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.

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).

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).

Bayesian Inference

VACCINE IMPACT MODELLING CONSORTIUM

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

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

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

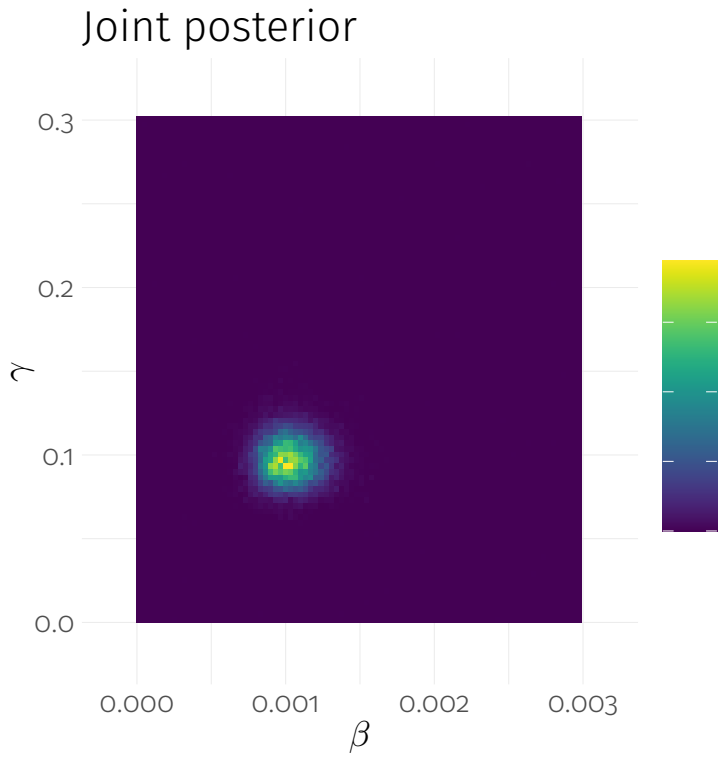
Hence $f(\beta, \gamma) = f(\beta)f(\gamma) = \frac{1}{0.003} \times \frac{1}{0.3} = 1111.11$, and thus

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

Posterior distribution

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

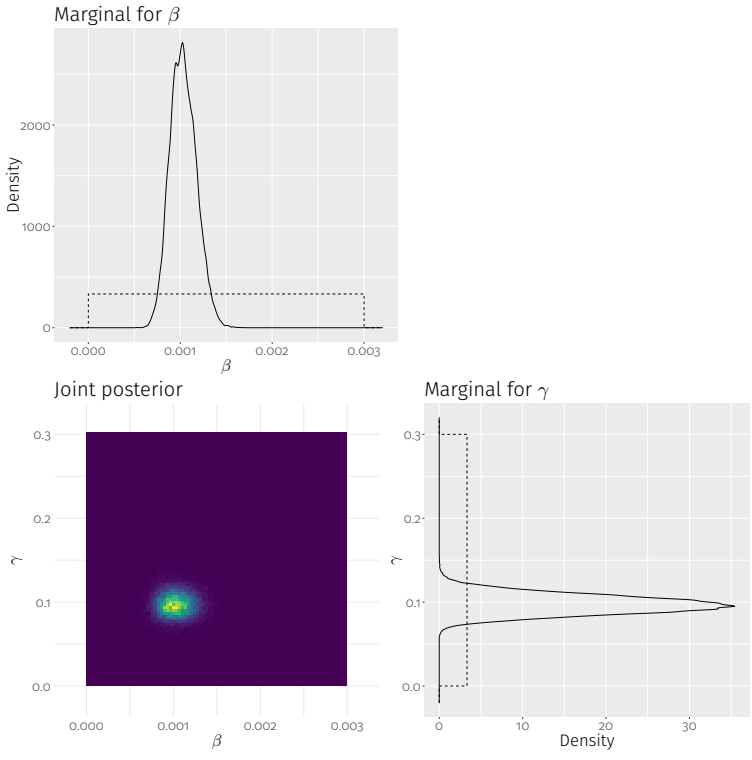
Even if the normalising constant $f(\mathbf{y})$ is **intractable**, it is nonetheless often possible to estimate the posterior distribution **numerically**[†]



[†]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.



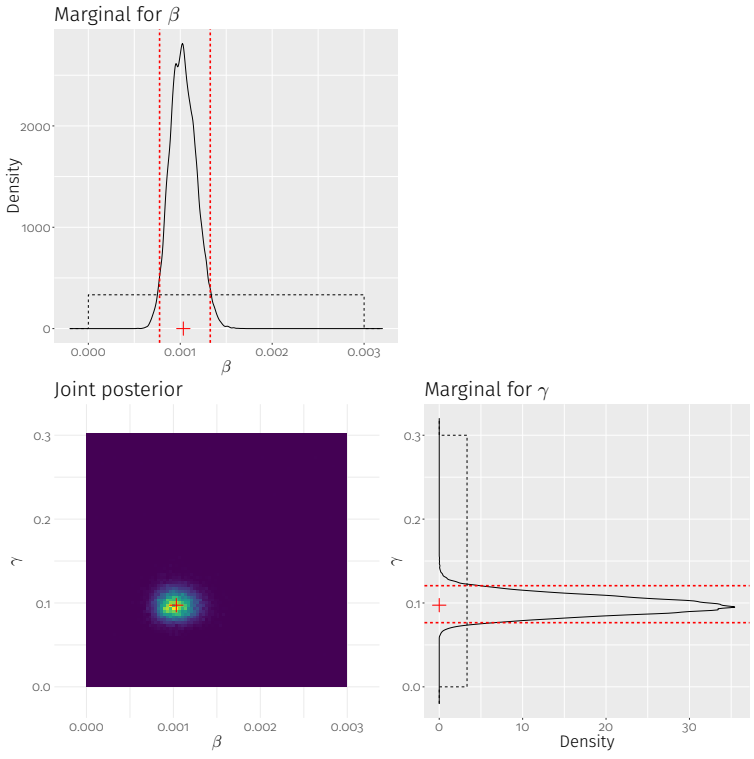
$$f(\beta \mid \mathbf{y}) = \int_{\gamma} f(\beta, \gamma \mid \mathbf{y}) d\gamma$$

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

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

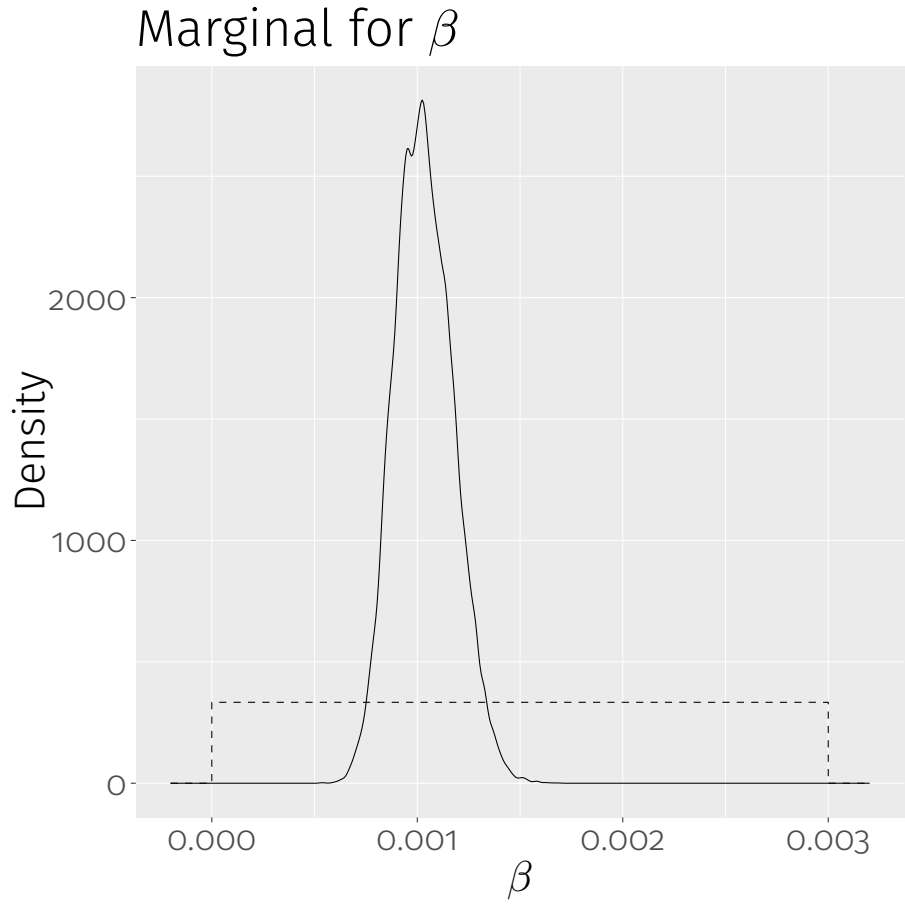
Prior distributions

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?

VACCINE IMPACT MODELLING CONSORTIUM

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?

VACCINE IMPACT MODELLING CONSORTIUM

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

VACCINE IMPACT MODELLING CONSORTIUM

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.