A Brief Introduction to Bayesian Statistics

Christian Mueller

Very Applied Methods MT 2019

Contents

1. Theory
2. Sampling
3. Priors
4. Summary
5. Further topics

How to Bayes?

Statistical Framework:

- Subjective Probability
- Likelihood Principle

Bayes' Theorem:

$$P(\theta \mid y) \propto P(y \mid \theta) \cdot P(\theta)$$

The posterior is proportional to the likelihood times the prior.

Sampling:

- Monte Carlo Principle
- Markov Chain Monte Carlo (MCMC)

Statistical Framework

This is arguably a philosophical debate about the ‘right’ way to do statistical inference.

Bayesian Probablity

The probability for event **A** is the **degree of belief** for the event to happen.

Likelihood Principle

The Likelihood Principle Berger and Wolpert (1988) states that all relevant for inference about a parameter comes through the likelihood

Bayesian vs. Frequentist Statistics

Scrapped because of time :(

See Jeff Gill: Bayesian Methods, page 64.

Bayesian Mantra

‘Inverting’ conditional probability for events **A** and **B**:

$$P(A \mid B) = \frac{P(A) \cdot P(B \mid A)}{P(B)}$$

Conditional probability for random variables

$$P(\theta \mid \mathbf{D}) = \frac{P(\theta) \cdot P(\mathbf{D} \mid \theta)}{P(\mathbf{D})}$$

$P(data)$ can be interpreted as a standardizing factor which ensure that the probability is well-defined, i.e. sums to 1.

Bayesian Mantra

$$P(\theta \mid \text{data}) \propto P(\theta) \cdot P(\text{data} \mid \theta)$$

The posterior is proportional to the prior times the likelihood.

Example: Estimating a simple linear regression

$$Y = \alpha + \beta \cdot X + \epsilon, \quad \text{with } \epsilon \sim \mathcal{N}(0, \sigma)$$

$$Y = \mathcal{N}(\alpha + \beta \cdot X, \sigma)$$

$$P(\alpha, \beta, \sigma \mid Y, X) = \frac{P(\alpha) \cdot P(\beta) \cdot P(\sigma) \cdot P(Y \mid \alpha, \beta, \sigma, X)}{P(Y, X)}$$

$$P(\alpha, \beta, \sigma \mid Y, X) \propto P(Y \mid \alpha, \beta, \sigma, X) \cdot P(\alpha) \cdot P(\beta) \cdot P(\sigma)$$

The posterior is proportional to the likelihood times the prior.

Inference

- We defined this beautiful multi-variate posterior distribution which comprises everything we know about the parameters (given prior and data).

$$P(\alpha, \beta, \sigma)$$

- What we are actually interested in are the marginal distribution of each parameter:

$$P(\alpha \,|\, \beta, \sigma) \quad P(\beta \,|\, \alpha, \sigma) \quad P(\sigma \,|\, \alpha, \beta)$$

- For the marginal distribution of one parameter, you need to integrate all the other variables out.
- Analytically this is very hard in the general case.
- Choosing conjugate prior distributions can help alleviate this because it is easier to analytically compute the marginal distribution.
- **Sampling to the rescue!**

MC, MC, and MCMC

Monte Carlo Principle

We can learn everything we want to know about a distribution by generating a lot of random samples from it.

Markov Chains

Mathematical tools to model how a system moves from one state to another state, where the new state depends on the previous state.

MCMC

- A system that describes how you iterate from one sample to the next (Markov chains) and take random samples in each iteration (Monte Carlo Principle) where the current random sample only depends on the previous random sample.
- The Markov chain framework is used to (mathematically) prove that MCMC will converge to the true distribution as (under certain conditions).

MCMC: Algorithms

Sampling a single variable

- Requirement: You need to be able to evaluate the density function
- Algorithm: (Adaptive) Rejection Sampling

Gibbs' Sampling

- Move through the sampling space by sampling a single variable by conditioning on the previous values of all other variables.
- This corresponds to a multivariate random sample for each full iteration!

Other algorithms

- Metropolis-Hastings
- Slice sampling
- Hamiltonian Monte Carlo

Convergence diagnostics

- The sampling algorithm needs to be stopped at some time $t_0 < \infty$
- Did the algorithm run 'long enough' to explore the whole space of the posterior distribution?

Methods

- Visual: Traceplot
- Test: Geweke
- Test: Heidelberger-Welch
- Test: Potential scale reduction factor (Gelman and Rubin)

Caveat

You cannot test for **convergence**. All tests are for **non-convergence**.

R Example: Bayesian simple linear regression

$$Y = \alpha + \beta \cdot X + \epsilon, \quad \text{with } \epsilon \sim \mathcal{N}(0, \sigma)$$

$$Y = \mathcal{N}(\alpha + \beta \cdot X, \sigma)$$

$$P(\alpha, \beta, \sigma \mid Y, X) = \frac{P(\alpha) \cdot P(\beta) \cdot P(\sigma) \cdot P(Y | \alpha, \beta, \sigma, X)}{P(Y, X)}$$

Choosing Priors

General rule: The more information in the likelihood, the less influential the prior.

Distribution

- The distribution is often already ‘fixed’ by …

… the implementation.

… prior research or standard models.

… mathematical tractability (conjugacy).

Mean and variance

- This is the part where you have to make all the decisions as a researcher.
- You often want to choose some kind of uninformative prior.
- The **mean** is usually set to 0 (in the case of parameters).
- The **variance** is usually set to be large.
- However, it does not need to be *extremely* large.

R Example: Informative and uninformative priors

Go back to the simple linear regression example.

- Play around with the prior variance.
- Play around with the prior mean.
- See how much information you need to give into the priors to sway the posterior distribution.
- Reduce the sample size of the data and compare again.

Controversies around Bayesian statistics

- Philosophical differences around what constitutes a probability.
- You have to make full parametric assumptions for priors and likelihood.
- You can estimate fundamentally non-identified models!
- You can inject information through the priors which will influence the posterior

R Code: Fundamentally unidentified model

Let $Y \sim \mathcal{N}(\mu_0, \sigma_0)$ be the true data generating process.

Estimate a model:

$$Y \sim \mathcal{N}(\hat{\mu}_1 + \hat{\mu}_2, \hat{\sigma})$$

$\hat{\mu}_1$ and $\hat{\mu}_2$ are not identified!

Estimates for $\hat{\mu}_1$ and $\hat{\mu}_2$ can be any values as long as $\hat{\mu}_0 = \hat{\mu}_1 + \hat{\mu}_2$ is satisfied.

$\hat{\mu}_1$ and $\hat{\mu}_2$ can be estimated in a Bayesian framework and $\hat{\mu}_1 + \hat{\mu}_2$ will be an estimate for μ_0!

Why Bayes?

- The likelihood is central to all parametric statistical inference, e.g. MLE.
- The more complicated the model becomes, the harder it becomes to maximise the likelihood function to estimate parameters.
- Bayesian estimation will not magically resolve those issues!
- BUT it allows you to ...

... replace the problem of finding a maximum with the problem of sampling from the posterior:

... 'identify' a model by injecting a small amount information into the model, often to constrain the posterior variance of the parameters to a 'sensible' range.

Why Bayes? (cont'd)

- Some statistical models lend themselves to Bayesian estimation because they easily translated from their specification, e.g. multilevel/hierarchical models
- For certain research questions you can only observe a limited amount of data to infer a complicated (hidden) theoretical process:
- Example IRT (Factor analysis): You estimate θ for each individual and α and β for each item.
- You cannot add additional data (individuals/items) without increasing the number of estimated parameters!
- Bayesian estimation allows you to make inferences about the hidden process in the absence of an overwhelming amount of information.
- You *buy* this possibility with more assumptions about the prior distribution of model parameters.
- Other examples: Exponential Random Graph Models for networks or topic models for text.

That's it!

Alternatives to sampling

Sampling has the favourable property that it is *guaranteed* to converge to the true posterior distribution as $n_{\mathsf{samples}} \to \infty$.
Alternatives have been developed with the main idea to change sampling (back) to an optimisation problem.

- MAP
- Collapsed sampling
- EM and Variational Inference

All methods share one important drawback:

- You get sensible estimates for the posterior mean (expected value)
- You usually **do not** get sensible estimates for the posterior variance

Alternatives to sampling: Variational Inference

- Idea: Approximate the true distribution by a distribution that is easier to work with but ‘close’ to the true distribution.
- Propose a target distribution from a family of distributions close to the actual one
- Minimise the (KL-) difference between the two distributions.
- Main advantage (compared to MCMC): Potentially very fast!
- Main downsides (compared to MCMC):
  - The approximation **underestimates** (usually by a large degree) the variation of the true distribution.
  - Writing a VI algorithm is complicated and specific to the model and priors.
  - The properties of variational approximation are not well studied.
- See here for a ‘gentle’ introduction: Blei et al. (2017)
- VI is used in many advanced models with large scale data, e.g. topic modelling

References

Berger, James O. and Robert L. Wolpert (1988). *The Likelihood Principle*. 2nd ed. Hayward, CA: Institute of Mathematical Statistics.

Blei, David M., Alp Kucukelbir and Jon D. McAuliffe (2017). 'Variational Inference: A Review for Statisticians'. In: *Journal of the American Statistical Association* 112.518, pp. 859–877. DOI: 10.1080/01621459.2017.1285773.