

# *Introducing Markov Chain Monte Carlo*

Mark Andrews & Thom Baguley

## Posterior distributions

- In general, given observed data  $D$  and a model  $\Omega$ , the posterior distribution over the parameters  $\theta$  of the model is

$$P(\theta|D, \Omega) = \frac{\overbrace{P(D|\theta)}^{\text{Likelihood}} \overbrace{P(\theta|\Omega)}^{\text{Prior}}}{\underbrace{\int P(D|\theta)P(\theta|\Omega) d\theta}_{\text{Marginal likelihood}}}.$$

where the *marginal likelihood* gives the likelihood of the model given the observed data.

- Given the posterior distribution  $P(\theta|D, \Omega)$ , our aim is often to characterise this distribution in terms of e.g. its mean, variance, etc.
- Likewise, we may aim to calculate *posterior predictive* distributions such as

$$P(x_{\text{new}}|D, \Omega) = \int P(x_{\text{new}}|\theta, \Omega)P(\theta|D, \Omega) d\theta.$$

## *Sampling from posterior distributions*

- ▶ In only rare situations can we determine the characteristics of the posterior distribution, or calculate posterior predictive distributions, in closed form.
- ▶ However, in general, if we can draw samples from  $P(\theta|D, \Omega)$  then we can approximate, e.g., the mean of the distribution by

$$\langle \theta \rangle = \int \theta P(\theta|D, \Omega) \approx \frac{1}{N} \sum_{i=1}^N \tilde{\theta}_i,$$

or the posterior predictive distribution by

$$P(x_{\text{new}}|D, \Omega) = \int P(x_{\text{new}}|\theta, \Omega) P(\theta|D, \Omega) d\theta \approx \frac{1}{N} \sum_{i=1}^N P(x_{\text{new}}|\tilde{\theta}_i, \Omega),$$

where

$$\tilde{\theta}_1, \tilde{\theta}_2 \dots \tilde{\theta}_N$$

are samples from  $P(\theta|D, \Omega)$ .

## Rejection sampling

- ▶ Rejection sampling is one of the simplest methods for sampling from posterior distributions.
- ▶ Let us denote  $P(\theta|D, \Omega)$  by  $f(\theta)$ .
- ▶ First we sample from  $\tilde{\theta}$  from another (simpler) distribution  $g(\theta)$ .
- ▶ The distribution  $g(\theta)$  can be any function so long as there exists a constant  $M$  such that

$$M \cdot g(\theta) \geq f(\theta),$$

for all possible values of  $\theta$ .

- ▶ Then draw  $u \sim U(0, 1)$ , a random sample from a uniform distribution between 0 and 1.
- ▶ If

$$u \leq \frac{f(\tilde{\theta})}{M \cdot g(\tilde{\theta})}$$

then keep  $\tilde{\theta}$ .

- ▶ Continue until  $N$  samples are collected.

# Gibbs sampling

- In a multivariate probability distribution, e.g.

$$P(x, y, z),$$

the univariate conditional distributions, e.g.  $P(x|y = \tilde{y}, z = \tilde{z})$ , may be straightforward to sample from.

- In Gibbs sampling, we set e.g.  $y$  and  $z$  to initial values  $\tilde{y}_0$  and  $\tilde{z}_0$  and then sample

$$\tilde{x}_0 \sim P(x|y = \tilde{y}_0, z = \tilde{z}_0),$$

$$\tilde{y}_1 \sim P(y|x = \tilde{x}_0, z = \tilde{z}_0),$$

$$\tilde{z}_1 \sim P(z|x = \tilde{x}_0, y = \tilde{y}_1),$$

and so on.

- After convergence, the samples e.g.  $\{\tilde{x}_N, \tilde{y}_N, \tilde{z}_N\}$  are draws from  $P(x, y, z)$ .

# Metropolis Hastings

- ▶ Let us denote  $P(\theta|D, \Omega)$  by  $f(\theta)$ .
- ▶ We sample from a symmetric *proposal* distribution  $Q(\cdot|\cdot)$ .
- ▶ We start with an initial  $\tilde{\theta}_0$ , and sample

$$\tilde{\theta} \sim Q(\theta|\tilde{\theta}_0).$$

- ▶ We then accept  $\tilde{\theta}$  with probability

$$\alpha = \min \left( 1.0, \frac{f(\tilde{\theta})}{f(\tilde{\theta}_0)} \right).$$

- ▶ After convergence, the accepted samples are draws from the distribution  $f(\theta)$ .
- ▶ For Metropolis Hastings, the distribution  $f(\theta)$  need be only known up to a proportional constant.