Approximate Bayesian Inference

Christos Dimitrakakis

November 28, 2023

Approximate Bayesian inference

The General problem

- Observations D.
- Nuisance variables z.
- Unknown parameter θ.
- Direct calculation of any of these terms can be infeasible:

$$\beta(\theta \mid D) = \frac{P_\theta(D)\beta(\theta)}{\sum_{\theta'} P_{\theta'}(D)\beta(\theta')}, \qquad P_\theta(D) = \sum_z P_\theta(D, z).$$

Common methods

- Monte Carlo
- Variational Bayes
- Approximate Bayesian Computation (ABC)
- Stochastic Variational Inference

Basic sampling theory

Inversion sampler

$F(u) = \mathbb{P}(x \geq u) = P(\{\omega : x(\omega) \geq u\})$ is the CDF of x.

- Sample u uniformly in $[0, 1]$
- Set $x = F^{-1}(u)$.

Rejection Sampler

- Input: Threshold ϵ, distribution Q
- Repeat:
- $\hat{x} \sim Q$.
- $u \sim \mathrm{Unif}[0, 1]$
- Until $u \leq P(\hat{x})/\epsilon Q(\hat{x})$.
- Return $\hat{x}$

Notes

- Useful for sampling from a known distribution P.
- Indirectly useful from sampling from unknown distributions.

Monte-Carlo sampling

$$\beta(B \mid D) = \frac{\int_B P_\theta(D) d\beta(\theta)}{\int_\Theta P_{\theta'}(D)\beta(\theta')}$$

We can approximate the integrals by sampling from the prior β:

$$\int_B P_\theta(D) d\beta(\theta) \approx \frac{1}{N} \sum_{n=1}^{N} \mathbb{I}\left\{\theta^{(n)} \in B\right\} P_{\theta^{(n)}}(D), \qquad \theta^{(n)} \sim \beta.$$

- Sampling from the prior is inefficient.
- The estimator has high bias and variance.
- So, we can use Markov Chain Monte Carlo. This lets us sample a sequence $\theta^{(n)}$ which converges asymptotically to $\beta(\theta^{(n)}|D)$.

Markov Chain Monte Carlo

MCMC for posterior sampling

- Form a Markov chain $P(\theta^{(n+1)} \mid \theta^{(n)}, D)$

MCMC for other latent variables

- Form a Markov chain $P(z^{(n+1)} \mid z^{(n)}, D)$

Metropolis-Hastings

Algorithm (symmetric version)

- Input: Proposal distribution $Q(x|x') = Q(x'|x)$
- At time n:
- $\hat{x} \sim Q(x|x^{(n)})$
- w.p. $P(\hat{x})/P(x^{(n)})$, $x^{(n+1)} = \hat{x}$ else $x^{(n+1)} = x^{(n)}$

Application to posterior sampling:

The denominator cancels out, leading to:

$$\frac{\beta(\theta' \mid D)}{\beta(\theta \mid D)} = \frac{P_{\theta'}(D)\beta(\theta')}{P_\theta(D)\beta(\theta)}$$

The only question is which proposal to use.

Metropolis-Hastings

Algorithm

- Input: Proposal distribution $Q(x|x')$ satisfying detailed balance, likelihood P.
- At time n:
- $\hat{x}|x^{(n)} \sim Q(x|x^{(n)})$
- With probability
$$\frac{P(\hat{x})Q(x^{(n)}|\hat{x})}{P(x^{(n)})Q(\hat{x}|x^{(n)})},$$
set $x^{(n+1)} = \hat{x}$
- Otherwise $x^{(n+1)} = x^{(n)}$

Application to posterior sampling:

The $\mathbb{P}_\beta(D)$ term cancels out, leading to:

$$\frac{\beta(\theta' \mid D)Q(\theta \mid \theta')}{\beta(\theta \mid D)Q(\theta' \mid \theta)} = \frac{P_{\theta'}(D)\beta(\theta')Q(\theta \mid \theta')}{P_\theta(D)\beta(\theta)Q(\theta' \mid \theta)}$$

M-H Theory

Stationary distribution

The Markov chain defined by the M-H algorithm must have a unique stationary distribution

$$\sigma = \sigma \boldsymbol{P},$$

where $\boldsymbol{P}$ is the transition kernel of the chain with

$$P_{ij} = \mathbb{P}(x^{(n+1)} = j \mid x^{(n)} = i).$$

In addition, $\lim_{n\to\infty} \boldsymbol{P}^k = 1\sigma$.

Sufficient conditions

- If the transition kernel satisfies detailed balance:

$$P(x'|x)\sigma(x) = P(x|x')\sigma(x')$$

then $\sigma(x)$ is a stationary distribution.

- If the Markov chain is ergodic then there is a unique σ.

The Gibbs sampler

This is used when we need to sample over only some variables $z_1, \ldots, z_k$, given some fixed variables x.

General algorithm

- Input: Factors $P(z_k \mid z_1, \ldots z_{k-1}, z_{k+1}, \ldots, z_K, x)$
- For $n \in [N]$:
- For $k \in [K]$: $z_k^{(n)} \sim P(z_k \mid z_1^{(n)}, \ldots z_{k-1}^{(n)}, z_{k+1}^{(n-1)}, \ldots, z_K^{(n-1)}, x)$

Application to posterior sampling with latent variables:

Latent variable z, parameter θ.

- Until convergence:
- $\theta^{(n)} \sim P(\theta \mid z^{(n-1)}, x)$
- $z^{(n)} \sim P(z \mid \theta^{(n)}, x)$

ABC: Approximate Bayesian Computation

When to use

- When we can sample from $P_\theta(D)$.
- When we cannot calculate $P_\theta(D)$.

A metric ρ over datasets

- $\rho(D, D')$ is distance between datasets.
- We can use that to define a rejection sampler

ABC Rejection Sampling

- **Input**: $\epsilon > 0$.
- Sample $\theta' \sim \beta(\theta)$
- Sample $D' \sim P_{\theta'}$.
- If $\rho(D, D') \leq \epsilon$, accept θ'

Theorem

If $\rho(D, D') = \|f(D) - f(D')\|$ and f is a sufficient statistic and $\epsilon = 0$ then ABC Rejection Sampling is exact.

Multi-platform

- STAN
- BUGS

Python

- PyMC3
- TensorFlow Probability
- PyStan
- Pyro (Torch)