MAST90125: Bayesian Statistical Learning

Lecture 4: Posteriors and prediction in Bayesian inference

Prepared by Feng Liu and Guoqi Qian

Introduction

In the previous lecture, we discussed prior distributions. However once we have data to test a postulated model, do we want to make inference based only on the prior?

Introduction

In the previous lecture, we discussed prior distributions. However once we have data to test a postulated model, do we want to make inference based only on the prior?

The answer is no. The inferential statements we want to make should be based on the posterior distribution, that is a distribution conditioned on the data available. This lecture will discuss various aspects of posterior based inference.

Point estimation

- When we use maximum likelihood to determine quantities of inferential interest, we aim to obtain
  - A point estimate, ideally unbiased, ($E(\hat{\theta}|\theta) = \theta$).
  - A measure of estimate precision, such as Var($\hat{\theta}$).
  - An interval based measure of uncertainty in the parameter estimate, often relying on large-sample theory.
- In Bayesian methods, inference is based on the posterior distribution, $p(\theta|y)$.
  - While it would be straight-forward to construct an interval expressing a range of possible θ from $p(\theta|y)$, what would be an appropriate point estimate? Is there even an appropriate point estimate?

Point estimation

- What properties should a point estimate determined from the posterior distribution have?

Point estimation

- What properties should a point estimate determined from the posterior distribution have?
  - Does it still make sense to want an unbiased estimator?

Point estimation

- What properties should a point estimate determined from the posterior distribution have?
  - Does it still make sense to want an unbiased estimator?
    - Not necessarily. The posterior can be viewed as a weighted average of prior and likelihood, as seen before, so unbiasedness based on data alone conditional on θ is not of overriding importance.
  - What is a point estimate supposed to be a measure of?

Point estimation

- What properties should a point estimate determined from the posterior distribution have?
  - Does it still make sense to want an unbiased estimator?
    - Not necessarily. The posterior can be viewed as a weighted average of prior and likelihood, as seen before, so unbiasedness based on data alone conditional on θ is not of overriding importance.
  - What is a point estimate supposed to be a measure of?
    - The unbiased MLE point estimate $\hat{\theta}$ is a measure of central tendency, specifically mean.
- The information given in $p(\theta|y)$ would allow for a variety of measures of central tendency to be constructed such as

$$E(\theta|y) \quad \text{Median}(\theta|y) \quad \text{Mode}(\theta|y)$$

Credible intervals

- The posterior distribution readily allows for the construction of $100(1-\alpha)$ % credible intervals:

$$\Pr(\theta \in C|y) = 1-\alpha: \ \textit{cont.} \ \int_C p(\theta|y)d\theta, \ \textit{disc.} \ \sum_{\theta' \in C} \Pr(\theta'|y),$$

where C denotes the credible region.

  - Question 1: In this setting, what is θ?

Credible intervals

- The posterior distribution readily allows for the construction of $100(1-\alpha)$ % credible intervals:

$$\Pr(\theta \in C|y) = 1-\alpha: \ \textit{cont.} \ \int_C p(\theta|y)d\theta, \ \textit{disc.} \ \sum_{\theta' \in C} \Pr(\theta'|y),$$

where C denotes the credible region.

  - Question 1: In this setting, what is θ?
    - θ is a random variable. Unlike in frequentist statistics, parameters are random variables in Bayesian statistics.
  - Question 2: If we consider the single parameter case, does C need to be a single interval?

Credible intervals

- The posterior distribution readily allows for the construction of $100(1-\alpha)$ % credible intervals:

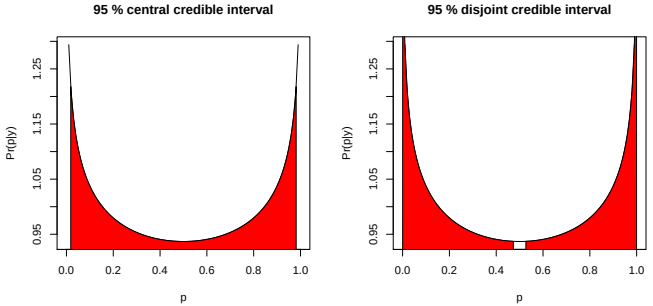
$$\Pr(\theta \in C|y) = 1-\alpha: \ \textit{cont.} \int_C p(\theta|y)d\theta, \ \textit{disc.} \sum_{\theta' \in C} \Pr(\theta'|y),$$

where C denotes the credible region.

  - Question 1: In this setting, what is θ?
    - θ is a random variable. Unlike in frequentist statistics, parameters are random variables in Bayesian statistics.
  - Question 2: If we consider the single parameter case, does C need to be a single interval?
    - No. In fact, there may be cases where to minimise the length of the credible interval it is desirable to report a union of disjoint intervals.

Example of Credible intervals

- Consider a bi-modal Beta distribution $\alpha = \beta = 0.9$.



- While the 95 % central credible interval for p is $(0.0185, 0.9815)$ with interval width 0.963, an 95 % disjoint credible interval $(0, 0.4733) \cup (0.5267, 1)$, has interval width $2 \times 0.4733 = 0.9466$.

Credible interval as Bayesian analogue to ?

- The credible interval should remind you of something very familiar from frequentist statistics. What is it?

Credible interval as Bayesian analogue to ?

- The credible interval should remind you of something very familiar from frequentist statistics. What is it?
  - The confidence interval, which is the mathematical statement,

$$\Pr(L(y) \leq \theta \cap U(y) \geq \theta | \theta) = 1 - \alpha$$

- The difference is in interpretation.
  - The probability used in constructing the confidence interval is conditional on the parameter θ. This means for any confidence interval you construct, whether θ is in the interval is a yes/no statement.

Credible interval as Bayesian analogue to ?

- The credible interval should remind you of something very familiar from frequentist statistics. What is it?
  - The confidence interval, which is the mathematical statement,

$$\Pr(L(y) \leq \theta \cap U(y) \geq \theta | \theta) = 1 - \alpha$$

- The difference is in interpretation.
  - The probability used in constructing the confidence interval is conditional on the parameter θ. This means for any confidence interval you construct, whether θ is in the interval is a yes/no statement.
  - The probability used in constructing the credible interval is conditional on the data y. This means the credible interval can be interpreted probabilistically, as in there is a $1 - \alpha$ probability that θ lies in the region C, based on the data available.

Types of credible interval

- As the beta posterior example mentioned earlier, there are multiple ways to define the credible region.
- One way would be to use the central credible region,

$$\Pr(q_{\alpha/2} \leq \theta \leq q_{1-\alpha/2}|y) = 1 - \alpha,$$

so that the limits $(q_{\alpha/2}, q_{1-\alpha/2})$ are the $\alpha/2$ and $1 - \alpha/2$ quantiles of the posterior distribution $p(\theta|y)$.

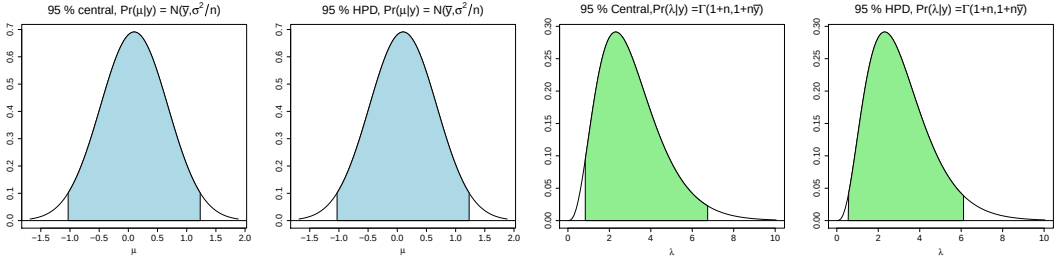
Types of credible interval

- As the beta posterior example mentioned earlier, there are multiple ways to define the credible region.
- One way would be to use the central credible region,

$$\Pr(q_{\alpha/2} \leq \theta \leq q_{1-\alpha/2}|y) = 1 - \alpha,$$

  so that the limits $(q_{\alpha/2}, q_{1-\alpha/2})$ are the $\alpha/2$ and $1 - \alpha/2$ quantiles of the posterior distribution $p(\theta|y)$.
- However, one of the main aims in inference is to maximise the precision of estimation. This suggests constructing a $100(1 - \alpha)$ % credible interval such that interval width is minimised. This is referred to as the highest posterior density (HPD) interval.

Example of credible interval

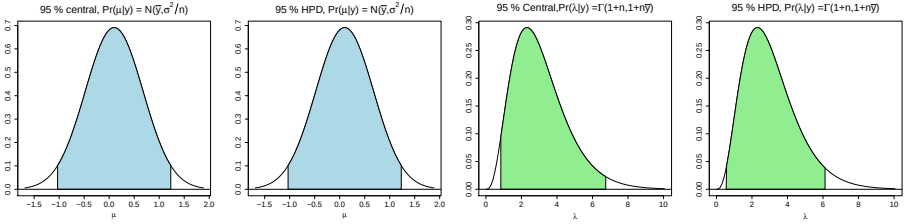


- While for the normal posterior, the central and HPD intervals are the same, $(-1.032, 1.232)$, the interval differs for the gamma posterior, with the central interval $(0.838, 6.744)$ being right shifted and about 6 % wider than the HPD interval $(0.548, 6.114)$.

Example of credible interval

- It turns out that the HPD interval is defined as follows,

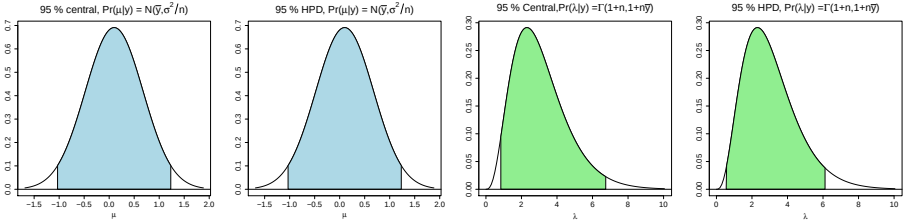
$$\Pr(\theta \in C|y) = 1 - \alpha \quad \text{s.t } p(\theta_1|y) \geq p(\theta_2|y)\ \forall \theta_1 \in C, \theta_2 \notin C,$$



Example of credible interval

- It turns out that the HPD interval is defined as follows,

$$\Pr(\theta \in C|y) = 1 - \alpha \quad \text{s.t } p(\theta_1|y) \geq p(\theta_2|y)\ \forall \theta_1 \in C, \theta_2 \notin C,$$



which in the case of the gamma posterior meant the limits of the HPD interval satisfy $p(\theta_L|y) = p(\theta_U|y)$ while for unimodal, symmetric distributions such as the normal, the $100(1-\alpha)$ % central and HPD intervals are equivalent.

Predictive distributions

In future lectures, we will use predictive distributions regarding the posterior distribution. Thus, we will define predictive distributions and consider an example.

- The prior predictive distribution:

$$p(y) = \int p(y, \theta) d\theta = \int p(y|\theta) p(\theta) d\theta,$$

also known as the marginal distribution of y, or the normalising constant.

Predictive distributions

In future lectures, we will use predictive distributions regarding the posterior distribution. Thus, we will define predictive distributions and consider an example.

- The prior predictive distribution:

$$p(y) = \int p(y, \theta) d\theta = \int p(y|\theta) p(\theta) d\theta,$$

  also known as the marginal distribution of y, or the normalising constant.

- The posterior predictive distribution:

$$p(\tilde{y}|y) = \int p(\tilde{y}, \theta|y) d\theta = \int p(\tilde{y}|\theta, y) p(\theta|y) d\theta = \int p(\tilde{y}|\theta) p(\theta|y) d\theta \quad \text{as } \tilde{y} \perp y|\theta,$$

  which is the distribution of a hypothetical new observation given the data already available. This is very useful in model checking.

Predictive distributions: Details

How to obtain the posterior predictive distribution is important for your understanding. Try to do it by yourselves. The posterior predictive distribution (assume $\tilde{y} \perp y|\theta$):

Predictive distributions: Details

How to obtain the posterior predictive distribution is important for your understanding. Try to do it by yourselves. The posterior predictive distribution (assume $\tilde{y} \perp y|\theta$):

$$\text{hint: } p(\tilde{y}|y) = \int p(\tilde{y}, \theta|y) d\theta = \int \frac{p(\tilde{y}, \theta, y)}{p(y)} d\theta$$

Predictive distributions: Details

How to obtain the posterior predictive distribution is important for your understanding. Try to do it by yourselves. The posterior predictive distribution (assume $\tilde{y} \perp y|\theta$):

$$\text{hint: } p(\tilde{y}|y) = \int p(\tilde{y}, \theta|y) d\theta = \int \frac{p(\tilde{y}, \theta, y)}{p(y)} d\theta$$

Which term do we want here?

Predictive distributions: Details

How to obtain the posterior predictive distribution is important for your understanding. Try to do it by yourselves. The posterior predictive distribution (assume $\tilde{y} \perp y|\theta$):

$$\text{hint: } p(\tilde{y}|y) = \int p(\tilde{y}, \theta|y) d\theta = \int \frac{p(\tilde{y}, \theta, y)}{p(y)} d\theta$$

Which term do we want here? The posterior $p(\theta|y) =$

Predictive distributions: Details

How to obtain the posterior predictive distribution is important for your understanding. Try to do it by yourselves. The posterior predictive distribution (assume $\tilde{y} \perp y|\theta$):

$$\text{hint: } p(\tilde{y}|y) = \int p(\tilde{y}, \theta|y) d\theta = \int \frac{p(\tilde{y}, \theta, y)}{p(y)} d\theta$$

Which term do we want here? The posterior $p(\theta|y) = p(\theta, y)/p(y)$.

Predictive distributions: Details

How to obtain the posterior predictive distribution is important for your understanding. Try to do it by yourselves. The posterior predictive distribution (assume $\tilde{y} \perp y|\theta$):

$$\text{hint: } p(\tilde{y}|y) = \int p(\tilde{y}, \theta|y) d\theta = \int \frac{p(\tilde{y}, \theta, y)}{p(y)} d\theta$$

Which term do we want here? The posterior $p(\theta|y) = p(\theta, y)/p(y)$.

$$p(\tilde{y}|y) = \int \frac{p(\tilde{y}, \theta, y)}{p(y)} d\theta = \int \frac{p(\tilde{y}|\theta, y)p(\theta, y)}{p(y)} d\theta = \int p(\tilde{y}|\theta, y)p(\theta|y) d\theta. \quad (1)$$

Then, using the assumption $\tilde{y} \perp y|\theta$, we conclude the result.

Predictive distributions: Details

How to obtain the posterior predictive distribution is important for your understanding. Try to do it by yourselves. The posterior predictive distribution (assume $\tilde{y} \perp y|\theta$):

$$\text{hint: } p(\tilde{y}|y) = \int p(\tilde{y}, \theta|y) d\theta = \int \frac{p(\tilde{y}, \theta, y)}{p(y)} d\theta$$

Which term do we want here? The posterior $p(\theta|y) = p(\theta, y)/p(y)$.

$$p(\tilde{y}|y) = \int \frac{p(\tilde{y}, \theta, y)}{p(y)} d\theta = \int \frac{p(\tilde{y}|\theta, y)p(\theta, y)}{p(y)} d\theta = \int p(\tilde{y}|\theta)p(\theta|y) d\theta. \qquad (2)$$

Then, using the assumption $\tilde{y} \perp y|\theta$, we conclude the result.

Predictive distribution example

Consider n conditional $i.i.d$ observations drawn from an exponential distribution with parameter λ. As the parameter $\lambda \in (0, \infty)$, a Gamma prior was chosen.

Predictive distribution example

Consider n conditional $i.i.d$ observations drawn from an exponential distribution with parameter λ. As the parameter $\lambda \in (0, \infty)$, a Gamma prior was chosen.

- The joint distribution $p(y_1, \dots y_n, \lambda)$ is

$$\prod_{i=1}^{n} p(y_i|\lambda)p(\lambda) = \lambda^n e^{-\lambda \sum_i^n y_i} \frac{\beta^\alpha \lambda^{\alpha-1} e^{-\beta\lambda}}{\Gamma(\alpha)} = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha+n-1} e^{-\lambda(\beta+\sum_i^n y_i)} \tag{3}$$

- From (3), we can deduce that the posterior distribution $p(\lambda|y_1, \dots y_n)$ is $\text{Ga}(\alpha + n, \beta + \sum_i^n y_i)$. The prior predictive distribution can be determined by marginalising λ from (3),

$$\begin{aligned} p(y_1, \dots y_n) &= \frac{\beta^\alpha}{\Gamma(\alpha)} \int \lambda^{\alpha+n-1} e^{-\lambda(\beta+\sum_i^n y_i)} d\lambda = \frac{\beta^\alpha}{\Gamma(\alpha)} \times 1 \Big/ \left(\frac{(\beta + \sum_i^n y_i)^{\alpha+n}}{\Gamma(\alpha+n)} \right) \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\alpha+n)}{(\beta+\sum_i^n y_i)^{\alpha+n}} = \frac{\Gamma(\alpha+n)}{\Gamma(\alpha)\beta^n} \left(1 + \frac{\sum_i^n y_i}{\beta}\right)^{-(\alpha+n)} \end{aligned}$$

Predictive distribution example

The posterior predictive distribution $p(\tilde{y}|y_1, \dots y_n)$ is

$$\begin{aligned}
\int p(\tilde{y}|\lambda)p(\lambda|y_1, \dots y_n)d\lambda &= \int \lambda e^{-\lambda\tilde{y}} \frac{(\beta + \sum_i^n y_i)^{\alpha+n}\lambda^{\alpha+n-1}e^{-(\beta+\sum_{i=1}^n y_i)\lambda}}{\Gamma(\alpha+n)} d\lambda \\
&= \frac{(\beta + \sum_i^n y_i)^{\alpha+n}}{\Gamma(\alpha+n)} \int \lambda^{\alpha+n+1-1} e^{-\lambda(\beta+\sum_i^n y_i+\tilde{y})} d\lambda \\
&= \frac{(\beta + \sum_i^n y_i)^{\alpha+n}}{\Gamma(\alpha+n)} \frac{\Gamma(\alpha+n+1)}{(\beta+\tilde{y}+\sum_i^n y_i)^{\alpha+n+1}} \\
&= \frac{\alpha+n}{\beta+\sum_i^n y_i} \left(1 + \frac{\tilde{y}}{\beta+\sum_i^n y_i}\right)^{-(\alpha+n+1)}, \\
&= \frac{a}{b}(1+\tilde{y}/b)^{-(a+1)},
\end{aligned}$$

- which is different from $p(y_1, \dots y_n)$ in the previous slide.