

Approximate Bayesian Inference

Václav Šmíd

March 4, 2022

Recapitulation

1. what principle is used in Jeffrey's prior
2. is improper prior $p(\theta) \propto 1$ always non-informative?
3. what is conjugate prior

What is the result of Bayesian inference?

The Bayes rule

$$p(\theta|Y) \propto p(Y|\theta)p(\theta)$$

in practice we appreciate moments, HPD regions:

$$E(\theta) = \int \theta p(\theta|Y) d\theta \qquad E(g(\theta)) = \int g(\theta) p(\theta|Y) d\theta$$

$$C : \{\mu \pm 3\sigma\}$$

These are available only for (most of) proper distributions.

What is the result of Bayesian inference?

The Bayes rule

$$p(\theta|Y) \propto p(Y|\theta)p(\theta)$$

in practice we appreciate moments, HPD regions:

$$E(\theta) = \int \theta p(\theta|Y) d\theta \quad E(g(\theta)) = \int g(\theta) p(\theta|Y) d\theta$$

$$C : \{\mu \pm 3\sigma\}$$

These are available only for (most of) proper distributions.

What if we do not have that?

1. choose different metrics
2. approximate $p(\theta|Y) \approx q(\theta|Y)$
3. combine the above

Alternative results of Bayesian inference

Maximum likelihood (ML): is Bayesian for improper prior $p(\theta) \propto 1$

$$\hat{\theta} = \arg \max p(Y|\theta)$$

Maximum a posteriori (MAP): is a point estimate for prior $p(\theta)$

$$\hat{\theta} = \arg \max p(Y|\theta)p(\theta)$$

Maximum marginal likelihood: is point estimate of a marginal $\theta = [\theta_1, \theta_2]$

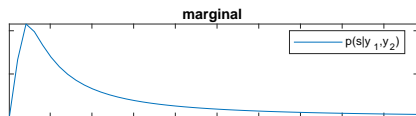
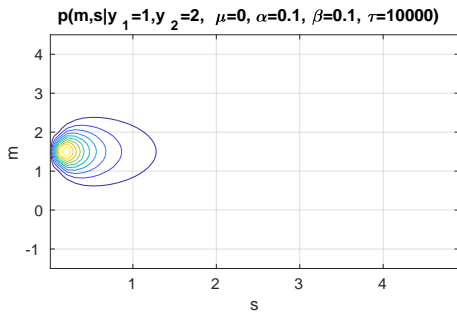
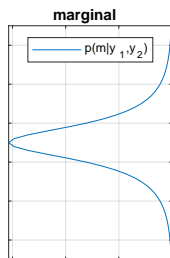
$$\hat{\theta}_1 = \arg \max_{\theta_1} p(\theta_1|Y) = \arg \max_{\theta_1} \int p(\theta_1, \theta_2|Y) d\theta_2$$

Empirical Bayes (EB): is a method of inferring $\theta = [\theta_1, \theta_2]$

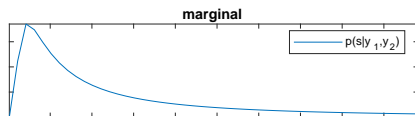
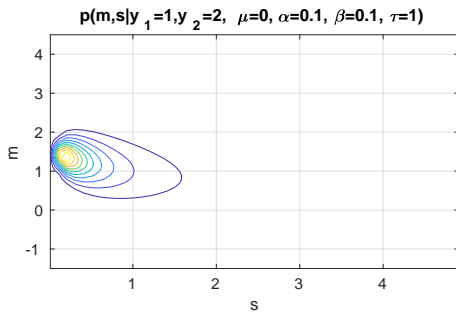
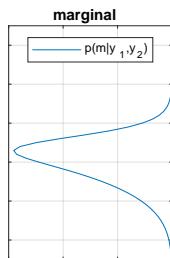
$$p(\theta_1|Y) \approx p(\theta_1|Y, \theta_2^*)$$

where $\theta_2^* = \arg \max_{\theta_2} p(\theta_2|Y)$.

Example



Example



EM algorithm classical form

Splits unknown vector into two parts:

1. quantity of interest θ (e.g. mean m)
2. missing data z (e.g. variance s)

General concept of missing data:

- ▶ nuisance parameter – we do not really care about it
- ▶ missing data – if we knew it, the solution would be easy
- ▶ latent variable

EM algorithm

General EM algorithm [Dempster, Laird, Rubin, 1977]. Maximum likelihood estimate:

$$\hat{\theta} = \arg \max_{\theta} \int p(Y|\theta, z)p(z)dz,$$

can be (approximately) found by alternating:

E-step: $q(\theta|\theta^{(i)}) = \int \log p(Y, z|\theta)p(z|\theta^{(i)}, Y)dz$

M-step: $\theta^{(i+1)} = \arg \max_{\theta} q(\theta|\theta^{(i)})$

where (i) is the iteration index, starting from initial condition $\theta^{(1)}$.

Necessary condition:

- ▶ we need to be able to compute

$$p(z|\theta^{(i)}, Y) = \frac{p(Y, z|\theta^{(i)})}{p(Y|\theta^{(i)})}$$

Toy example: $\theta = m$, $z = s$

Necessary condition:

$$p(z|\theta^{(i)}, Y) \equiv p(s|m^{(i)}, Y)$$

already done:

$$\begin{aligned} & p(s|m, y_1, y_2, \mu, \alpha, \beta, \tau) \\ & \propto \frac{1}{s^{\alpha_0+2}} \exp\left(-\frac{1}{2} \frac{(m-y_1)^2}{s} - \frac{1}{2} \frac{(m-y_2)^2}{s} - \frac{\beta_0}{s}\right) \\ & = i\mathcal{G}(\underbrace{\alpha_0+1}_{\alpha}, \underbrace{0.5(m-y_1)^2 + 0.5(m-y_2)^2 + \beta_0}_{\beta}) \\ & = i\mathcal{G}(\alpha, \beta) \end{aligned}$$

E-step: $\theta = m$, $z = s$

Proxy distribution

$$\begin{aligned}q(m|m^{(i)}) &= \int \log p(y_1, y_2, m, s) p(s|m^{(i)}|Y) ds \\ &= \mathbb{E}_{p(s|m^{(i)}, Y)}(\log p(y_1, y_2, m, s))\end{aligned}$$

using

$$p(y_1, y_2, m, s) \propto \frac{1}{s} \exp\left(-\frac{1}{2} \frac{(m - y_1)^2}{s} - \frac{1}{2} \frac{(m - y_2)^2}{s} - \frac{1}{2} \frac{(m - \mu)^2}{\tau}\right)$$

$$\log p(y_1, y_2, m, s) = c + \log s - \frac{1}{2} \frac{(m - y_1)^2}{s} - \frac{1}{2} \frac{(m - y_2)^2}{s} - \frac{1}{2} \frac{(m - \mu)^2}{\tau},$$

$$q(m|m^{(i)}) = c - \frac{1}{2} \mathbb{E}\left(\frac{1}{s}\right) [(m - y_1)^2 + (m - y_2)^2] - \frac{1}{2} \frac{(m - \mu)^2}{\tau}$$

$$\mathbb{E}\left(\frac{1}{s}\right) = \frac{\alpha}{\beta} \quad \hat{s} = \frac{\beta}{\alpha}$$

$$q(m|m^{(i)}) = c - \frac{1}{2} \left[\frac{(m - y_1)^2}{\hat{s}} + \frac{(m - y_2)^2}{\hat{s}} \right] - \frac{1}{2} \frac{(m - \mu)^2}{\tau}$$

M-step

New point estimate $m^{(i+1)} = \arg \max_m q(m|m^{(i)})$

1. set first derivative equal to 0
2. completion of squares

$$m^{(i+1)} = \left(\frac{1}{\tau} + \frac{2}{\hat{s}} \right)^{-1} \left(\frac{\mu}{\tau} + \frac{y_1 + y_2}{\hat{s}} \right),$$

based on \hat{s} estimate of conditional.

Algorithm:

1. set $i = 1$, choose $m^{(1)}$
2. compute

$$\alpha = \alpha_0 + 1,$$

$$\beta = 0.5(m^{(i)} - y_1)^2 + 0.5(m^{(i)} - y_2)^2 + \beta_0$$

3. Recompute $m^{(i+1)}$

Homework assignment

Note: meaning of variables can be swapped. E-step over m and M-step on s .

Working code for EM estimation of toy problem:

1. $\hat{m} = \arg \max p(m|y_1, y_2)$ (2 points)
2. $\hat{s} = \arg \max p(s|y_1, y_2)$ (8 points)

Distributional approximations

Principles of approximations:

Laplace Approximation: by Taylor approximation

$$\log p(\theta) \approx \log p(\hat{\theta}) + [\nabla \log p(\hat{\theta})]^T (\theta - \hat{\theta}) - \frac{1}{2} (\theta - \hat{\theta})^T H (\theta - \hat{\theta})$$
$$H = -\nabla \nabla \log p(\hat{\theta})$$

Divergence Minimization: approximation by optimization

Monte Carlo: approximation by sampling

Divergence minimization

We seek best approximation of intractable distribution $p(\theta)$ in the chosen class of parametric functions, $q(\theta|)$, such that

$$\theta^* = \arg \min_{\theta} D(p, q),$$

where $D(p, q)$ is a statistical divergence.

Divergence minimization

We seek best approximation of intractable distribution $p(\theta)$ in the chosen class of parametric functions, $q(\theta|)$, such that

$$\theta^* = \arg \min_{\theta} D(p, q),$$

where $D(p, q)$ is a statistical divergence.

Different results for different choices of: i) $q(\theta)$, and ii) D .

Divergence minimization

We seek best approximation of intractable distribution $p(\theta)$ in the chosen class of parametric functions, $q(\theta)$, such that

$$\theta^* = \arg \min_{\theta} D(p, q),$$

where $D(p, q)$ is a statistical divergence.

Different results for different choices of: i) $q(\theta)$, and ii) D .

Variational Bayes:

1. conditional independence $q(\theta_1, \theta_2) = q(\theta_1)q(\theta_2)$,
2. (reverse) Kullback-Leibler. $\text{KL}(q||p) \neq \text{KL}(p||q)$

Kullback Leibler divergence

Measure of divergence between two probability densities

$$KL(q||p) = E_q \left(\log \frac{q}{p} \right)$$

Not a metric!

$$E_q \left(\log \frac{q}{p} \right) \neq E_p \left(\log \frac{p}{q} \right)$$

also known as relative/free entropy

$$KL(q||p) = E_q (\log q) - E_q (\log p)$$

with properties:

1. $KL(q||p) \geq 0$,
2. $KL(q||p) = 0$, $\iff q = p$