Session 2.2: Posterior Predictive Distribution and Monte Carlo computation

Bayesian modelling for Spatial and Spatio-temporal data, Imperial College

January-March 2023

Learning objectives

After this lecture you should be able to

- Explain what is the posterior predictive distribution

- Explain how it is computable

ADD MC objectives

Outline

1. Bayesian prediction
2. Computation of PPD
3. Example

Posterior Predictive Distribution

Bayesian prediction

- Often the objective of our analysis is to predict a future event

Bayesian prediction

- Often the objective of our analysis is to predict a future event
- Consider this example:

We estimate the prevalence of a disease in a UK hospital using a sample of n = 58 individuals.

We find that $y = 10$ individuals have the disease.

What is the probability that, if we additionally sample `(k = 30)` individuals this year, at least 5 will have the disease?

Bayesian prediction

- Often the objective of our analysis is to predict a future event
- Consider this example:

We estimate the prevalence of a disease in a UK hospital using a sample of n = 58 individuals.

We find that $y = 10$ individuals have the disease.

What is the probability that, if we additionally sample `(k = 30)` individuals this year, at least 5 will have the disease?

- As usually we start specifying the data distribution:

$$y \sim \text{Binomial}(\theta, n = 58)$$

- Let's θ be the true disease prevalence and y^* be the predicted value

Bayesian prediction

- Often the objective of our analysis is to predict a future event

- Consider this example:

We estimate the prevalence of a disease in a UK hospital using a sample of n = 58 individuals.

We find that $y = 10$ individuals have the disease.

What is the probability that, if we additionally sample `(k = 30)` individuals this year, at least 5 will have the disease?

- As usually we start specifying the data distribution:

$$y \sim \text{Binomial}(\theta, n = 58)$$

- Let's θ be the true disease prevalence and y^* be the predicted value
- If θ were known, then we would predict

$$y^*|\theta \sim \text{Binomial}(30, \theta)$$

thus $\text{P}(y \geq 5) = 1 - (\sum_{j=0}^{4} \theta^j (1-\theta)^{30-j})$

BUT `(\dots\theta)` is unknown

Source of variation in prediction:

MRC Centre for Environment & Health

Medical Research Council

Imperial College London

- We don't know the true value of the parameters and we specify a prior on it:

$$\theta \sim \mathrm{Beta}(a, b)$$

- There is sampling variability ($\rightarrow$ choice of the data distribution)

Source of variation in prediction:

- We don't know the true value of the parameters and we specify a prior on it:

$$\theta \sim \text{Beta}(a, b)$$

- There is sampling variability ($\rightarrow$ choice of the data distribution)

To account for the sources of variation we iterate the following steps:

1. Sample from the posterior distribution $\theta \sim p(\theta \mid y)$
2. Sample new values $y^* \sim p(y \mid \theta)$

- By repeating these steps a large number of times, we eventually obtain a reasonable approximation to the posterior predictive distribution.

Posterior Predictive Distribution (PPD)

- The PPD represents our uncertainty over the outcome of a future data collection, accounting for the observed data and model choice

- For the sake of prediction, the parameters are not of interest. They are vehicles by which the data inform about the predictive model

- The PPD averages over their posterior uncertainty

$$p(y^*|y) = \int p(y^*|\theta)p(\theta|y)d\theta$$

- This properly accounts for parametric uncertainty

- The input is data, the output is a prediction distribution

Computation

Computing the PPD

- Say $\theta^{(1)}, \ldots, \theta^{(M)}$ are samples from the posterior

- If we make a sample for y^* for each $\theta^{(m)}$,

$$y^{*(m)} \sim p(y|\theta^{(m)})$$

  then the $y^{*(m)}$ are samples from the PPD

- The posterior predictive mean is approximated by the sample mean of the $y^{*(m)}$

- The probability that $y^* \geq 5$ is approximated by the sample proportion of the $y^{*(m)}$ that are equal or above 5

Example

Example

- We estimate the prevalence of a disease in the UK population using a sample of $n = 58$ individuals.
- We find that $y = 10$ individuals have the diseases.
- What is the probability that, if we additionally sample `(k=30)` individuals this year, at least 5 will have the disease?

1. Likelihood: $y \sim \text{Binomial}(\theta, 58)$
2. Prior: $\theta \sim \text{Beta}(1, 1)$
3. Posterior: $\theta \mid y \sim \text{Beta}(10 + 1, 58 - 10 + 1)$
4. PPD: $y^* \sim \text{Binomial}(\theta \mid y, 30)$
5. $P(y \geq 5) = \sum_{j=5}^{30} P(y^* = j)$