MSA101/MVE187 2018 Lecture 15

Petter Mostad

Chalmers University

October 23, 2018

Approximate computing using Variational Bayes

- Assume we can write down the posterior $\pi_{post}(\theta \mid y) \propto_\theta \pi(y \mid \theta)\pi(\theta)$ up to a constant factor, but we are not able to compute or simulate from it.
- An alternative is to find and use an approximate function $q(\theta) \approx \pi_{post}(\theta \mid y)$.
- Specifically, we try to find the function $q \in \mathcal{Q}$ (where $\mathcal{Q}$ is some set of density functions defined on θ) minimizing the Kullback Leibler distance $KL[q||\pi_{post}]$.
- Note, if $\pi_{post} \in \mathcal{Q}$, the KL distance is minimized (with value 0) when $q = \pi_{post}$.
- Most commonly, $\mathcal{Q}$ consists of all functions factorizing over a specific partition of the variables in θ: Writing $\theta = (\theta_1, \dots, \theta_k)$, we have, for $q \in \mathcal{Q}$,

$$q((\theta_1, \theta_2, \dots, \theta_k)) = q_1(\theta_1)q_2(\theta_2)\cdots q_k(\theta_k)$$

Variational Bayes

We can write

$$\log \pi(y) = \log \pi(y, \theta) - \log \pi_{post}(\theta \mid y)$$

which, for any $q \in \mathcal{Q}$, gives rise to

$$\log \pi(y) = \mathcal{L}(q) + KL[q||\pi_{post}]$$

where

$$\begin{aligned} \mathcal{L}(q) &= \int q(\theta) \log\left(\frac{\pi(y, \theta)}{q(\theta)}\right) d\theta \\ KL[q||\pi_{post}] &= -\int q(\theta) \log\left(\frac{\pi_{post}(\theta \mid y)}{q(\theta)}\right) d\theta \end{aligned}$$

Writing $q(\theta) = \prod_{i=1}^{k} q_i(\theta_i)$, we get

$$\mathcal{L}(q) = \int \prod_{i=1}^{k} q_i(\theta_i) \log \pi(y, \theta)\, d\theta - \sum_{i=1}^{k} \int q_i(\theta_i) \log(q_i(\theta_i))\, d\theta_i$$

Variational Bayes

Selecting some $j \in \{1, \dots, k\}$, we get that the q_j *maximizing* $\mathcal{L}(q)$ subject to q_i being fixed for all $i \neq j$ is the q_j *maximizing*

$$\int q_j(\theta_j)\,\mathrm{E}_{-j}\left[\log \pi(y, \theta)\right]\, d\theta_j - \int q_j(\theta_j)\log(q_j(\theta_j))\, d\theta_j,$$

i.e., the q_j *minimizing* the KL distance $KL\left[q_j || w\right]$, where $w(\theta_j)$ is the density on θ_j whose log-density is, up to a constant, equal to $\mathrm{E}_{-j}\left[\log \pi(y, \theta)\right]$, where E_{-j} indicates the expectation under the density where all q_i, $i \neq j$, are fixed.

- The algorithm starts with some $q \in \mathcal{Q}$.
- For all $j \in \{1, \dots, k\}$, find q_j as the density proportional to $\exp\left(\mathrm{E}_{-j}\left[\log \pi(y, \theta)\right]\right)$.
- Find the densities q_j fulfilling these joint equations, either directly or using iteration.
- For more details, see Chapter 10 in Bishop.

Variational Bayes: Example

- Consider the following example:

$$\begin{aligned} y_1, \ldots, y_n &\sim \text{Normal}(\mu, \tau^{-1}) \\ \pi(\mu) &\propto 1 \\ \pi(\tau) &\propto 1/\tau \end{aligned}$$

- We know that the exact posterior is given by

$$\begin{aligned} \tau \mid y_1, \ldots, y_n &\sim \text{Gamma}\left(\frac{n-1}{2}, \frac{n-1}{2}s^2\right) \\ \mu \mid \tau, y_1, \ldots, y_n &\sim \text{Normal}\left(\bar{y}, (n\tau)^{-1}\right) \end{aligned}$$

  where s^2 is the sample variance.

- As an illustration, we find the Variational Bayes approximate posterior. Note:

$$\begin{aligned} \pi(y_1, \ldots, y_n, \mu, \tau) &\propto \frac{1}{\tau} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi/\tau}} \exp\left(-\frac{\tau}{2}(y_i - \mu)^2\right) \\ \log(\pi(y_1, \ldots, y_n, \mu, \tau)) &= C + \left(\frac{n}{2} - 1\right)\log\tau - \frac{\tau}{2}(n-1)s^2 - \frac{n\tau}{2}(\bar{y} - \mu)^2 \end{aligned}$$

Variational Bayes: Example

- Assume $q(\mu,\tau) = q_\mu(\mu)q_\tau(\tau)$; let E_μ and E_τ be the expectations under q_μ and q_τ, respectively. Taking E_τ, the logposterior becomes, as a function of μ,

$$C' - \frac{n}{2}\,\mathrm{E}_\tau(\tau)(\overline{y}-\mu)^2$$

corresponding to a Normal $\left(\overline{y}, (n\,\mathrm{E}_\tau(\tau))^{-1}\right)$ distribution for μ.

- Taking E_μ, the logposterior becomes, as a function of τ,

$$C + \left(\frac{n}{2}-1\right)\log\tau - \frac{\tau}{2}(n-1)s^2 - \frac{n\tau}{2}\,\mathrm{E}_\mu\left[(\overline{y}-\mu)^2\right]$$

corresponding to Gamma $\left(\frac{n}{2}, \frac{1}{2}\left((n-1)s^2 + n\,\mathrm{E}_\mu\left((\overline{y}-\mu)^2\right)\right)\right)$ for τ.

- Solving for the expectations, we get the Variational Bayes solution

$$\begin{aligned}\tau \mid y_1,\ldots,y_n &\sim \text{Gamma}\left(\frac{n}{2}, \frac{ns^2}{2}\right)\\ \mu \mid y_1,\ldots,y_n &\sim \text{Normal}\left(\overline{y}, \frac{s^2}{n}\right)\end{aligned}$$

Approximate Bayesian Computations (ABC)

- In our Bayesian inference methods so far, simulation from the posterior $\pi(\theta \mid x)$ is based on being able to compute, for various θ, $\pi(x \mid \theta)\pi(\theta)$, (at least up to a constant).
- What if we do not have a formula for the likelihoood $\pi(x \mid \theta)$?
- Example: Our stochastic "model" could be some very complex stochastic computer simulation program $R(\theta)$ producing a value for x given a value for θ.
- Idea for simulating from the posterior: Simulate θ from the prior, and keep only those θ with $R(\theta) = x$.

ABC cont.

- Example:
  - θ is binary with $P(\theta = 1) = 0.6$
  - x is binary with $\Pr(x = 1 \mid \theta = 1) = 0.9$, $\Pr(x = 1 \mid \theta = 0) = 0.1$
  - If the data is $x = 1$ then simulated values $\theta = 1$ would be kept with probability 0.9, simulated values $\theta = 0$ would be kept with probability 0.1.
  - We see the result corresponds to simulating $\theta = 1$ with probability $0.54/0.58 = 0.93$; correct according to Bayes formula.
- For continuous variables x we would get zero acceptance probability unless we replace the acceptance criterion $R(\theta) = x$ with $R(\theta) \approx x$.
- The most basic ABC algorithm defines a distance function ρ on the set where x lives, and an acceptance threshold ϵ. Then $\theta_1, \dots, \theta_n$ are simulated from the prior, and those θ_i with $\rho(R(\theta_i), x) \leq \epsilon$ are accepted.

ABC: Using sufficient statistic or similar

- In any (Bayesian) analysis, the likelihood $\pi(x \mid \theta)$ can be replaced by the corresponding likelihood $\pi(S(x) \mid \theta)$ of a sufficient statistic $S(x)$.
- Simple example: The likelihood of data $x = (x_1, \ldots, x_n)$, where $x_i \sim \text{Normal}(\theta, 1)$ can be replaced with the likelihood of $S(x) = \overline{x} \sim \text{Normal}(\theta, 1/n)$.
- If we can only simulate $x = R(\theta)$ we are unlikely to be able to prove that a statistic is sufficient. However, we may specify a function S we believe "summarizes" the features of the data that depend on θ. Then we replace x with $S(x)$.

ABC: Modelling the likelihood

- In realistic examples the acceptance rate or the accuracy will still be too low.
- A solution: Try to simulate the "correct" θ:
  - Example: If $R(\theta_1)$ and $R(\theta_2)$ are "on either side of x", maybe $(\theta_1 + \theta_2)/2$ will result in a value closer to x.
- Note: Targeting the simulation of θ in this way means the acceptance must be adjusted accordingly.