Why Bayes

June 7, 2021

Table of contents

Introduction

Bayesian approaches

- Typically contrasted with **frequentist** approaches
- Treat parameters as uncertain, data as fixed

Bayes' Rule

$$\underbrace{p(\theta|x)}_{posterior} = \frac{\underbrace{p(x|\theta)}_{likelihood} \; \underbrace{p(\theta)}_{prior}}{\underbrace{p(x)}_{evidence}} = \frac{p(x|\theta)p(\theta)}{\int p(x|\theta)p(\theta)\, d\theta}$$

Posterior

The posterior distribution is proportional to the prior times the likelihood:

$$p(\theta|x) \propto p(x|\theta)p(\theta)$$

The posterior distribution *is a distribution* over θ.

Evidence

The evidence, or *marginal likelihood*, can be used for model comparison.

Motivations

Motivations

Avoiding overfitting

Maximum likelihood and overfitting

Maximum likelihood can have problems with **over-fitting**.

Maximum likelihood and overfitting

Maximum likelihood can have problems with **over-fitting**.

The approach can be seen as *over-committing* to a single, fixed parameter value.

Maximum likelihood and overfitting

Maximum likelihood can have problems with **over-fitting**.

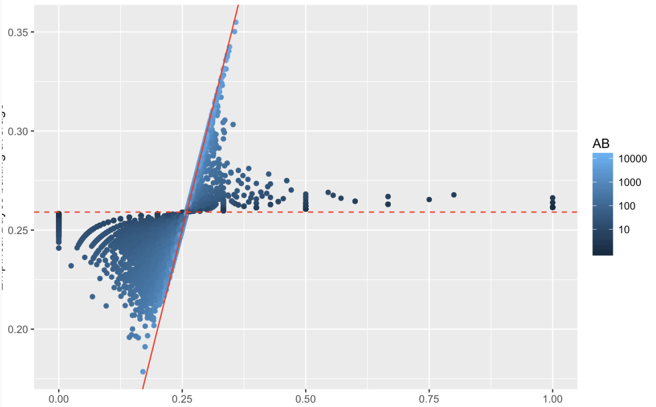
The approach can be seen as *over-committing* to a single, fixed parameter value.

Bayesian methods can correct this by treating parameters as random variables.

Bayesian estimation of batting averages

Let

- x be observed data (batting average after n at bats)
- θ be parameters (a player's 'true' batting average)

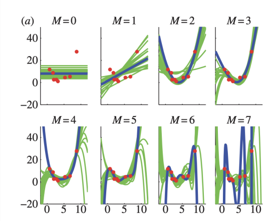


$\mathbb{E}[p(\theta \mid X)]$

Observed data (x)

Bayesian Occam's Razor

Maximum Likelihood (ML) solutions tend to overfit. Bayesian marginalization reduces overfitting.

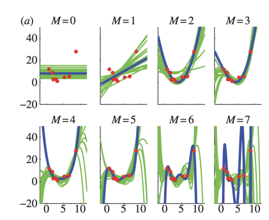


Models $y = f(x) + \epsilon$ of various complexity (polynomials of various order, M) were fit to 8 data points sampled from a quadratic model.

- Plotted are ML polynomials (least squares fits to the data under Gaussian noise) and posterior samples from a Bayesian model (which used a Gaussian prior for the coefficients, and an inverse gamma prior on the noise).
- How would you compare them?

Bayesian Occam's Razor

Maximum Likelihood (ML) solutions tend to overfit. Bayesian marginalization reduces overfitting.

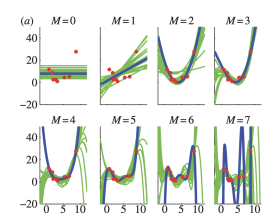


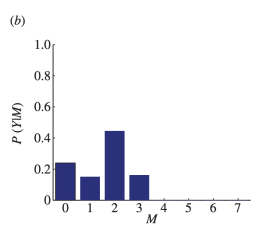
Models $y = f(x) + \epsilon$ of various complexity (polynomials of various order, M) were fit to 8 data points sampled from a quadratic model.

- Plotted are ML polynomials (least squares fits to the data under Gaussian noise) and posterior samples from a Bayesian model (which used a Gaussian prior for the coefficients, and an inverse gamma prior on the noise).
- How would you compare them? The ML estimate can look very different from a typical sample from the posterior!

Bayesian Occam's Razor

Maximum Likelihood (ML) solutions tend to overfit. Bayesian marginalization reduces overfitting.





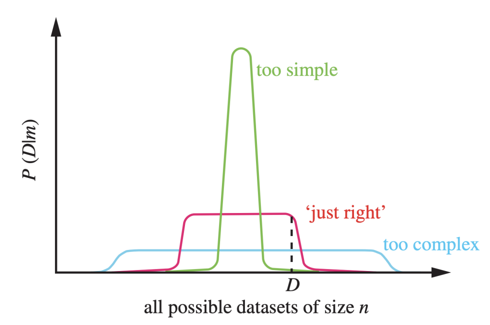
Models $y = f(x) + \epsilon$ of various complexity (polynomials of various order, M) were fit to 8 data points sampled from a quadratic model.

- Plotted are ML polynomials (least squares fits to the data under Gaussian noise) and posterior samples from a Bayesian model (which used a Gaussian prior for the coefficients, and an inverse gamma prior on the noise).
- How would you compare them? The ML estimate can look very different from a typical sample from the posterior!

The evidence is plotted as a function of model order. Model orders M=0 to M=3 have considerably higher evidence than other model orders. We see that Bayesian marginalization has reduced overfitting. (The maximum likelihood model, the $M = 7$ model, fits the data perfectly, but overfits wildly, predicting the function will shoot up or down between neighboring data points.)

Ghahramani, Z. (2013). Bayesian non-parametrics and the probabilistic approach to modelling. Philosophical Transactions of the Royal

Bayesian Occam's Razor



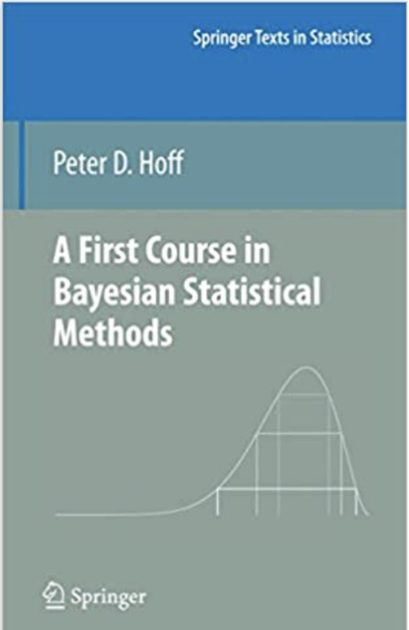
Competing probabilistic models correspond to alternative distributions over the datasets. Here, we have illustrated three possible models that spread their probability mass in different ways over these possible datasets. A *complex* model (shown in blue) spreads its mass over many more possible datasets, whereas a *simple* model (shown in green) concentrates its mass on a smaller fraction of possible data. Because probabilities have to sum to one, the complex model spreads its mass at the cost of not being able to model simple datasets as well as a simple model—this normalization is what results in an automatic Occam razor. Given any particular dataset, here indicated by the dotted line, we can use the marginal likelihood to reject both overly simple models, and overly complex models.

Ghahramani, Z. (2013). Bayesian non-parametrics and the probabilistic approach to modelling. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 371(1984), 20110553.

Motivations

Estimating the probability of a rare event

References


Springer Texts in Statistics
Peter D. Hoff
A First Course in Bayesian Statistical Methods
Springer

Description of problem

- Want to estimate the prevalence of an infectious disease in a small town.
- The higher the prevalence, the more public health precautions will be recommended.
- A small random sample of 20 individuals are checked for infection.

Description of problem

- **Parameter** θ, the fraction of infected individuals in the city.
- **Parameter space**: $\Theta = [0, 1]$
- **Sample**: Y the number of infected individuals in the sample
- **Sample space**: $\mathcal{Y} = \{0, 1, ..., 20\}$

Sampling model

If the value of θ were known, a reasonable sampling model for Y would be

$$Y \mid \theta \sim \text{Binomial}(20, \theta)$$

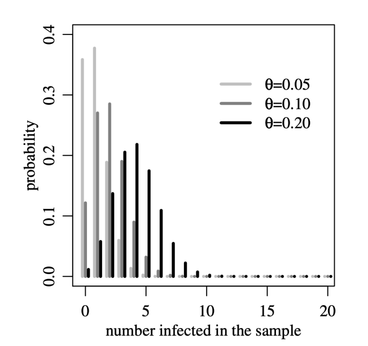


Figure 1: Binomial$(20, \theta)$ distributons for three values of θ.

Prior distribution

Other studies from various parts of the country indicate that the infection rate in comparable cities range from about 0.05 to 0.20, with an average prevalence of 0.10.

Prior distribution

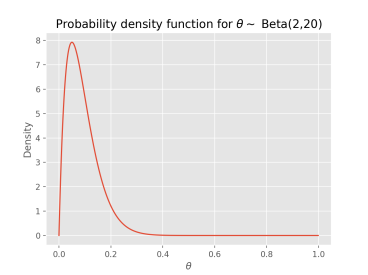
Other studies from various parts of the country indicate that the infection rate in comparable cities range from about 0.05 to 0.20, with an average prevalence of 0.10.

This suggests we use a prior distribution $p(\theta)$ that assigns a substantial amount of probability to the interval $(0.05, 0.20)$.

Prior distribution

Other studies from various parts of the country indicate that the infection rate in comparable cities range from about 0.05 to 0.20, with an average prevalence of 0.10.

This suggests we use a prior distribution $p(\theta)$ that assigns a substantial amount of probability to the interval $(0.05, 0.20)$.



We can encode this prior information using

$$\theta \sim \text{Beta}(2, 20)$$

From prior to posterior

Suppose $Y = 0$. How should we update our beliefs about θ?

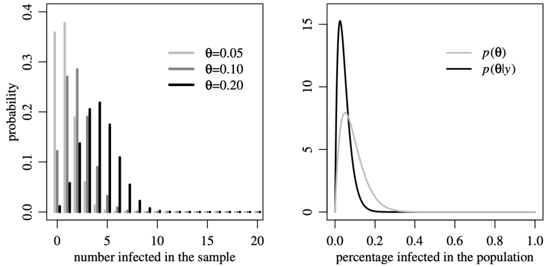


Fig. 1.1. Sampling model, prior and posterior distributions for the infection rate example. The plot on the left-hand side gives binomial$(20, \theta)$ distributions for three values of θ. The right-hand side gives prior (gray) and posterior (black) densities of θ.

Prior	**Posterior**
$\theta \sim \text{Beta}(2, 20)$	$\theta \mid \{Y = 0\} \sim \text{Beta}(4, 20)$
$\mathbb{E}[\theta] = 0.09$	$\mathbb{E}[\theta \mid \{Y = 0\}] = 0.048$
$\text{mode}[\theta] = 0.05$	$\text{mode}[\theta \mid \{Y = 0\}] = 0.025$
$P(\theta < 0.10) = 0.64$	$P(\theta < 0.10 \mid \{Y = 0\}) = 0.93$

Sensitivity analysis

Suppose we consider beliefs represented by Beta(a, b) distributions for (a, b) other than $(2, 20)$.

Sensitivity analysis

Suppose we consider beliefs represented by $\text{Beta}(a, b)$ distributions for (a, b) other than $(2, 20)$.

If $\theta \sim \text{Beta}(a, b)$, then $\theta \mid Y = y \sim \text{Beta}(a + y, b + n - y)$.

Sensitivity analysis

Suppose we consider beliefs represented by Beta(a, b) distributions for (a, b) other than $(2, 20)$.

If $\theta \sim \text{Beta}(a, b)$, then $\theta \mid Y = y \sim \text{Beta}(a + y, b + n - y)$.

The posterior expectation is

$$\begin{aligned}\mathbb{E}[\theta \mid Y = y] &= \frac{a + y}{a + b + n} \\ &= \frac{n}{a + b + n}\frac{y}{n} + \frac{a + b}{a + b + n}\frac{a}{a + b} \\ &= \frac{n}{w + n}\bar{y} + \frac{w}{w + n}\theta_0\end{aligned}$$

where $\theta_0 = a/(a + b)$ is the prior expectation of θ and $w = a + b$.

Sensitivity analysis

Suppose we consider beliefs represented by Beta(a, b) distributions for (a, b) other than $(2, 20)$.

If $\theta \sim \text{Beta}(a, b)$, then $\theta \mid Y = y \sim \text{Beta}(a + y, b + n - y)$.

The posterior expectation is

$$\begin{aligned} \mathbb{E}[\theta \mid Y = y] &= \frac{a + y}{a + b + n} \\ &= \frac{n}{a + b + n} \frac{y}{n} + \frac{a + b}{a + b + n} \frac{a}{a + b} \\ &= \frac{n}{w + n} \bar{y} + \frac{w}{w + n} \theta_0 \end{aligned}$$

where $\theta_0 = a/(a + b)$ is the prior expectation of θ and $w = a + b$.

Interpretation?

Sensitivity analysis

Suppose we consider beliefs represented by Beta(a, b) distributions for (a, b) other than $(2, 20)$.

If $\theta \sim \text{Beta}(a, b)$, then $\theta \mid Y = y \sim \text{Beta}(a + y, b + n - y)$.

The posterior expectation is

$$\begin{aligned}
\mathbb{E}[\theta \mid Y = y] &= \frac{a + y}{a + b + n} \\
&= \frac{n}{a + b + n} \frac{y}{n} + \frac{a + b}{a + b + n} \frac{a}{a + b} \\
&= \frac{n}{w + n} \bar{y} + \frac{w}{w + n} \theta_0
\end{aligned}$$

where $\theta_0 = a/(a + b)$ is the prior expectation of θ and $w = a + b$.

Interpretation? The posterior expectation is a compromise between the prior expectation θ_0 and sample mean $\bar{y}$. The weights on each depend on the sample size, n, and our prior confidence in this guess, w.

Sensitivity analysis

If someone provides us with a prior guess θ_0 and degree of confidence w, then we can approximate their prior beliefs about θ with

$$Beta\left(a = w\theta_0, \quad b = w(1 - \theta_0)\right)$$

And their posterior beliefs are represented with

$$Beta\left(a = w\theta_0 + y, \quad b = w(1 - \theta_0) + n - y\right)$$

Sensitivity analysis

If someone provides us with a prior guess θ_0 and degree of confidence w, then we can approximate their prior beliefs about θ with

$$Beta\left(a = w\theta_0, \quad b = w(1 - \theta_0)\right)$$

And their posterior beliefs are represented with

$$Beta\left(a = w\theta_0 + y, \quad b = w(1 - \theta_0) + n - y\right)$$

We can compute such a posterior distribution for a wide range of θ_0 and w values to perform a *sensitivity analysis*, an exploration of how posterior information is affected by differences in prior opinion.

Sensitivity analysis

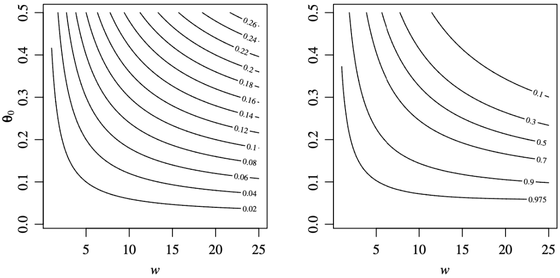


Fig. 1.2. Posterior quantities under different beta prior distributions. The left- and right-hand panels give contours of $\mathrm{E}[\theta|Y=0]$ and $\Pr(\theta < 0.10|Y=0)$, respectively, for a range of prior expectations and levels of confidence.

The second plot may be of use if, e.g., city officials would like to recommend a vaccine to the general public unless they were reasonably sure the current infection rate was less than 0.10.

Sensitivity analysis

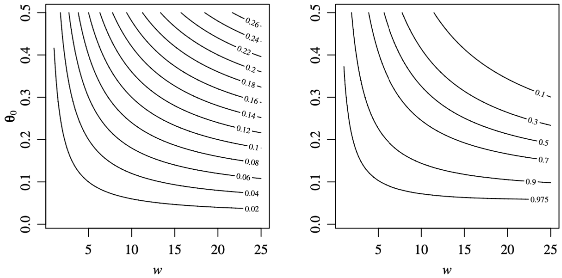


Fig. 1.2. Posterior quantities under different beta prior distributions. The left- and right-hand panels give contours of $\mathrm{E}[\theta|Y=0]$ and $\Pr(\theta < 0.10|Y=0)$, respectively, for a range of prior expectations and levels of confidence.

The second plot may be of use if, e.g., city officials would like to recommend a vaccine to the general public unless they were reasonably sure the current infection rate was less than 0.10.

A high degree of certainty (say 97.5%) is only achieved by people who already thought the infection rate was lower than the average of other cities.

Comparison to non-Bayesian methods

A 95% confidence interval for population proportion θ is the *Wald interval*, given by

$$\bar{y} \pm 1.96\sqrt{\bar{y}(1-\bar{y})/n}$$

Comparison to non-Bayesian methods

A 95% confidence interval for population proportion θ is the *Wald interval*, given by

$$\bar{y} \pm 1.96\sqrt{\bar{y}(1-\bar{y})/n}$$

The interval has *correct asymptotic frequentist coverage*, meaning that if n is large, then with probability approximately equal to 95%, Y will take on a value y such that the above interval contains θ.

Comparison to non-Bayesian methods

A 95% confidence interval for population proportion θ is the *Wald interval*, given by

$$\bar{y} \pm 1.96\sqrt{\bar{y}(1-\bar{y})/n}$$

The interval has *correct asymptotic frequentist coverage*, meaning that if n is large, then with probability approximately equal to 95%, Y will take on a value y such that the above interval contains θ.

For our sample in which $\bar{y} = 0$, the Wald confidence interval comes out to be just a single point: 0.

Comparison to non-Bayesian methods

A 95% confidence interval for population proportion θ is the *Wald interval*, given by

$$\bar{y} \pm 1.96\sqrt{\bar{y}(1-\bar{y})/n}$$

The interval has *correct asymptotic frequentist coverage*, meaning that if n is large, then with probability approximately equal to 95%, Y will take on a value y such that the above interval contains θ.

For our sample in which $\bar{y} = 0$, the Wald confidence interval comes out to be just a single point: 0.

In fact, the 99.99% Wald interval also comes out to be zero.

Comparison to non-Bayesian methods

People have suggested alternatives to avoid this type of behavior.

Comparison to non-Bayesian methods

People have suggested alternatives to avoid this type of behavior.

The "adjusted" Wald interval suggested by Agresti and Coull (1998) is given by

$$\widehat{\theta} \pm \sqrt{\widehat{\theta}(1 - \widehat{\theta})/n}, \quad \text{where}$$
$$\widehat{\theta} = \frac{n}{n+4}\bar{y} + \frac{4}{n+4}\frac{1}{2}$$

Comparison to non-Bayesian methods

People have suggested alternatives to avoid this type of behavior.

The "adjusted" Wald interval suggested by Agresti and Coull (1998) is given by

$$\widehat{\theta} \pm \sqrt{\widehat{\theta}(1-\widehat{\theta})/n}, \quad \text{where}$$
$$\widehat{\theta} = \frac{n}{n+4}\bar{y} + \frac{4}{n+4}\frac{1}{2}$$

While not motivated as such, the interval is clearly related to Bayesian inference: $\widehat{\theta}$ is equivalent to the posterior mean for θ under a Beta$(2,2)$ prior, which represents weak prior information centered around $\theta = 1/2$.

Comparison to non-Bayesian methods

Compared to the post-hoc "adjustment" approach, the Bayesian formalism provides

- Reasonable conclusions which fall naturally out of the framework
- Flexibility to other choice of priors than $\mathrm{Beta}(2, 2)$
- Sensitivity analysis to consider the sets of conclusions that would be reached by people with different priors.
- Simultaneous access to various functionals of the posterior – not just $\mathbb{E}[\theta \mid Y = y]$ but also $\mathbb{P}[\theta < 0.10 \mid Y = 0]$.

Extensions: Hierarchical models

- Hierarchical models use "surrounding data" as a prior in a more formal way.
- In Hoff's disease prevalence example, we constructed our beta prior manually, by taking a couple of basic facts about similar towns and then converting that into beta parameters.
- A hierarchical model could let the prior expectation be tied more exactly to those surrounding towns. We can automatically set the strength of the prior expectation according to the relative uncertainty within and between towns, and to automatically adapt as data rolls in.
- Hierarchical regressions allow the prior expectation to be more strongly influenced by towns that are similar w.r.t. relevant characteristics, such as size, SES, etc

Conclusion

Summary

What are some advantages of the Bayesian approach?

Summary

What are some advantages of the Bayesian approach?

- Reduces overfitting

Summary

What are some advantages of the Bayesian approach?

- Reduces overfitting
- Automatic complexity control

Summary

What are some advantages of the Bayesian approach?

- Reduces overfitting
- Automatic complexity control
- Exploits prior knowledge (previous results, reasonable values of data, etc.)

Summary

What are some advantages of the Bayesian approach?

- Reduces overfitting
- Automatic complexity control
- Exploits prior knowledge (previous results, reasonable values of data, etc.)
- Immediate access to many inferential quantities of inference

Summary

What are some advantages of the Bayesian approach?

- Reduces overfitting
- Automatic complexity control
- Exploits prior knowledge (previous results, reasonable values of data, etc.)
- Immediate access to many inferential quantities of inference
- Natural (and more flexible) solutions to frequentist problems:

  (adjustments to confidence intervals, regularization, etc.)

Summary

What are some advantages of the Bayesian approach?

- Reduces overfitting
- Automatic complexity control
- Exploits prior knowledge (previous results, reasonable values of data, etc.)
- Immediate access to many inferential quantities of inference
- Natural (and more flexible) solutions to frequentist problems:
  (adjustments to confidence intervals, regularization, etc.)
- Can be easier in practice to extend to more complex models

Utility for large datasets

- Still useful for larger models

- Especially complex models – what really matters is the information we get about a given parameter.

  A big dataset is just a bunch of small datasets

- Example: biometric profiling. (A given bigram may be rare to type, but in the end, you're typing *some* rare bigram a high percentage of the time!)