

# Chapter IV

## Bayesian inference

### Choice of $g$ and the normal approximation

If wish to sample from  $f(\boldsymbol{\theta}|\mathbf{x})$ , could choose  $g(\boldsymbol{\theta}) = f(\boldsymbol{\theta})$ . If we do not know  $f(\mathbf{x})$  then have

$$\tilde{w}_i = f(\mathbf{x}|\boldsymbol{\theta}_i) \quad \text{and} \quad w(\boldsymbol{\theta}_i) = \frac{f(\mathbf{x}|\boldsymbol{\theta}_i)}{\sum_{i=1}^n f(\mathbf{x}|\boldsymbol{\theta}_i)}.$$

- ▶ Simulate  $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_n$  from the prior  $f(\boldsymbol{\theta})$
- ▶ Set  $\tilde{w}_i = f(\mathbf{x}|\boldsymbol{\theta}_i)$
- ▶ Set  $w_i = \tilde{w}_i / \sum \tilde{w}_i$  and estimate  $\mathbb{E}(\boldsymbol{\theta}|\mathbf{x})$  by

$$\sum_{i=1}^n w_i \boldsymbol{\theta}_i$$

This is inefficient if the prior is very different to the posterior as we will spend too much time sampling  $\boldsymbol{\theta}_i$  where the likelihood is very small, and so the weights  $w(\boldsymbol{\theta}_i)$  will also be very small.

If this is the case, then the effective sample size will be small, and our estimates of  $E(\boldsymbol{\theta}|\mathbf{x})$  will be dominated by just a few of the  $\boldsymbol{\theta}$  samples.

Unnormalised densities frequently occur when we are doing Bayesian inference.

Suppose we are interested in some posterior expectation, for example, the posterior mean:

$$I = \mathbb{E}(\theta|x) = \int \theta f(\theta|x) d\theta$$

where

$$f(\theta|x) = \frac{f(\theta)f(x|\theta)}{f(x)} \quad \text{by Bayes theorem.}$$

The denominator  $f(x) = \int f(\theta)f(x|\theta)dx$  is often intractable and unknown, and so we instead work with the unnormalised density

$$f_1(\theta|x) = f(\theta)f(x|\theta) = \text{prior} \times \text{likelihood}$$

A more efficient alternative to using the prior distribution for  $g$ , is to build a normal approximation to the posterior and use this as  $g$

Let  $h(\boldsymbol{\theta}) = \log f(\boldsymbol{\theta}|\mathbf{x})$ . Now define  $\mathbf{m}$  to be posterior mode of  $\boldsymbol{\theta}$ , so  $\mathbf{m}$  maximises both  $f(\boldsymbol{\theta}|\mathbf{x})$  and  $h(\boldsymbol{\theta})$ .

We may need to use numerical optimisation (such as the `optim` command in R) to find  $\mathbf{m}$ , but note that we don't need to know  $f(\mathbf{x})$  to do this.

We can then use a Taylor expansion of  $h(\boldsymbol{\theta})$  around  $\mathbf{m}$

$$h(\boldsymbol{\theta}) = h(\mathbf{m}) + (\boldsymbol{\theta} - \mathbf{m})^T \mathbf{h}'(\mathbf{m}) + \frac{1}{2}(\boldsymbol{\theta} - \mathbf{m})^T M(\boldsymbol{\theta} - \mathbf{m}) + \dots$$

to build a Gaussian approximation to the posterior (known as the Laplace approximation).

Here,  $\mathbf{h}'(\mathbf{m})$  the vector of first derivatives of  $h(\boldsymbol{\theta})$ , and  $M$  the matrix of second derivatives of  $h(\boldsymbol{\theta})$ , both evaluated at  $\boldsymbol{\theta} = \mathbf{m}$ .

## Assessing convergence

Since  $\mathbf{m}$  maximises  $h(\mathbf{m})$  we have  $h'(\mathbf{m}) = \mathbf{0}$ . Hence

$$f(\boldsymbol{\theta}|\mathbf{x}) = \exp\{h(\boldsymbol{\theta})\} \simeq \exp\{h(\mathbf{m})\} \exp\left\{-\frac{1}{2}(\boldsymbol{\theta} - \mathbf{m})^T V^{-1}(\boldsymbol{\theta} - \mathbf{m})\right\}, \quad (1)$$

where  $-V^{-1} = M$ .

Thus, our approximation of  $f(\boldsymbol{\theta}|\mathbf{x})$  is a multivariate normal distribution, mean vector  $\mathbf{m}$ , variance matrix  $-M^{-1}$ . This will be a good approximation if posterior mass concentrated around  $\mathbf{m}$ .

NB: We do not need  $f(\mathbf{x})$  to obtain  $M$ , since

$$h(\boldsymbol{\theta}) = \log f(\boldsymbol{\theta}|\mathbf{x}) = \log f(\boldsymbol{\theta}) + \log f(\mathbf{x}|\boldsymbol{\theta}) - \log f(\mathbf{x}),$$

so  $\log f(\mathbf{x})$  will disappear when we differentiate  $h(\boldsymbol{\theta})$ .

Suppose we wish to estimate  $\mathbb{E}\{r(\boldsymbol{\theta})|\mathbf{x}\}$  for some  $r(\boldsymbol{\theta})$ . If  $f(\mathbf{x})$  known, then

$$\hat{\mathbb{E}}\{r(\boldsymbol{\theta})|\mathbf{x}\} = \frac{1}{n} \sum_{i=1}^n r(\boldsymbol{\theta}_i) w(\boldsymbol{\theta}_i),$$

and can use central limit theorem to obtain a confidence interval for  $\mathbb{E}\{r(\boldsymbol{\theta})|\mathbf{x}\}$ , as in MC integration.

We can check our estimate by

- 1) Increasing the sample size  $n$  to check the stability of any estimate.
- 2) Increasing the standard deviation in the  $g(\boldsymbol{\theta})$  density, to check stability to the choice of  $g$ , e.g., if we're using a normal approximation, we could multiply  $V$  by 4 etc.

## Example: leukaemia data

Patients suffering from leukaemia are given a drug, 6-mercaptopurine (6-MP), and the number of days  $x_i$  until freedom from symptoms is recorded of patient  $i$ :

6\*, 6, 6, 6, 7, 9\*, 10\*, 10, 11\*, 13, 16, 17\*,

19\*, 20\*, 22, 23, 25\*, 32\*, 32\*, 34\*, 35\*.

A \* denotes censored observation.

Will suppose that time  $x$  to the event of interest follows a *Weibull* distribution:

$$f(x|\alpha, \beta) = \alpha\beta(\beta x)^{\alpha-1} \exp\{-(\beta x)^\alpha\}$$

for  $x > 0$ .

For censored observations, we have

$$P(x > t|\alpha, \beta) = \exp\{-(\beta t)^\alpha\}.$$

## Example: leukaemia data

### Likelihood

Define

- ▶  $d$ : number of uncensored observations,
- ▶  $\sum_u \log x_i$ : sum of logs of all uncensored observations.

Writing  $\boldsymbol{\theta} = (\alpha, \beta)^T$ , the log likelihood is then given by

$$\log f(\mathbf{x}|\boldsymbol{\theta}) = d \log \alpha + \alpha d \log \beta + (\alpha - 1) \sum_u \log x_i - \beta^\alpha \sum_{i=1}^n x_i^\alpha.$$

Suppose our prior distributions for  $\alpha$  and  $\beta$  are both exponential with

$$\begin{aligned} f(\alpha) &= 0.001 \exp(-0.001\alpha), \\ f(\beta) &= 0.001 \exp(-0.001\beta). \end{aligned}$$

## Example: leukaemia data

Building an approximation to the posterior

1) **Obtain the posterior mode of  $\theta$ .** Maximise log posterior, i.e.

$$h(\theta) = d \log \alpha + \alpha d \log \beta + (\alpha - 1) \sum_u \log x_i - \beta^\alpha \sum_{i=1}^n x_i^\alpha - 0.001\alpha - 0.001\beta +$$

for some constant  $K$ .

In R, we can find the mode to be  $\mathbf{m} = (1.354, 0.030)$  using the `optim` command.

3) **Obtain the normal approximation to use as  $g(\theta)$ .**

$g(\theta)$ : bivariate normal, mean  $\mathbf{m}$ , variance matrix  $V = -M^{-1}$ :

$$\theta \sim N \left\{ \begin{pmatrix} 1.354 \\ 0.030 \end{pmatrix}, \begin{pmatrix} 0.0334 & 0.0003 \\ 0.0003 & 0.00006 \end{pmatrix} \right\}$$

4) **Sample  $\theta_1, \dots, \theta_n$  from  $g(\theta)$  and compute the importance weights  $w(\theta_1), \dots, w(\theta_n)$ .** The weights are given by

$$w(\theta_i) = \frac{\tilde{w}(\theta_i)}{\sum_{i=1}^n \tilde{w}(\theta_i)}, \quad \text{with} \quad \tilde{w}(\theta_i) = \frac{f(\theta_i)f(\mathbf{x}|\theta_i)}{g(\theta_i)}$$

NB the Gaussian approximation may give us negative samples. Since  $\alpha > 0$  and  $\beta > 0$ , we should simply discard negative  $\theta$  values, i.e., use a truncated normal density for  $g(\theta)$ .

Note that when we compute  $w(\theta_i)$ , it is not necessary to rescale  $g(\theta)$  so that it integrates to 1, as any normalising constant in  $g(\theta)$  will cancel.

2) **Derive the matrix of second derivatives of  $h(\theta)$ .**

$$M = \begin{pmatrix} \frac{\partial^2}{\partial \alpha^2} h(\theta) & \frac{\partial^2}{\partial \alpha \partial \beta} h(\theta) \\ \frac{\partial^2}{\partial \alpha \partial \beta} h(\theta) & \frac{\partial^2}{\partial \beta^2} h(\theta) \end{pmatrix},$$

evaluated at  $\theta = \mathbf{m}$ .

$$\frac{\partial^2}{\partial \alpha^2} h(\theta) = -\frac{d}{\alpha^2} - \sum (\beta x_i)^\alpha (\log(\beta x_i))^2$$

$$\frac{\partial^2}{\partial \beta^2} h(\theta) = \frac{1}{\beta^2} \left\{ \beta^\alpha \alpha (1 - \alpha) \sum_{i=1}^n x_i^\alpha - d\alpha \right\},$$

$$\frac{\partial^2}{\partial \alpha \partial \beta} h(\theta) = \frac{1}{\beta} \left[ d - \beta^\alpha \left\{ \alpha \log \beta \sum_{i=1}^n x_i^\alpha + \sum_{i=1}^n x_i^\alpha + \alpha \sum_{i=1}^n x_i^\alpha \log x_i \right\} \right]$$

$$M = \begin{pmatrix} -31.618 & 175.442 \\ 175.442 & -18806.085 \end{pmatrix}.$$

5) **Estimate the posterior mean of  $\theta$**

We compute the estimate

$$\hat{E}(\theta|\mathbf{x}) = \sum_{i=1}^n \theta_i w(\theta_i).$$

In R, with  $n = 100000$ , this gives  $\hat{E}(\theta|\mathbf{x}) = (1.346, 0.031)^T$ .

6) **Check for convergence**

We repeat steps 4 and 5 with more dispersion in  $g(\theta)$ :

$g(\theta)$	$\hat{E}(\theta \mathbf{x})$
$N(\mathbf{m}, V)$	$(1.346, 0.031)^T$
$N(\mathbf{m}, 4V)$	$(1.384, 0.031)^T$
$N(\mathbf{m}, 16V)$	$(1.380, 0.031)^T$

Finally, double the sample size (no effect observed).

For percentiles, we can do resampling in R.

See computer class 5 for more details and code to implement this approach.