldsymbol{x})}{g(\theta)}) = \frac{1}{n}\sum_{i=1}^{n} \frac{\pi(\theta^i|\boldsymbol{x})}{g(\theta^i)} f(\theta^i) = \frac{1}{n}\sum_{i=1}^{n} w(\theta^i)f(\theta^i).$$

where we have now defined "importance" weights, for the θ^i sampled from $g(\theta)$,

$$w(\theta^i) = \frac{\pi(\theta^i|\boldsymbol{x})}{g(\theta^i)}.$$

Importance sampling - Advantages

- The advantage of this method is that we can use it for any densities provided that they are continuous and have the same support.
- In addition, it can be used even when the constant of proportionality for π is unknown.
- Assume that $\pi^*(\theta|\boldsymbol{x})$ is the known expression, up to proportionality. Then estimate,

$$\hat{E}_\pi(f(\theta)) = \frac{\sum_{i=1}^n w^*(\theta^i) f(\theta^i)/n}{\sum_{i=1}^n w^*(\theta^i)/n}.$$

where,

$$w^*(\theta^i) = \frac{\pi^*(\theta^i|\boldsymbol{x})}{g(\theta^i)}.$$

- This works because the denominator in $\hat{E}_\pi(f(\theta))$ is an importance sampling estimator of $\int \pi^*(\theta|\boldsymbol{x})d\theta$, and when this divides $\pi^*(\theta^i|\boldsymbol{x})$ we obtain an estimate of $\pi(\theta^i|\boldsymbol{x})$.

Importance sampling - Disadvantages

- The variance of the estimator can be very large, when $g$is not suitable for the problem at hand, leading to estimates that are not reliable.
- Without the constant of proportionality for π, the variance of the estimator can be even larger.
- Note that importance sampling can still be very efficient, and reduce the variance of Monte Carlo estimates.
- However, the choice of the g density is crucial, and the curve of an appropriate distribution g depends on $\pi(\theta|\boldsymbol{x})$ and the different functions of interest to be estimated.
- (In the example below, Importance sampling significantly reduces the variability of the estimate.)
- You can now see the part in Section 2.7.1 where Importance sampling is used to obtain MC estimates of posterior model probabilities.

Outline

Example (1)

- Suppose that we wish to estimate the probability $\mathbb{P}(\theta > 2)$, where θ follows a Cauchy distribution, with known density

$$\pi(\theta) = \frac{1}{\pi(1+\theta^2)}, \qquad \theta \in \mathbb{R}$$

  so we require

$$\int_2^\infty \pi(\theta) d\theta = \int_{-\infty}^\infty I(\theta > 2)\pi(\theta) d\theta,$$

  where I denotes the indicator function.

- We could simulate from the Cauchy distribution directly, but the variance of the ergodic average in this case, is very large.

Example (2)

- Alternatively, we observe that, for large θ, $\pi(\theta)$ is similar in behaviour to the density

$$g(\theta) = 2/\theta^2 \quad \theta > 2.$$

- We can simulate from this distribution directly using the method of inversion. Let $U^i \sim U(0,1)$ and set $\theta^i = 2/u^i$ for $i = 1, \ldots, n$ (you should check this!).
- Note that g does not have the same support as π, but g does have the same support as $I(\theta > 2)\pi(\theta)$ and so we can still use importance sampling. (Samples $\theta^i \leq 2$ from some other g would be removed from the MC estimation as $f(\theta^i) = I(\theta^i > 2) = 0$.)

Example (3)

- Suppose that we sample $\theta^1, \ldots, \theta^n$ from g. We define importance sampling weights,

$$w_i = \frac{\pi(\theta^i)}{g(\theta^i)} = \frac{(\theta^i)^2}{2\pi(1+(\theta^i)^2)}.$$

- Then, since each $\theta^i > 2$ we have that $f(\theta^i) = I(\theta^i > 2) = 1$ for all i.
- Thus, our estimator becomes:

$$\frac{1}{n}\sum_{i=1}^{n} \frac{(\theta^i)^2}{2\pi(1+(\theta^i)^2)},$$

where $\theta^i = 2/u^i$, for u^i a realised $U^i \sim U[0,1]$.
- This can be easily coded in, for example, R. See demonstration in lecture, and code uploaded on Moodle.

Sampling Importance Resampling (SIR)

- Sampling importance resampling (SIR) is an extension of Importance sampling, where we first sample using Importance sampling, and then resample with replacement the n simulated θ values, where the probability of simulating θ^i is given by w_i.
- The set of resampled values, denoted by $\phi^1, \ldots, \phi^n$, can then be used to obtain Monte Carlo estimates of summary statistics of interest.
- See Section 2.8.3 in the lecture notes for more details.

Finally...

- All direct sampling algorithms suffer from the problem of dimensionality.
- These methods can be generally implemented in one dimension (without too many problems) but become significantly more difficult (often impossible) to implement efficiently in higher dimensions.
- This is why we considered earlier Markov chain Monte Carlo, the most common approach for implementing Bayesian analyses and obtaining inference on the parameters of interest in multiple dimensions.