Introduction to Statistical Modelling

T.J. McKinley (t.mckinley@exeter.ac.uk)

Statistical inference

Statistical inference can be thought of the ***inverse*** of simulation.

That is, we observe some data and want to know:

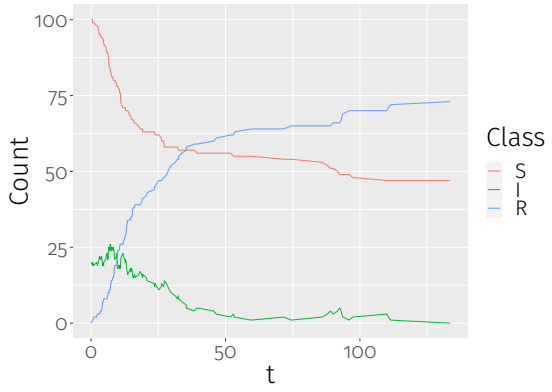
> *What **parameter values** for a model produce the 'best fit' to the data?*

We can use this to provide insights into key epidemiological processes (e.g. e.g. estimating the transmission rate, R_0 etc.). We can also use this to produce **predictions** and **forecasts**.

> *Key aspect is that we wish to quantify **uncertainty**.*

Example: SIR model

As an example, let's look at some data that we've simulated from a simple SIR model in a closed population of size $N = 120$, with the introduction of 20 initial infectives at time $t = 0$.



Example: SIR model

If we assume these data come from a **stochastic** SIR model of the form:

$$P(S_{t+\delta t} \to S_t - 1 \text{ and } I_{t+\delta t} \to I_t + 1) \approx \beta S_t I_t,$$
$$P(I_{t+\delta t} \to I_t - 1 \text{ and } R_{t+\delta t} \to R_t + 1) \approx \gamma I_t$$

for small δt. We can then ask the question:

> *"What values of β and γ produce epidemic curves that are the most consistent with the **observations**?"*

Likelihood functions

*"What values of β and γ produce epidemic curves that are the most consistent with the **observations**?"*

This question can be tackled by appealing to the **likelihood function**.

The *likelihood function*, $f(\mathbf{y} \mid \theta)$, gives the **likelihood**† of observing the data ($\mathbf{y}$) ***given*** a set of parameters (θ).

*The exact form of the **likelihood** function depends on the **specific model** and **data**.*

†if the data, $\mathbf{y}$, are **discrete**, then this is a **probability**

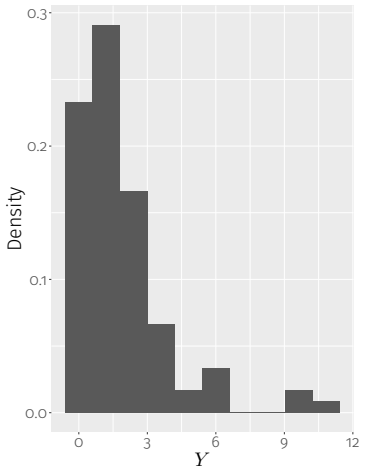
Likelihood functions

The exact form of the **likelihood** function depends on the **specific model** and **data**.

For example, imagine we have $n = 100$ **independent** samples from an **exponential** distribution:

$$Y_i \sim \text{Exp}(\lambda)$$

where λ is **unknown**.



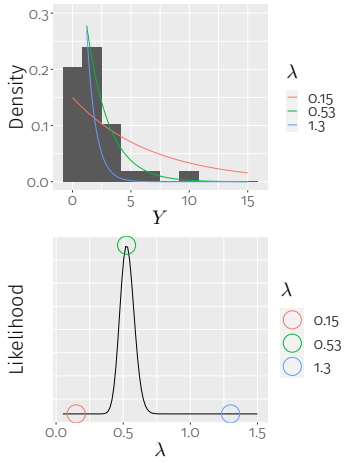
Likelihood functions

The exact form of the **likelihood** function depends on the **specific model** and **data**.

If the data are **independent**, then

$$f(\mathbf{y} \mid \lambda) = \prod_{i=1}^{n} f(y_i \mid \lambda)$$
$$= \prod_{i=1}^{n} \lambda e^{-\lambda y_i}$$

which is a function of λ and is dependent on the **probability density function** for each **observation** y_i.

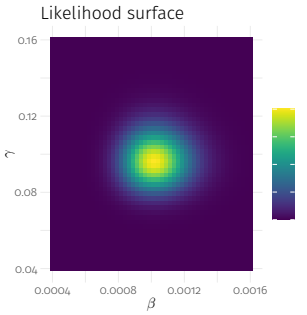


Likelihood functions

The **likelihood function** can be thought of as a **function** of the **unknown parameters** θ.

In the case of our SIR model, we have $\theta = (\beta, \gamma)$.

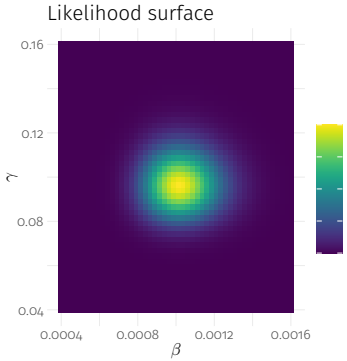
The **likelihood surface** (for different values of β and γ) looks like the plot opposite.



Note that in general **likelihoods** for compartmental models like this are **intractable**†, but in this simulated setting we can write it down directly.

†since **data** points are generally **not independent**, and typically the likelihood also depends on **unobserved variables**—we will return to this later

Likelihood functions



We can see that parameter values in the **yellow** region, produce **higher** likelihood values than parameter values in the **dark blue** regions.

*This means that parameters in the **yellow** region would produce simulations that are **more consistent** with the observed data than parameters in the **dark blue** regions.*

Maximum likelihood

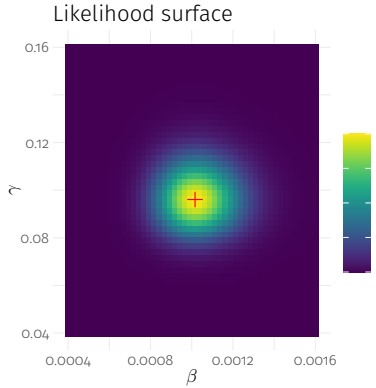
A natural way to estimate the parameters is to ask:

*What parameter values **maximise** the likelihood function[†]?*

Here the **maximum likelihood** estimates are shown with a **red cross**, and are given by:

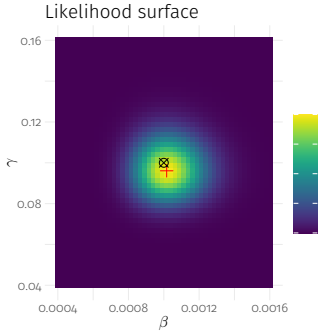
$$\hat{\beta} = 0.00102 \text{ and } \hat{\gamma} = 0.0961,$$

to 3 *significant figures*.



[†]we will see an alternative approach—using the **Bayesian** framework—later

Likelihood functions



- The **absolute value** of the likelihood is rarely interpretable, only **relative** values.
- The likelihood is based on the **data** and the choice of **model**, and thus will change for different data sets and different models.
- ML estimates do not guarantee a **good fit**.
- Similar parameter values can give similar fits (**uncertainty**).

Confidence intervals

Uncertainties in the parameter estimates can be quantified using **confidence intervals**. **Wider** confidence intervals signal **larger** uncertainties.

Here 95% confidence intervals[†] are:

- β: (0.000743, 0.00129)
- γ: (0.074, 0.118)

***Note**: these do **not** correspond to a 95% probability that the true value is between the limits. Rather, it means that if the experiment were to be conducted an **infinite** number of times, 95% of the time the calculated CI would contain the true value[‡].*

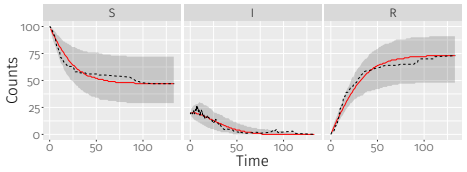
[†]based on a **large sample** approximation

[‡]this is so-called **frequentist** inference, as opposed to **Bayesian** inference that we will cover shortly

Model checking and prediction

We can check the model fit using the ML estimates to seed a large number of simulations from the model, and plot these against the observed data.

Median curves (red) plus 95% prediction intervals (grey) against observed data (dashed)



Here the model produces simulations that are consistent with the data[†].

Note that the uncertainty bounds here **do not** account for the **parameter uncertainty**[‡]; to calculate a **true prediction interval** for these types of model is harder (see Gelman and Hill (2007) for *simulation-based* approaches).

[†]be careful, simulations from stochastic models can be tricky—see McKinley, Cook, and Deardon (2009)

[‡]the parameters are **fixed** at the MLEs

Practical

In the first practical we will explore fitting the **catalytic model** for endemic diseases to serology data for ***rubella***.

To do this, we will need to write down a **likelihood function**, and then use one of R's in-build **optimisation** functions (`optim()`) to maximise with respect to the parameters to find the **maximum likelihood estimates**.

References i

Gelman, Andrew, and Jennifer Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

McKinley, Trevelyan J., Alex R. Cook, and Robert Deardon. 2009. "Inference in Epidemic Models Without Likelihoods." *The International Journal of Biostatistics* 5 (1). https://doi.org/10.2202/1557-4679.1171.

Introduction to Bayesian Modelling

T.J. McKinley (t.mckinley@exeter.ac.uk)

Bayesian inference

In the classical (**frequentist**) framework, the parameters of the model are considered **fixed**.

The **Bayesian** framework acknowledges that there is uncertainty in our **knowledge** of the parameters, and models this explicitly by treating the parameters as random variables (i.e. they follow a **probability distribution**)[†].

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

[†]Bayes (1763)

Bayesian inference

In the classical (**frequentist**) framework, the parameters of the model are considered **fixed**.

The **Bayesian** framework acknowledges that there is uncertainty in our **knowledge** of the parameters, and models this explicitly by treating the parameters as random variables (i.e. they follow a **probability distribution**).

We are interested in estimating the **posterior distribution** of the **parameters** (θ), *given* the **data** ($\mathbf{y}$).

$$f(\theta \mid \mathbf{y}) = \frac{f(\mathbf{y} \mid \theta)f(\theta)}{f(\mathbf{y})}$$

Bayesian Inference

$$f(\theta \mid \mathbf{y}) = \frac{f(\mathbf{y} \mid \theta)f(\theta)}{f(\mathbf{y})}$$

Here the **parameters** are denoted θ and the **data** are denoted $\mathbf{y}$.

- $f(\theta \mid \mathbf{y})$ is the **posterior** distribution for the parameters *given* the data.
- $f(\mathbf{y} \mid \theta)$ is the **likelihood** function (the distribution of the data *given* the parameters).
- $f(\theta)$ is the **prior** distribution (representing our belief in the values of the parameters in the *absence* of data).
- $f(\mathbf{y})$ is the **marginal likelihood**.

Bayesian Inference

The **marginal likelihood** is defined as:

$$f(\mathbf{y}) = \int_{\theta} f(\mathbf{y} \mid \theta) f(\theta) d\theta,$$

where the integral is across as many dimensions as θ.

This integral is often **analytically intractable**. However, importantly, $f(\mathbf{y})$ is **constant** (it is a function of the **data**, and the data are **observed**). Hence we often write:

$$f(\theta \mid \mathbf{y}) \propto f(\mathbf{y} \mid \theta) f(\theta)$$

Example: SIR model

The posterior distribution for our SIR model is:

$$f(\beta, \gamma \mid \mathbf{y}) \propto f(\mathbf{y} \mid \beta, \gamma) f(\beta, \gamma),$$

where $\mathbf{y}$ corresponds to all event **times** and **types** (e.g. infection, removal etc.)[†].

Assuming that we can calculate the **likelihood function**, $f(\mathbf{y} \mid \beta, \gamma)$, then we need to specify a **prior** distribution, $f(\beta, \gamma)$, for the parameters.

[†]note that in practice we rarely (if ever) have these data, and so usually the likelihood is **intractable**; however, since we have simulated the data we can actually calculate this here

Prior distributions

We know that:

- both β and γ must be **positive**;
- both β and γ are **continuous**.

Let's assume that we also know (from previous studies) that $\beta < 0.003$ and $\gamma < 0.3$[†]. We will also assume that β and γ are *a priori* **independent**, meaning that we can write:

$$f(\beta, \gamma) = f(\beta) f(\gamma).$$

If we then assume that all values of β are equally likely in $(0, 0.003)$, and similarly for γ in $(0, 0.3)$, then we can set:

$$\beta \sim U(0, 0.003) \quad \text{and} \quad \gamma \sim U(0, 0.3).$$

[†]for the sake of exposition

Prior distributions

Therefore if $\beta \sim U(0, 0.003)$ then

$$f(\beta) = \begin{cases} \frac{1}{0.003} & \text{for } 0 < \beta < 0.003, \\ 0 & \text{otherwise.} \end{cases}$$

Also, if $\gamma \sim U(0, 0.3)$ then

$$f(\gamma) = \begin{cases} \frac{1}{0.3} & \text{for } 0 < \gamma < 0.3, \\ 0 & \text{otherwise.} \end{cases}$$

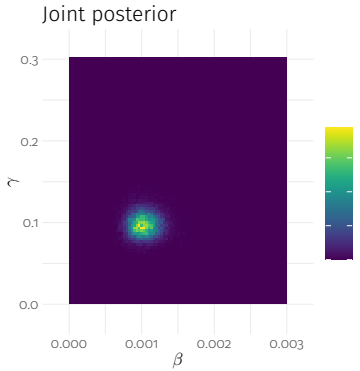
Hence $f(\beta, \gamma) = f(\beta) f(\gamma) = \frac{1}{0.003} \times \frac{1}{0.3} = 1111.11$, and thus

$$f(\beta, \gamma) = \begin{cases} 1111.11 & \text{for } 0 < \beta < 0.003 \text{ and } 0 < \gamma < 0.3, \\ 0 & \text{otherwise.} \end{cases}$$

Posterior distribution

$$f(\beta, \gamma \mid \mathbf{y}) \propto f(\mathbf{y} \mid \beta, \gamma) f(\beta, \gamma)$$

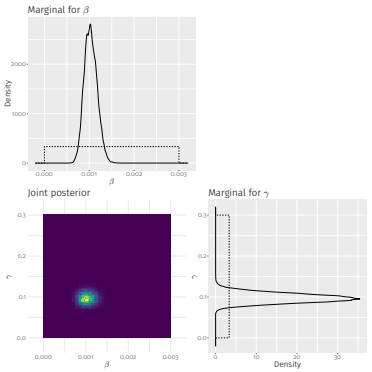
Even if the normalising constant $f(\mathbf{y})$ is **intractable**, it is nonetheless often possible to estimate the posterior distribution **numerically**†



†here I've used **Markov chain Monte Carlo**—more on that later

Joint and marginal distributions

It is often of interest to extract the **marginal distributions** for each parameter also.



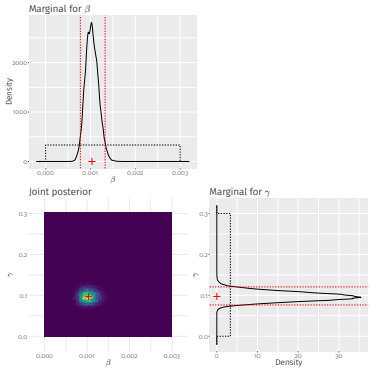
$$f(\beta \mid \mathbf{y}) = \int_{\gamma} f(\beta, \gamma \mid \mathbf{y}) d\gamma$$

$$f(\gamma \mid \mathbf{y}) = \int_{\beta} f(\beta, \gamma \mid \mathbf{y}) d\beta$$

These are often challenging to obtain *analytically*, but can be obtained ***numerically***.

Posterior distributions

Since the posterior is a **probability distribution** with respect to the **parameters**, we can produce estimates of **posterior means**, **standard deviations** and **credible intervals**†.



Parameter	Mean	SD	2.5%	97.5%
β	0.001	0.00014	0.00078	0.0013
γ	0.097	0.011	0.077	0.12

We can compare these to our ML estimates:

Parameter	Mean	2.5%	97.5%
β	0.001	0.00074	0.0013
γ	0.096	0.074	0.12

†**credible intervals** are actually probability statements about the parameters, as opposed to frequentist **confidence** intervals

Interpretation of the posterior

$$f(\theta \mid \mathbf{y}) \propto f(\mathbf{y} \mid \theta) f(\theta)$$

Hence we can view the posterior as reflecting the degree to which our prior beliefs change in the presence of data (captured through the likelihood).

Naturally, this leads to a discussion about the impact of the choice of prior distribution.

Prior distributions

Some statisticians argued that it is better to assume that the **data** (through the likelihood function) contains all necessary information regarding the parameters, and that the inclusion of a subjective choice of prior distribution will bias the results[†].

Bayesians on the other hand, argue that often we do have information about the parameters *a priori*, and it makes sense to build this into the estimates.

Generally, if there is a lot of information in the **data** relative to the **prior**, then the posterior will reflect the likelihood more than the prior, and vice-versa.

[†]it is worth remembering that formulating the likelihood also involves subjective decisions

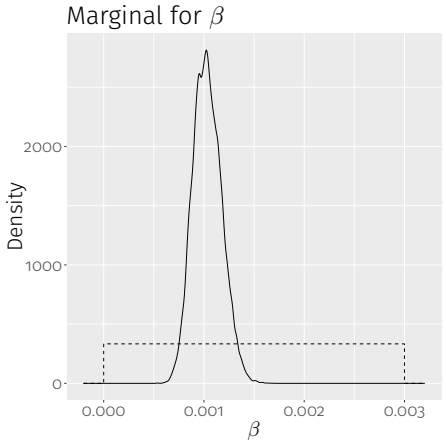
Prior distributions

To this end, if we have negligible prior information, we can choose some **vague** distribution (i.e. with **large variance**).

Care must be taken when specifying prior distributions, since poor choices can unduly influence the posterior.

Vague priors are usually safer, but can be more difficult to evaluate.

If the impact of the prior on the posterior is strong, then some form of sensitivity analysis to the choice of prior might also be advantageous.



Why Bayesian methods?

Why should we care about Bayesian methods? I will avoid a full-on philosophical debate here, but will put forward three key arguments.

1. Firstly, I would argue that the Bayesian treatment of probability is more natural than the classical treatment, and this extends to predictive distributions.

   e.g. (Rubin 1984): "...consumers of statistical answers... almost uniformly interpret them Bayesianly [sic]."

2. Secondly, the use of prior distributions, and by extension **hierarchical structures** (used extensively for e.g. missing data problems, spatio-temporal modelling etc.), enable analysis of models that are difficult to fit using classical methods.

Why Bayesian methods?

3. The use of prior distributions can also be very useful for inference when some parameters are **unidentifiable** (i.e. the likelihood contains negligible information about the parameter).

   In this case we cannot make inference about the unidentifiable parameter, but we can make inference about the other parameters conditional on the prior for the unidentifiable parameter.

4. It is straightforward to propagate uncertainties through to predictions to produce consistent **prediction intervals** that account for *parameter* and *stochastic* uncertainty.

Frequentist vs. Bayesian statistics

We will not enter into a philosophical debate here, rather we focus on a pragmatic comparison of the two schools of thought:

Frequentist	**Bayesian**
Parameters treated as **fixed** quantities (i.e. the "true" value exists but is unknown, and we wish to estimate it).	Uses **probability distributions** to explicitly model uncertainty around **knowledge** of the parameters.
The frequentist interpretation of probability corresponds to the uncertainty in an outcome of an experiment if the experiment were to be repeated *ad nauseum*.	The Bayesian interpretation of probability corresponds to how a degree of belief in a proposition changes due to available evidence.
Confidence intervals	**Credible** intervals
	Use of **prior** information

References i

Bayes, Thomas. 1763. "An Essay Towards Solving a Problem in the Doctrine of Chances." *Philosophical Transactions of the Royal Society* 53 (0): 370–418.

Gelman, Andrew, John B. Carlin, Hal S. Stern, and Donald B. Rubin. 2004. *Bayesian Data Analysis*. 2nd ed. Chapman; Hall/CRC.

Rubin, Donald B. 1984. "Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician." *The Annals of Statistics* 12: 1151–72.

Markov chain Monte Carlo

T.J. McKinley (t.mckinley@exeter.ac.uk)

Fitting Bayesian models

Now we return to the problem of how to fit Bayesian models. Recall that we only know the posterior distribution up to a normalising constant:

$$f(\theta \mid \mathbf{y}) \propto f(\mathbf{y} \mid \theta) f(\theta)$$

This means that we cannot write down an analytical form for the **posterior** distribution, so instead we need to appeal to **numerical** methods.

Fitting Bayesian models

It turns out that we can generate **empirical estimates** of probability distributions by **random sampling**.

That is, we take *large numbers of random samples* from a distribution, and then use these to estimate key aspects of the distribution, such as the mean, variance, quantiles, shape etc.

The general concept of estimating distributions (and key summary measures) through random sampling is known generally as **Monte Carlo** estimation[†].

[†]though Monte Carlo methods can be applied to a wider variety of problems than simply those explored here—for example Monte Carlo integration

Example: expected values

Consider a **random variable** X, that comes from a probability distribution with **probability density function** $f(x)$[†].

The **expected value** of X is defined as:

$$E(X) = \int_{\mathcal{X}} x f(x) dx,$$

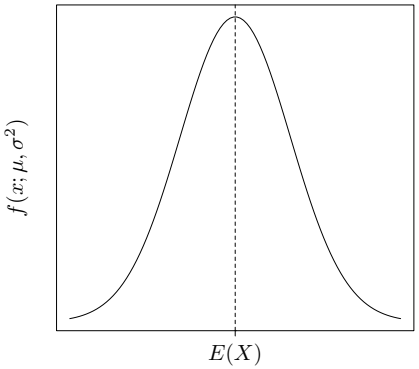
where the integral is over all possible values of X (denoted $\mathcal{X}$).

The **expected value**, $E(X)$, is the formal definition of the **mean** of a probability distribution.

[†]analogous formulations exist for **discrete** X, with probability *density* functions replaced by a probability *mass* functions, and *integration* replaced with *summation*

Aside: expected values

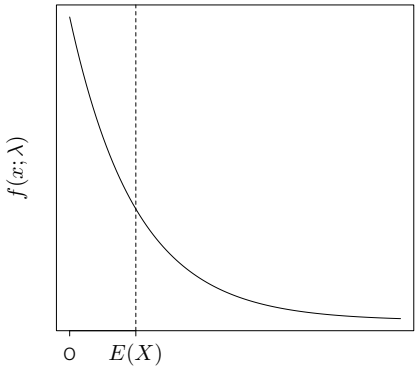
$$
\begin{aligned}
E(X) &= \int_x x f(x) dx \\
&= \int_{-\infty}^{\infty} x \cdot \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx \\
&= \mu.
\end{aligned}
$$



One can think about the expected value as being a **weighted average**, where each possible value of x is ***weighted*** by the **probability density function**, $f(x)$, evaluated at x.

Aside: expected values

$$
\begin{aligned}
E(X) &= \int_x x f(x) dx \\
&= \int_0^{\infty} x \cdot \lambda e^{-\lambda x} dx \\
&= \frac{1}{\lambda}.
\end{aligned}
$$



One can think about the expected value as being a **weighted average**, where each possible value of x is ***weighted*** by the **probability density function**, $f(x)$, evaluated at x.

Aside: expected values

Recall: the **sample mean**:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i,$$

where the x_i are considered **random samples** from a distribution with a **probability density function** $f(x)$[†].

It can be shown that as $n \to \infty$, then $\bar{x} \to E(X)$.

> *The **sample mean** $\bar{x}$ is an **unbiased** and **consistent** estimator of $E(X)$.*

[†]in a slight abuse of notation, we will denote this as $x_i \sim f(x)$

Monte Carlo estimation

Therefore, as long as you can produce **random samples**, $x_i \sim f(x)$, from a distribution, then you can ***estimate*** the **mean** of that distribution ($\mu = E(X)$) as:

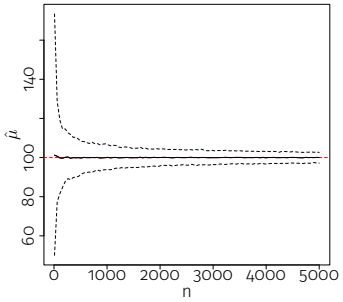
$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i,$$

regardless of the distribution of X.

This is the essence of **Monte Carlo** estimation.

Example: exponential distribution

Assume that we can sample from an **exponential** distribution with mean $\mu = 100$. Take different numbers of samples from $n = 5, \ldots, 5000$ and calculate the sample mean each time. Repeat this 1,000 times for each n[†] and plot the mean and 95% quantiles:



Notice:

- always **unbiased**;
- small $n \rightarrow$ large variance;
- large $n \rightarrow$ small variance.

[†] a Monte Carlo estimate of a Monte Carlo estimate, if you like!

The Law of the Unconscious Statistician

In general, if X is a random variable defined on some range of values $\mathcal{X}$, with probability density function $f(x)$, and $g(x)$ is some function of X, then

$$E\left[g(X)\right] = \int_{\mathcal{X}} g(x) f(x) dx.$$

This is known as the **Law of the Unconscious Statistician**[†].

Therefore we can generate Monte Carlo estimates of **general expectations**:

$$\begin{aligned} E\left[g(X)\right] &= \int_{\mathcal{X}} g(x) f(x) dx \\ &\approx \frac{1}{n} \sum_{i=1}^{n} g(x_i), \end{aligned}$$

where $x_i \sim f(x_i)$.

[†] $E(X)$ is a special case where $g(X) = X$

Estimating posterior expectations

Let's think about calculating the **posterior mean** for a parameter θ.

$$f(\theta \mid \mathbf{y}) \propto f(\mathbf{y} \mid \theta) f(\theta).$$

The posterior mean is:

$$E(\theta) = \int_{\Theta} \theta f(\theta \mid \mathbf{y}) d\theta,$$

where Θ corresponds to all possible values of θ. Hence we can approximate $E(\theta)$ as:

$$E(\theta) \approx \frac{1}{n} \sum_{i=1}^{n} \theta_i,$$

where θ_i are samples from the **posterior**, $f(\theta \mid \mathbf{y})$—**as long as we can sample from the posterior**!

Estimating posterior expectations

So, we wish to estimate the **posterior mean** for a parameter θ

$$E(\theta) = \int_{\Theta} \theta f(\theta \mid \mathbf{y}) d\theta \approx \frac{1}{n} \sum_{i=1}^{n} \theta_i,$$

where $\theta_i \sim f(\theta \mid \mathbf{y})$.

The problem is that we only know the **posterior** up to some normalising constant i.e.

$$f(\theta \mid \mathbf{y}) = \frac{f(\mathbf{y} \mid \theta) f(\theta)}{f(\mathbf{y})},$$

where $f(\mathbf{y})$ is **unknown**. Thus simple methods of sampling, such as inverse transform sampling, do not work.

Instead, we will use a technique called **Markov chain Monte Carlo**.

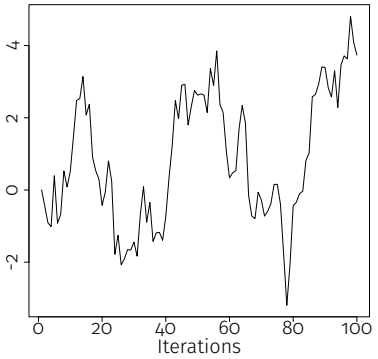
Markov chains

A Markov chain is simply a **sequence** of numbers, where each number in the sequence relates to the ***previous*** number[†]

For example, a simple random-walk:

$$x_i \sim N(x_{i-1}, 1)$$

for $i = 1, \dots, n$ and $x_0 = 1$ (say).



[†]we will only deal with **first-order** Markov chains here

MCMC

A simple **random-walk** like the one shown previously will just iterate around at random.

However, it is possible to generate Markov chains that **converge** to some distribution, known as the **stationary distribution**.

There are lots of variations of **MCMC**, but the one we will explore here is known as the **Metropolis-Hastings** algorithm[†] (Metropolis et al. 1953; Hastings 1970).

[†]technically the **random-walk Metropolis-Hastings** algorithm

Random-walk Metropolis-Hastings

Require: $\theta^{(0)}$.
  for $i = 1, \dots, n$ **do**
    Propose **candidate** $\theta' \sim q\left(\cdot \mid \theta^{(i-1)}\right)$.
    Calculate the **acceptance probability**:

$$\alpha = \min\left(1, \frac{f\left(\theta'\right) q\left(\theta^{(i-1)} \mid \theta'\right)}{f\left(\theta^{(i-1)}\right) q\left(\theta' \mid \theta^{(i-1)}\right)}\right)$$

    Sample $u \sim U(0,1)$
    if $u < \alpha$ **then**
      $\theta^{(i)} = \theta'$
    else
      $\theta^{(i)} = \theta^{(i-1)}$
    end if
  end for

$f(\cdot)$ is the **target** distribution.

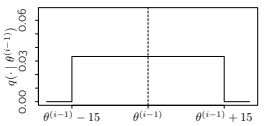
$q\left(\cdot \mid \theta^{(i-1)}\right)$ is a **proposal** distribution.

Example: normal distribution

Let's sample from a **normal distribution** with mean 10 and standard deviation 5, with

$$f(\theta) = \frac{1}{10\sqrt{\pi}} e^{-\frac{1}{2}\left(\frac{\theta - 10}{5}\right)^2}$$

Let's start at $\theta^{(0)} = 10$ and use a **uniform** proposal distribution $\theta' \sim U(\theta^{(i-1)} - 15, \theta^{(i-1)} + 15)$.



Symmetric proposals gives the **Metropolis** algorithm.

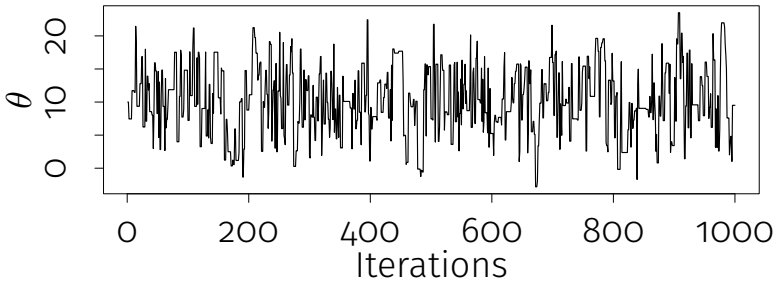
Require: $\theta^{(0)}$.
  for $i = 1, \dots, n$ **do**
    Propose **candidate** $\theta' \sim q\left(\cdot \mid \theta^{(i-1)}\right)$.
    Calculate the **acceptance probability**:

$$\alpha = \min\left(1, \frac{f\left(\theta'\right) \cancel{q\left(\theta^{(i-1)} \mid \theta'\right)}}{f\left(\theta^{(i-1)}\right) \cancel{q\left(\theta' \mid \theta^{(i-1)}\right)}}\right)$$

    Sample $u \sim U(0,1)$
    if $u < \alpha$ **then**
      $\theta^{(i)} = \theta'$
    else
      $\theta^{(i)} = \theta^{(i-1)}$
    end if
  end for

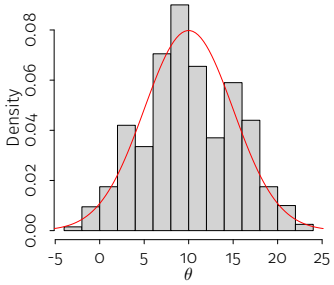
Example: normal distribution

Running this algorithm from $n = 1,000$ iterations gives the following chain:



Example: normal distribution

Plotting a **histogram** of these 1,000 samples provides an **empirical estimate** of the target distribution.



From these samples we can derive a **Monte Carlo** estimate of the mean:

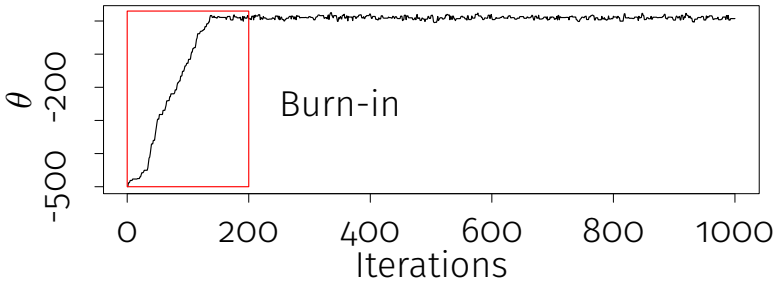
$$E(\theta) \approx \frac{1}{n}\sum_{i=1}^{1000}\theta_i = 10.15,$$

which can be compared to the true value $E(\theta) = 10$.

We can also derive **intervals**, **variances** etc.

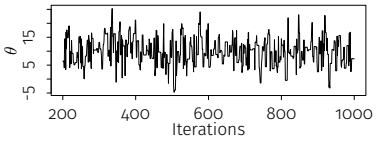
Example: normal distribution

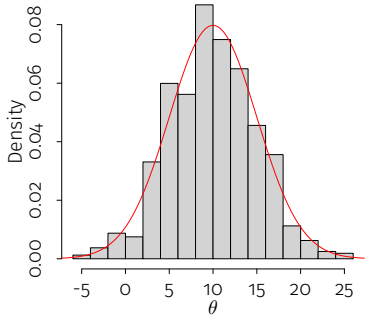
Remarkably, it doesn't matter where we start the chain from. Here $\theta^{(0)} = -500$:



Example: normal distribution

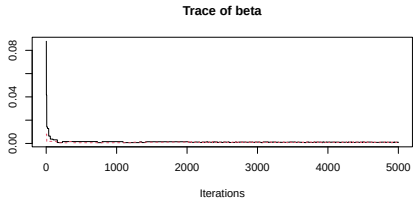
If we discard the **burn-in**, then we are left with samples from the **target** distribution as required.

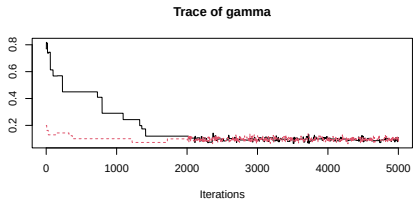




Convergence

In order for us to make robust inference, we need to be sure that the chain has **converged**. There is no consensus[†], but a common approach is to run **multiple** chains from different **initial values**:

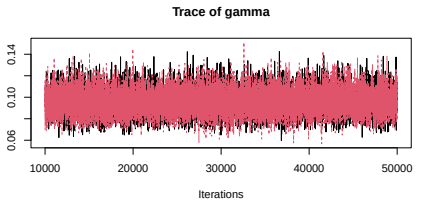




[†]or **guarantee**

Mixing

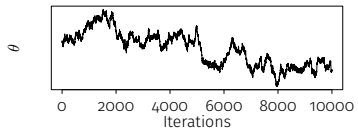
We also want to make sure that the chain is **exploring** the parameter space *efficiently*. This is known as **mixing**. After the **burn-in**, your **trace** plots should look like:



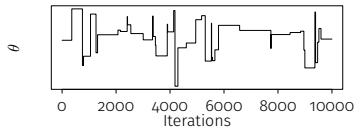
Improving efficiency

The **proposal distribution**, $q(\cdot)$, is key to efficient **mixing** and **convergence**.

In *random-walk* Metropolis-Hastings, the size of the **proposal variances** are important.



Here the proposal variance is **small** → **poor mixing**[q].



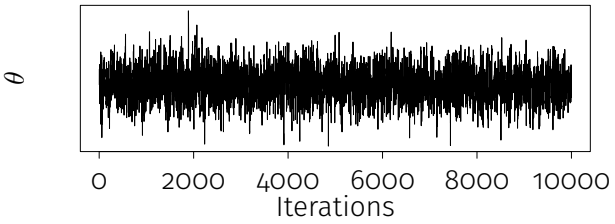
Here the proposal variance is **large** → **poor mixing**.

[q]high **autocorrelation**

Improving efficiency

Poor mixing means that the chain takes a long time to explore the posterior. This means you will have to run the chain for **much longer** to assess **convergence**.

Optimising MCMC is **hard**. In many problems we can use **adaptive proposal distributions** (e.g. Roberts and Rosenthal 2009), or alternative MCMC algorithms such as Hamiltonian Monte Carlo[†].



[†]such as used in the Stan software

Estimating posterior distributions

In Bayesian inference, the **target** distribution is the **posterior**:

$$f(\theta \mid \mathbf{y}) = \frac{f(\mathbf{y} \mid \theta) f(\theta)}{f(\mathbf{y})},$$

and the Metropolis-Hastings algorithm has **acceptance probability**:

$$\alpha = \min\left(1, \frac{\frac{f\left(\mathbf{y}|\theta'\right) f(\theta')}{\cancel{f(\mathbf{y})}}}{\frac{f\left(\mathbf{y}|\theta^{(i-1)}\right) f(\theta^{(i-1)})}{\cancel{f(\mathbf{y})}}} \times \frac{q\left(\theta^{(i-1)} \mid \theta'\right)}{q\left(\theta' \mid \theta^{(i-1)}\right)}\right)$$

Hence the **normalising constant** cancels in the ratio!

Posteriors for functions of random variables

One of the beautiful aspects of using MCMC for estimation, is that it is trivial to generate **posterior samples** for any function of the parameters[†].

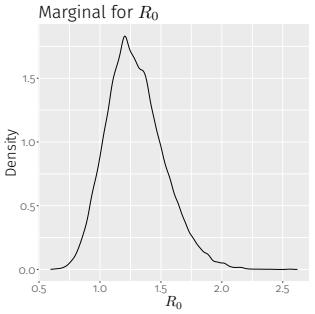
For example, given samples of $\beta^{(i)}$ and $\gamma^{(i)}$ $(i = 1, \ldots, N)$ from an SIR model, we can generate N samples of R_0 as:

$$R_0^{(i)} = \frac{\beta^{(i)} N}{\gamma^{(i)}}.$$

for $i = 1, \ldots, N$.

```
##         beta      gamma          ##        R0
##  0.001101749 0.10350498          ##  1.277329
##  0.001101749 0.10350498          ##  1.277329
##  0.001188174 0.08525523    →     ##  1.672401
##  0.001188174 0.08525523          ##  1.672401
##  0.001188174 0.08525523          ##  1.672401
##  0.001188174 0.08525523          ##  1.672401
```

Mean	SD	2.5%	97.5%
1.3	0.23	0.89	1.8



[†]from the Law of the Unconscious Statistician

Posterior predictive distributions

This idea carries through to **posterior predictive distributions**:

$$f\left(\mathbf{y}^* \mid \mathbf{y}\right) = \int_\Theta f\left(\mathbf{y}^* \mid \theta\right) f\left(\theta \mid \mathbf{y}\right) d\theta,$$

where $f\left(\mathbf{y}^* \mid \mathbf{y}, \theta\right)$ is the predictive distribution for a new observation $\mathbf{y}^*$ (defined by the **model**), given parameters θ; with $f\left(\theta \mid \mathbf{y}\right)$ the posterior density for θ.

Hence we **integrate** (or average) over the **posterior distribution**, and hence propagate uncertainty in the parameters directly through to the predictive density.

We can produce random samples from $f\left(\mathbf{y}^* \mid \mathbf{y}\right)$ by:

- taking N random samples from the **posterior** (via MCMC), $\theta^{(i)}$, then
- simulating $\left(\mathbf{y}^*\right)^{(i)} \sim f\left(\cdot \mid \theta^{(i)}\right)$.

These can be plotted / summarised in the usual way.

Software

There are various **general-purpose** Bayesian modelling packages, most notably:

- WinBUGS
- `jags`
- OpenBUGS
- `MCMCpack`
- `nimble`
- Stan / `rstan`

And various wrapper packages to make fitting simple (e.g. regression models much easier) e.g. the R packages `brms` and `rstanarm`.

Practical

In the first practical we will explore fitting the **catalytic model** for endemic diseases to serology data for ***rubella***.

We will use the R package `MCMCpack` to do this, since we can re-use the **likelihood function** we will derive for the MLEs.

For most use cases I would recommend `nimble` or Stan, which are under active development and very powerful. However, we don't have time to go over how to use these here (especially since the likelihood is non-standard).

References i

Gamerman, Dani, and Hedibert Freitas Lopes. 2006. *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. 2nd ed. CRC Press.

Gelman, Andrew, John B. Carlin, Hal S. Stern, and Donald B. Rubin. 2004. *Bayesian Data Analysis*. 2nd ed. Chapman; Hall/CRC.

Gilks, W. R., S. Richardson, and D. J. Spiegelhalter, eds. 1996. *Markov Chain Monte Carlo in Practice*. Chapman; Hall.

Hastings, W. K. 1970. "Monte Carlo Sampling Methods Using Markov Chains and Their Applications." *Biometrika* 57: 97–109.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. 1953. "Equations of State Calculations by Fast Computing Machine." *Journal of Chemical Physics* 21: 1087–91.

References ii

Roberts, Gareth O., and Jeffrey S. Rosenthal. 2009. "Examples of Adaptive MCMC." *Journal of Computational and Graphical Statistics* 18 (2): 349–67. https://doi.org/10.1198/jcgs.2009.06134.

(Particle) Markov chain Monte Carlo

T.J. McKinley (t.mckinley@exeter.ac.uk)

Inference and prediction

Once the model is fitted and the model fit assessed, we can use the model / parameter estimates in various ways:

- **Inference**: interpreting the parameter estimates.
- **Prediction**: to predict what might happen if the outbreak were to occur under the same conditions again.
- **Forecasting**: to predict what might happen in the future, based on data available now.

So what's the problem?

This is straightforward right?

The likelihood for compartmental models relies on having **exact** observations of event **times** and **types**. In practice events are rarely observed in detail:

- **Surveillance**: e.g. under-reporting, imperfect coverage, imperfect diagnosis, mis-diagnosis;
- **Rounding error**: e.g. data often collated daily / weekly;
- **Hidden states**: some epidemiological processes never observed (e.g. you might know *roughly* when you started feeling sick with flu, but not when you were infected or when you became infectious).

Dealing with these challenges is **hard**! (But we will have a go!)

Intractable likelihoods

To deal with the **partially observed** data, we can introduce a set of **latent** variables, $\mathbf{x} = (\mathbf{t}, \delta)$, where $\mathbf{t}$ is a vector of **hidden** event *times*, and δ is a vector of **hidden** event *types*.

Then the **likelihood** can be expressed as:

$$f(\mathbf{y} \mid \theta) = \int_{\mathcal{X}} f(\mathbf{y} \mid \mathbf{x}, \theta) f(\mathbf{x} \mid \theta) d\mathbf{x},$$

where

- $f(\mathbf{y} \mid \mathbf{x}, \theta)$ is an **observation** process (or **measurement error** / **model discrepancy**);
- $f(\mathbf{x} \mid \theta)$ is the **likelihood function** based on the **latent** variables $\mathbf{x}$.

Intractable likelihoods

$$f(\mathbf{y} \mid \theta) = \int_{\mathcal{X}} f(\mathbf{y} \mid \mathbf{x}, \theta) f(\mathbf{x} \mid \theta) d\mathbf{x},$$

This **marginalises** (*averages*) across the hidden variables $\mathbf{x}$.

This is a complex integral, over all possible combinations of events, and all possible event times consistent with the data.

It may also be the case that the **number** of hidden events is **unknown**, in which can we have to repeat the integration for every possible number of hidden events.

Data augmentation

One approach is therefore to include the **hidden** variables $\mathbf{x}$ as **additional parameters** in the model.

We can then estimate the ***joint* posterior** distribution for $(\theta, \mathbf{x})$, and then derive the ***marginals*** for the parameters of interest (θ) *numerically*.

This is usually done using MCMC methods; an approach known as **data-augmented MCMC** (e.g. Gibson and Renshaw 1998; Philip D. O'Neill and Roberts 1999; Jewell et al. 2009).

It is very powerful, but difficult to code, scale and optimise.

Simulation-based approaches

Alternatively, we can build inference algorithms around **simulating** directly from the model-of-interest, and then searching for parameter sets that are more consistent with the **observed data**.

These **simulation-based methods** are also powerful and flexible:

- Don't have to store all of the latent variables (so memory requirements are lower).
- Are often straightforward to parallelise.
- Simulation can often be easier than calculating the likelihood.
- Implementation often easier than DA (e.g. "plug-and-play")

However, there are also practical difficulties:

- The probability of matching the data exactly (i.e. getting a non-zero likelihood) is often very low.
- Often require some form of approximation to obtain a match.

Alternative fitting methods

Examples of latent variable methods:

- **Data-augmented MCMC** (e.g. Gibson and Renshaw 1998; Philip D. O'Neill and Roberts 1999; S. Cauchemez and Ferguson 2008; Jewell et al. 2009)
- **Sequential Monte Carlo** (Simon Cauchemez et al. 2008)

Examples of simulation-based methods:

- **Maximum likelihood via iterated filtering** (Ionides, Bretó, and King 2006)
- **Approximate Bayesian Computation** (e.g. Toni et al. 2009; McKinley, Cook, and Deardon 2009; Conlan et al. 2012; Brooks Pollock, Roberts, and Keeling 2014)
- **Pseudo-marginal methods** (e.g. P. D. O'Neill et al. 2000; Beaumont 2003; Andrieu and Roberts 2009; McKinley et al. 2014)
- **Particle MCMC** (Andrieu, Doucet, and Holenstein 2010; Drovandi, Pettitt, and McCutchan 2016)
- **Synthetic likelihood** (Wood 2010)
- **History matching** (with **emulation**) (e.g. Andrianakis et al. 2015; McKinley et al. 2018)

Pseudo-marginal MCMC

Require: $\theta^{(0)}$.
  for $i = 1, \ldots, n$ **do**
    Propose **candidate** $\theta' \sim q\left(\cdot \mid \theta^{(i-1)}\right)$.
    Calculate the **acceptance probability**:

$$\alpha = \min\left(1, \frac{\hat{f}\left(\mathbf{y} \mid \theta'\right) f\left(\theta'\right)}{\hat{f}\left(\mathbf{y} \mid \theta^{(i-1)}\right) f\left(\theta^{(i-1)}\right)} \times \frac{q\left(\theta^{(i-1)} \mid \theta'\right)}{q\left(\theta' \mid \theta^{(i-1)}\right)}\right)$$

    Sample $u \sim U(0, 1)$
    if $u < \alpha$ **then**
      $\theta^{(i)} = \theta'$
    else
      $\theta^{(i)} = \theta^{(i-1)}$
    end if
  end for

One option is to simply plug this **estimate** into a standard Metropolis-Hastings algorithm in place of the true likelihood.

Remarkably, as long as this estimate is **unbiased**, this will still converge to the **true** posterior.

This approach is known as **pseudo-marginal MCMC**.

Beaumont (2003); Andrieu and Roberts (2009).

Simulation-based approximations

One option is to replace the likelihood, $f(\mathbf{y} \mid \theta)$, by a **Monte Carlo** estimate:

$$\begin{aligned} f(\mathbf{y} \mid \theta) &= \int_{\mathcal{X}} f(\mathbf{y} \mid \mathbf{x}, \theta) f(\mathbf{x} \mid \theta) d\mathbf{x} \\ &\approx \frac{1}{M} \sum_{i=1}^{M} f(\mathbf{y} \mid \mathbf{x}_i, \theta), \end{aligned}$$

where $\mathbf{x}_i \sim f(\mathbf{x} \mid \theta)$ are simulations from the underlying model.

This provides an **unbiased** estimate for $f(\mathbf{y} \mid \theta)$.

Efficiency of pseudo-marginal MCMC

The efficiency (i.e. **mixing**) of pseudo-marginal MCMC relies on the **variance** of the **estimator** $\hat{f}(\mathbf{y} \mid \theta)$.

- If the variance is **small**, then mixing will be **improved**.
- If the variance is **large**, then mixing will be **poor**.

We can reduce the variance by:

- increasing the number of simulations $M \rightarrow$ higher computational burden;
- improving the estimator.

Particle MCMC

This leads on to the idea of **particle MCMC** (Andrieu, Doucet, and Holenstein 2010).

In essence this aims to use **Sequential Monte Carlo**[†] to produce an **unbiased** estimate of the likelihood that has **lower variance** than a vanilla Monte Carlo estimate.

One of the earliest and most widely used particle filters is known as the **bootstrap particle filter** (Gordon, Salmond, and Smith 1993).

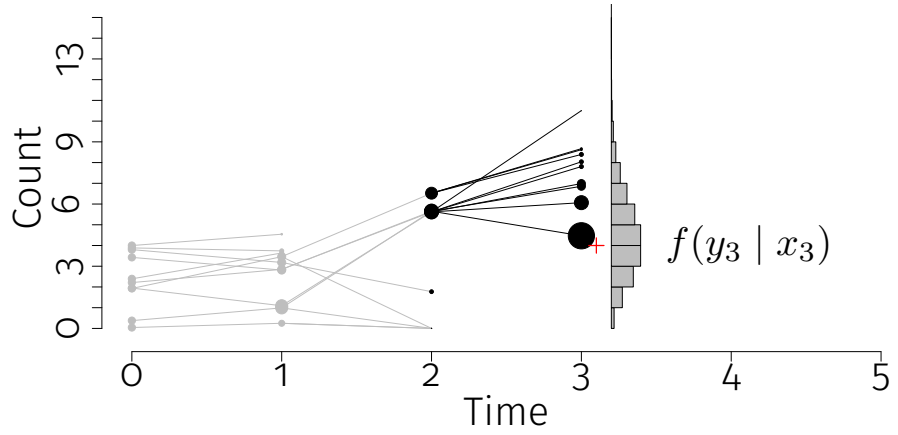
[†] i.e. **particle filtering**

Bootstrap particle filter

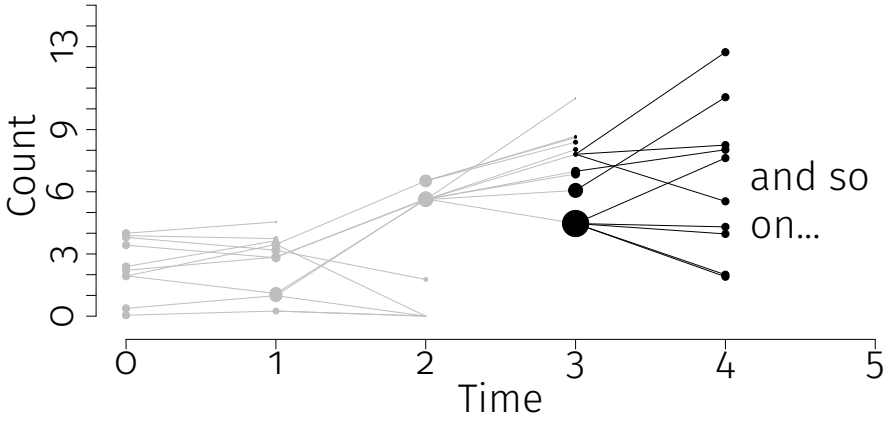
Each **particle** now corresponds to the **unobserved states** of the system at time 0, $\mathbf{x}_0 = (\mathbf{x}_0^1, \dots, \mathbf{x}_0^M)$. The parameters are **fixed**.

1. Each particle m is propagated forwards in time by **simulating** from the model $\mathbf{x}_1^m \sim f(\mathbf{x} \mid \mathbf{x}_0^m, \theta)$.
2. Each new particle is **weighted** according to the **observation process**, $f(\mathbf{y} \mid \mathbf{x}_1^m, \theta)$.
3. These weights are **normalised**, and a **re-sampling** step undertaken.
4. The new set of particles are propagated forwards to time $t+1$ and so on...

Bootstrap particle filter



Bootstrap particle filter



Bootstrap particle filter

We can generate an **unbiased estimate** of the conditional densities:

$$\hat{f}\left(y_t \mid y_{0:(t-1)}\right) = \frac{1}{M} \sum_{m=1}^{M} f\left(y_t \mid \mathbf{x}_t^m, \theta\right),$$

where $y_{0:(t-1)}$ corresponds to the observed time-series counts at time $t_0, t_1, \dots, t_{t-1}$.

It turns out that we can also derive an **unbiased** estimate of the overall **likelihood** as:

$$\hat{f}\left(\mathbf{y} \mid \theta\right) = f\left(y_0\right) \prod_{t=1}^{T} \hat{f}\left(y_t \mid y_{0:(t-1)}\right).$$

Particle MCMC

Hence we can generate an **unbiased** estimate of the likelihood which **numerically** integrates over the **hidden states**.

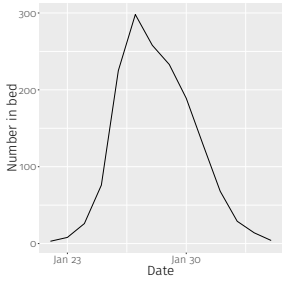
We can then plug this estimate into a standard Metropolis-Hastings algorithm to produce a **pseudo-marginal** MCMC routine that will converge to the *correct posterior distribution in probability*.

This approach only requires a **simulation** model, and an **observation process**.

The bootstrap particle filter we've used is defined for **time-series** counts, and can be extended in various ways.

Example: flu in boarding school

To illustrate some of these ideas we can use a case study of influenza in a boarding school. These data are from a paper in the BMJ in 1978 (Anonymous 1978) and provided in the `outbreaks` package. We use a simple $SIRR_1$ model:



The event probabilities are:

$$P\left[S_{t+\delta t} = S_t - 1, I_{t+\delta t} = I_t + 1\right] \approx \beta SI/N$$
$$P\left[I_{t+\delta t} = I_t - 1, R_{t+\delta t} = R_t + 1\right] \approx \gamma I$$
$$P\left[R_{t+\delta t} = R_t - 1, R_{1,t+\delta t} = R_{1,t} + 1\right] \approx \gamma_1 R$$

Example: flu in boarding school

Here we will place a Poisson error process around the R curve, such that:

$$R_t \sim \text{Po}(R'_t + 10^{-6}),$$

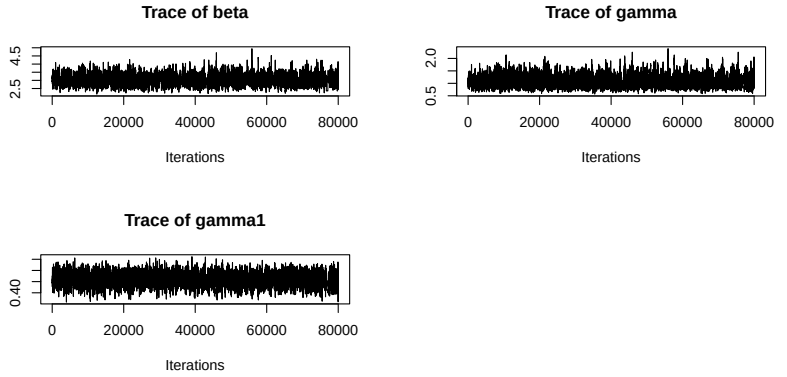
where R_t is the **observed** R count at time t, R'_t is the simulated count[†].

The initial population size is 763 pupils, and we assume an initial introduction of infection of a single child at day 0.

[†]see e.g. Funk et al. (2016) or here for similar ideas in practice

Example: flu in boarding school

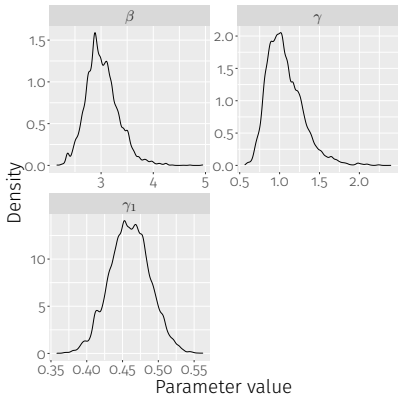
We ran a PMCMC algorithm for 100,000 iterations, discarding the first 20,000 as burn-in. We used 75 particles for the particle filter.



Example: flu in boarding school

Summaries of the marginal posterior distributions are:

Parameter	Mean	2.5%	97.5%
β	3	2.5	3.7
γ	1.1	0.73	1.6
γ_1	0.46	0.41	0.52



Summary

Particle MCMC is a powerful approach for inference in **partially observed** systems (see e.g Wilkinson 2012 or his associated blog for fantastic explanations of these methods).

It is often used when there is some form of **stochastic** discrepancy / observation process mapping the **hidden** states to the **observed** states.

Other particle filters exist, such as the **Alive Particle Filter** (Jasra et al. 2013), and the system can be extended to the **ABC** setting, where approximate matching around data points is used (Drovandi, Pettitt, and Lee 2014; McKinley et al. 2020).

Software

Partially Observed Markov Processes:

- `pomp`
- `SimBIID`[†]
- `SimInf`[‡]
- `nimble`[§]
- `hmer`[¶]

[†]designed mostly for teaching purposes, but should work for simple models
[‡]now implements ABC-SMC (e.g. Toni et al. 2009; McKinley, Cook, and Deardon 2009)
[§]now supports state-space models (although I've not used it for these)
[¶]hot-off-the-press! Implements emulation and history matching for epidemic models

References i

Andrianakis, Ioannis, Ian Vernon, Nicky McCreesh, Trevelyan J. McKinley, Jeremy E. Oakley, Rebecca N. Nsubuga, Michael Goldstein, and Richard G. White. 2015. "Bayesian History Matching of Complex Infectious Disease Models Using Emulation: A Tutorial and a Case Study on HIV in Uganda." *PLoS Computational Biology* 11 (1): e1003968.

Andrieu, Christophe, Arnaud Doucet, and Roman Holenstein. 2010. "Particle Markov Chain Monte Carlo Methods." *Journal of the Royal Statistical Society, Series B (Methodological)* 72 (3): 269–342.

Andrieu, Christophe, and Gareth O. Roberts. 2009. "The Pseudo-Marginal Approach for Efficient Monte Carlo Simulation." *The Annals of Statistics* 37 (2): 697–725.

References ii

Anonymous. 1978. "Influenza in a Boarding School." *British Medical Journal* 1: 578.

Beaumont, Mark A. 2003. "Estimation of Population Growth and Decline in Genetically Monitored Populations." *Genetics* 164: 1139–60.

Brooks Pollock, Ellen, Gareth O. Roberts, and Matt J. Keeling. 2014. "A Dynamic Model of Bovine Tuberculosis Spread and Control in Great Britain." *Nature* 511: 228–31. https://doi.org/10.1038/nature13529.

Cauchemez, S., and Neil M. Ferguson. 2008. "Likelihood-Based Estimation of Continuous-Time Epidemic Models from Time-Series Data: Application to Measles Transmission in London." *Journal of the Royal Society Interface* 5 (25): 885–97.

References iii

Cauchemez, Simon, Alain-Jacques Valleron, Pierre-Yves Boëlle, Antoine Flahault, and Neil M. Ferguson. 2008. "Estimating the Impact of School Closure on Influenza Transmission from Sentinel Data." *Nature* 452: 750–55. https://doi.org/10.1038/nature06732.

Conlan, Andrew J. K., Trevelyan J. McKinley, Katerina Karolemeas, Ellen Brooks Pollock, Anthony V. Goodchild, Andrew P. Mitchell, Colin P. D. Birch, Richard S. Clifton-Hadley, and James L. N. Wood. 2012. "Estimating the Hidden Burden of Bovine Tuberculosis in Great Britain." *PLoS Computational Biology* 8 (10): e1002730.

Drovandi, Christopher C., Anthony N. Pettitt, and Anthony Lee. 2014. "Bayesian Indirect Inference Using a Parametric Auxiliary Model." *Statistical Science* 30 (1): 72–95.

References iv

Drovandi, Christopher C., Anthony N. Pettitt, and Roy A. McCutchan. 2016. "Exact and Approximate Bayesian Inference for Low Integer-Valued Time Series Models with Intractable Likelihoods." *Bayesian Analysis* 11 (2): 325–52.

Funk, Sebastian, Adam J. Kucharski, Anton Camacho, Rosalind M. Eggo, Laith Yakob & Lawrence M. Murray, and W. John Edmunds. 2016. "Comparative Analysis of Dengue and Zika Outbreaks Reveals Differences by Setting and Virus." *PLoS Neglected Tropical Diseases* 10 (12): e0005173.

Gibson, Gavin J., and Eric Renshaw. 1998. "Estimating Parameters in Stochastic Compartmental Models Using Markov Chain Methods." *IMA Journal of Mathematics Applied in Medicine and Biology* 15: 19–40.

References v

Gordon, N. J., D. J. Salmond, and A. F. M. Smith. 1993. "Novel Approach to Nonlinear/Non-Gaussian Bayesian State Estimation." *Radar and Signal Processing, IEE Proceedings F.* 140 (2): 107–13. https://doi.org/10.1049/ip-f-2.1993.0015.

Ionides, E. L., C. Bretó, and A. A. King. 2006. "Inference for Nonlinear Dynamical Systems." *Proceedings of the National Academy of Sciences USA* 103: 18438–43.

Jasra, Ajay, Anthony Lee, Christopher Yau, and Xiaole Zhang. 2013. "The Alive Particle Filter." https://arxiv.org/abs/1304.0151.

Jewell, Chris P., Theodore Kypraios, Peter Neal, and Gareth O. Roberts. 2009. "Bayesian Analysis for Emerging Infectious Diseases." *Bayesian Analysis* 4 (4): 465–96.

References vi

McKinley, Trevelyan J., Alex R. Cook, and Robert Deardon. 2009. "Inference in Epidemic Models Without Likelihoods." *The International Journal of Biostatistics* 5 (1). https://doi.org/10.2202/1557-4679.1171.

McKinley, Trevelyan J., Peter Neal, Simon E. F. Spencer, Andrew J. K. Conlan, and Laurence Tiley. 2020. "Efficient Bayesian Model Choice for Partially Observed Processes: With Application to an Experimental Transmission Study of an Infectious Disease." *Bayesian Analysis* 15 (3): 839–70. https://doi.org/10.1214/19-BA1174.

McKinley, Trevelyan J., Joshua V. Ross, Rob Deardon, and Alex R. Cook. 2014. "Simulation-Based Bayesian Inference for Epidemic Models." *Computational Statistics and Data Analysis* 71: 434–47.

References vii

McKinley, Trevelyan J., Ian Vernon, Ioannis Andrianakis, Nicky McCreesh, Jeremy E. Oakley, Rebecca N. Nsubuga, Michael Goldstein, and Richard G. White. 2018. "Approximate Bayesian Computation and Simulation-Based Inference for Complex Stochastic Epidemic Models." *Statistical Science* 33 (1): 4–18. https://doi.org/10.1214/17-STS618.

O'Neill, P. D., D. J. Balding, N. G. Becker, M. Eerola, and D. Mollison. 2000. "Analyses of Infectious Disease Data from Household Outbreaks by Markov Chain Monte Carlo Methods." *Applied Statistics* 49: 517–42.

O'Neill, Philip D., and Gareth O. Roberts. 1999. "Bayesian Inference for Partially Observed Stochastic Epidemics." *Journal of the Royal Statistical Society. Series A (General)* 162: 121–29.

References viii

Toni, Tina, David Welch, Natalja Strelkowa, Andreas Ipsen, and Michael P. H. Strumpf. 2009. "Approximate Bayesian Computation Scheme for Parameter Inference and Model Selection in Dynamical Systems." *Journal of the Royal Society Interface* 6: 187–202.

Wilkinson, Darren J. 2012. *Stochastic Modelling for Systems Biology*. 2nd ed. Chapman; Hall / CRC.

Wood, Simon N. 2010. "Statistical Inference for Noisy Nonlinear Ecological Dynamic Systems." *Nature* 466: 1102–4.