Pseudo-Marginal Markov Chain Monte Carlo for Bayesian Inference

Angus Lewis

An approach for Bayesian inference

- We want to fit the model to the data
  - i.e. find parameters θ^* that explain the data
- Find the posterior distribution,

$$P(\theta|\mathbf{x}) = \frac{L(\theta)P(\theta)}{P(\mathbf{x})} \propto L(\theta)P(\theta).$$

Where,

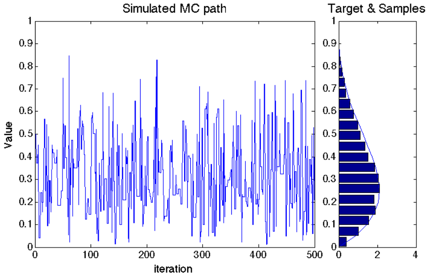
- $P(\mathbf{x}) = \int_\Theta L(\theta)P(\theta)d\theta$, which often cannot be computed,
- $L(\theta) = P(\mathbf{x}|\theta)$ is the likelihood,
- and $P(\theta)$ is the prior distribution.

Markov Chain Monte Carlo (MCMC)

How to deal with $P(\theta|\mathbf{x}) \propto L(\theta)P(\theta)$.

- Construct a Markov Chain that has the same stationary distribution as the posterior

$$\pi(\theta) = P(\theta|\mathbf{x})$$



Makov Chain Monte Carlo (MCMC)

Metropolis Hasting Algorithm

1. Initialise, $n = 0$ and θ_0.
2. Given the current state of the chain θ_n sample a candidate θ' from predetermined proposal $R(\theta'|\theta_n)$.
3. With probability

$$\alpha(\theta_n, \theta') = \min\left\{\frac{\frac{L(\theta')P(\theta')}{P(\mathbf{x})}R(\theta'|\theta_n)}{\frac{L(\theta_n)P(\theta_n)}{P(\mathbf{x})}R(\theta_n|\theta')}, 1\right\}$$
$$= \min\left\{\frac{L(\theta')P(\theta')R(\theta'|\theta_n)}{L(\theta_n)P(\theta_n)R(\theta_n|\theta')}, 1\right\}$$

   set $\theta_{n+1} = \theta'$
   else set $\theta_{n+1} = \theta_n$
4. Set $n = n + 1$ and return to 2

MCMC stationary distribution

- Transition density is

$$\kappa(\theta', \theta_n) = \alpha(\theta_n, \theta')R(\theta'|\theta_n) + (1 - \alpha^*(\theta_n))\delta_{\theta_n}(\theta'),$$

where $\alpha(\theta_n)^* = \int \alpha(\theta_n, \theta)R(\theta|\theta_n))d\theta$

MCMC stationary distribution

- Transition density is

$$\kappa(\theta', \theta_n) = \alpha(\theta_n, \theta')R(\theta'|\theta_n) + (1 - \alpha^*(\theta_n))\delta_{\theta_n}(\theta'),$$

  where $\alpha(\theta_n)^* = \int \alpha(\theta_n, \theta)R(\theta|\theta_n))d\theta$

- The transition density, κ, satisfies the Global Balance Equations,

$$P(\theta_n|\mathbf{x})\kappa(\theta', \theta_n) = P(\theta'|\mathbf{x})\kappa(\theta_n, \theta').$$

MCMC stationary distribution

- Transition density is

$$\kappa(\theta', \theta_n) = \alpha(\theta_n, \theta')R(\theta'|\theta_n) + (1 - \alpha^*(\theta_n))\delta_{\theta_n}(\theta'),$$

  where $\alpha(\theta_n)^* = \int \alpha(\theta_n, \theta)R(\theta|\theta_n))d\theta$
- The transition density, κ, satisfies the Global Balance Equations,

$$P(\theta_n|\mathbf{x})\kappa(\theta', \theta_n) = P(\theta'|\mathbf{x})\kappa(\theta_n, \theta').$$

- In addition, if

$$P(\alpha(\theta_n, \theta') < 1|\theta_n) > 0 \text{ and } R(\theta'|\theta_n) > 0 \quad \forall \quad \theta', \theta_n$$

MCMC stationary distribution

- Transition density is

$$\kappa(\theta', \theta_n) = \alpha(\theta_n, \theta')R(\theta'|\theta_n) + (1 - \alpha^*(\theta_n))\delta_{\theta_n}(\theta'),$$

  where $\alpha(\theta_n)^* = \int \alpha(\theta_n, \theta)R(\theta|\theta_n))d\theta$
- The transition density, κ, satisfies the Global Balance Equations,

$$P(\theta_n|\mathbf{x})\kappa(\theta', \theta_n) = P(\theta'|\mathbf{x})\kappa(\theta_n, \theta').$$

- In addition, if

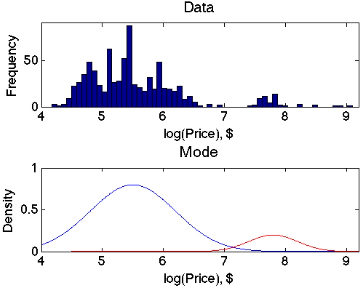
$$P(\alpha(\theta_n, \theta') < 1|\theta_n) > 0 \text{ and } R(\theta'|\theta_n) > 0 \quad \forall \quad \theta', \theta_n$$

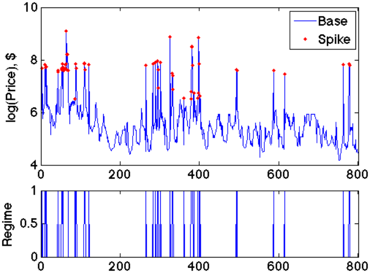
- Then $P(\theta|\mathbf{x})$ is the stationary distribution of the chain.

A state space model

- We have a model, M
  - The model has some unknown parameters, θ
  - And also some unkown data **R** which is a Markov chain

E.g.





Bayesian inference for state space models

- Find the posterior distribution,

$$P(\theta|\mathbf{x}) = \frac{L(\theta)P(\theta)}{P(\mathbf{x})} \propto L(\theta)P(\theta).$$

Where,

- $P(\mathbf{x}) = \int_{\Theta} L(\theta)P(\theta)d\theta$, which often cannot be computed,
- $L(\theta) = P(\mathbf{x}|\theta) = \sum_{\mathbf{R}} P(\mathbf{x}|\mathbf{R},\theta)P(\mathbf{R}|\theta)$ is the likelihood and is also hard to compute,
- and $P(\theta)$ is the prior distribution.

Bayesian inference for state space models

- Find the posterior distribution,

$$P(\theta, \mathbf{R}|\mathbf{x}) = \frac{L(\theta, \mathbf{R})P(\theta, \mathbf{R})}{P(\mathbf{x})} \propto L(\theta, \mathbf{R})P(\theta, \mathbf{R}).$$

Where,

- $P(\mathbf{x}) = \int_{\Theta} L(\theta)P(\theta)d\theta$, which often cannot be computed,
- $L(\theta, \mathbf{R}) = P(\mathbf{x}|\theta, \mathbf{R})$ is the likelihood,
- and $P(\theta)$ is the prior distribution.

Makov Chain Monte Carlo (MCMC)

Metropolis Hasting Algorithm agian

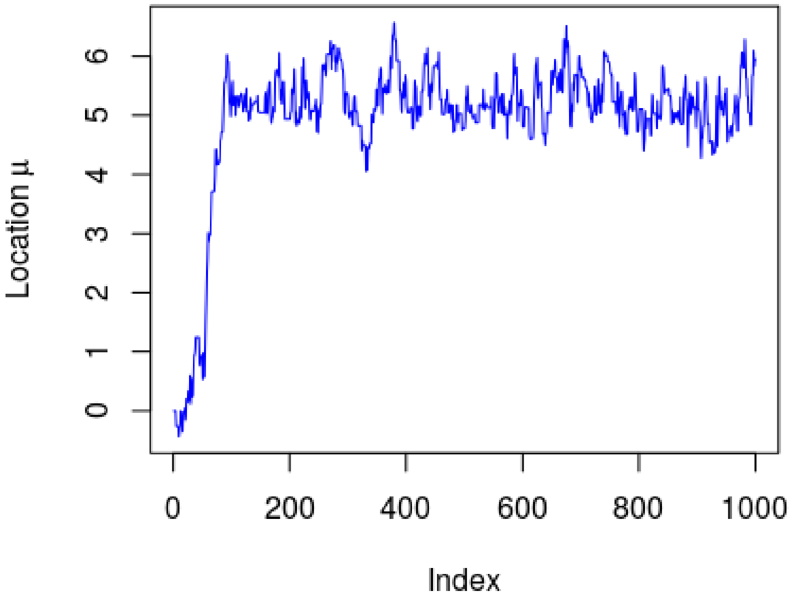
1. Initialise, $n = 0$ and $(\theta_0, \mathbf{R_0})$.
2. Given the current state of the chain $(\theta_n, \mathbf{R_n})$ sample a candidate $(\theta', \mathbf{R}')$ from predetermined proposal $R(\theta', \mathbf{R}'|\theta_n, \mathbf{R_n})$.
3. With probability

$$\alpha = \min\left\{\frac{L(\theta', \mathbf{R}')P(\theta', \mathbf{R}')R(\theta', \mathbf{R}'|\theta_n, \mathbf{R_n})}{L(\theta_n, \mathbf{R_n})P(\theta_n, \mathbf{R_n})R(\theta_n, \mathbf{R_n}|\theta', \mathbf{R}')}, 1\right\}$$

   set $\theta_{n+1} = \theta'$
   else set $\theta_{n+1} = \theta_n$
4. Set $n = n + 1$ and return to 2

This mixes slowly unless a good proposal is chosen

Mixing



Bayesian inference for state space models

- Find the posterior distribution,

$$P(\theta|\mathbf{x}) = \frac{L(\theta)P(\theta)}{P(\mathbf{x})} \propto L(\theta)P(\theta).$$

Where,

- $P(\mathbf{x}) = \int_{\Theta} L(\theta)P(\theta)d\theta$, which often cannot be computed,
- $L(\theta) = P(\mathbf{x}|\theta) = \sum_{\mathbf{R}} P(\mathbf{x}|\mathbf{R},\theta)P(\mathbf{R}|\theta)$ is the likelihood and is also hard to compute,
- and $P(\theta)$ is the prior distribution.

The Pseudo Marginal Part

- The likelihood:

$$L(\theta) = P(\mathbf{x}|\theta) = \sum_{\mathbf{R}} P(\mathbf{x}|\mathbf{R},\theta)P(\mathbf{R}|\theta),$$

  is hard to compute.
- What if we could sum over some of the **R**'s instead.
  - Randomly sample $\mathbf{R}_j$, $j = 1, 2, ..., m$.
  - Calculate

$$\hat{L}(\theta) = \frac{1}{m}\sum_{j=1}^{m} P(\mathbf{x}|\mathbf{R}_j,\theta)P(\mathbf{R}_j|\theta)$$

A Pseudo Marginal MCMC Algoritm

1. Initialise, $n = 0$ and θ_0.
2. Given the current state of the chain θ_n sample a candidate θ' from predetermined proposal $R(\theta'|\theta_n)$.
3. Sample some $\mathbf{R}_j$'s and calculate

$$\hat{L}(\theta') = \frac{1}{m}\sum_{j=1}^{m} P(\mathbf{x}|\mathbf{R}_j, \theta')P(\mathbf{R}_j|\theta')$$

4. With probability

$$\alpha(\theta_n, \theta') = \min\left\{\frac{\hat{L}(\theta')P(\theta')R(\theta'|\theta_n)}{\hat{L}(\theta_n)P(\theta_n)R(\theta_n|\theta')}, 1\right\}$$

   set $\theta_{n+1} = \theta'$, else set $\theta_{n+1} = \theta_n$. Save $\hat{L}(\theta_{n+1})$.

   It is important that $\hat{L}(\theta_{n+1})$ is saved.

5. Set $n = n + 1$ and return to 2

Proof of Pseudo Marginal Algorithm

- Define $W = \frac{\hat{L}(\theta)}{L(\theta)}$, the noise in the estimate of the likelihood, with density $p(w)$

Proof of Pseudo Marginal Algorithm

- Define $W = \frac{\hat{L}(\theta)}{L(\theta)}$, the noise in the estimate of the likelihood, with density $p(w)$
- Assume $\mathbb{E}[W|\theta] = c$ where $c > 0$, so that $\mathbb{E}[\hat{L}(\theta)] = cL(\theta)$

Proof of Pseudo Marginal Algorithm

- Define $W = \frac{\hat{L}(\theta)}{L(\theta)}$, the noise in the estimate of the likelihood, with density $p(w)$
- Assume $\mathbb{E}[W|\theta] = c$ where $c > 0$, so that $\mathbb{E}[\hat{L}(\theta)] = cL(\theta)$
- Consider each iteration as a joint update of (θ', w'), with proposal density $P((\theta', w')|(\theta_n, w_n)) = R(\theta'|\theta_n)p(w')$

Proof of Pseudo Marginal Algorithm

- Define $W = \frac{\hat{L}(\theta)}{L(\theta)}$, the noise in the estimate of the likelihood, with density $p(w)$
- Assume $\mathbb{E}[W|\theta] = c$ where $c > 0$, so that $\mathbb{E}[\hat{L}(\theta)] = cL(\theta)$
- Consider each iteration as a joint update of (θ', w'), with proposal density $P((\theta', w')|(\theta_n, w_n)) = R(\theta'|\theta_n)p(w')$
- Then the acceptance ratio is

$$\begin{aligned}\alpha &= \min\left\{\frac{\hat{L}(\theta')p(\theta')P((\theta_n, w_n)|(\theta', w'))}{\hat{L}(\theta_n)p(\theta_n)P((\theta', w')|(\theta_n, w_n))}, 1\right\} \\ &= \min\left\{\frac{L(\theta')w'p(w'|\theta')}{L(\theta_n)w_n p(w_n|\theta_n)}\frac{p(\theta')P((\theta_n, w_n)|(\theta', w'))}{p(\theta_n)P((\theta', w')|(\theta_n, w_n))}, 1\right\}\end{aligned}$$

Proof of Pseudo Marginal Algorithm

- Transition density is

$$\kappa(\theta', w', \theta_n, w_n) = \alpha R(\theta'|\theta_n)p(w') + (1 - \alpha^*)\delta_{\theta_n, w_n}(\theta', w'),$$

- By the same arguments as before the global balance equations are satisfied

$$L(\theta_n)w_n p(w_n|\theta_n)\kappa(\theta', w', \theta_n, w_n) = L(\theta')w' p(w'|\theta')\kappa(\theta_n, w_n, \theta', w')$$

- The target distribution is proportional to $L(\theta)p(\theta)wp(w|\theta)$.
- Integrating over w leaves a distribution proportional to $P(\theta|\mathbf{x})$.

Notes

- For an estimate of $L(\theta)$ with $\mathbb{E}[W|\theta] = c$ where $c > 0$ so that $\mathbb{E}[\hat{L}|\theta] = cL(\theta)$ this still works! (But it may mix poorly)
- We can use a bad estimate of $L(\theta)$ i.e. with just a single sample of $\mathbf{R}$. (But it may mix poorly)
- We *must* use the estimate $\hat{L}(\theta_n)$ from the previous iteration.

- This idea was first introduced by Mark Beaumont in Beaumont (2003), where he was using an approximate likelihood in the context of a statistical genetics example. This was later picked up by Christophe Andrieu and Gareth Roberts, who studied the technical properties of the approach in Andrieu and Roberts (2009).