Session 2.2: Posterior Predictive Distribution and Monte Carlo computation

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

January-March 2023





Imperial College London

After this lecture you should be able to

- Explain what is the posterior predictive distribution
- Explain how it is computable

ADD MC objectives



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



Imperial College London

Posterior Predictive Distribution

Medical Imperial College Research Council London

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



- 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:

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

ullet Let's heta 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:

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

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

$$y^* | heta \sim ext{Binomial}(30, heta)$$

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

BUT (\dots\theta) is unknown

edical Imperial College Isearch Duncil London

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

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

ullet There is sampling variability (o 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:

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

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

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

- Sample from the posterior distribution $heta \sim p(heta \mid y)$
- Sample new values $y^* \sim p(y \mid heta)$
- By repeating these steps a large number of times, we eventually obtain a reasonable approximation to the posterior predictive distribution.

Posterior Predictive Distribution (PPD)

MRC

- 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





- ullet Say $heta^{(1)},\ldots, heta^{(M)}$ are samples from the posterior
- If we make a sample for y^* for each $\theta^{(m)}$,

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

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

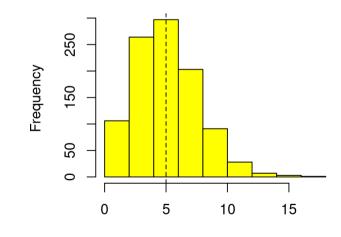
- The posterior predictive mean is approximated by the sample mean of the $y^{st(m)}$
- ullet The probability that $y^* \geq 5$ is approximated by the sample proportion of the $y^{*(m)}$ that are equal or above 5



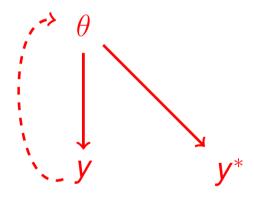
Example

Example

- ullet We estimate the prevalence of a disease in the UK population using a sample of n=58 individuals.
- ullet 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: $\theta \sim \mathrm{Beta}(1,1)$
 - 3 Posterior: $heta \mid y \sim \mathrm{Beta}(10+1,58-10+1)$
 - $oldsymbol{\theta}$ PPD: $y^* \sim \operatorname{Binomial}(\theta \mid y, 30)$
 - $P(y \ge 5) = \sum_{j=5}^{30} P(y^* = j)$







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