Bayesian workflow

The (Bayesian) workflow for modeling can be summarized in the following steps:

1. Scope the problem
2. Specify likelihood and priors
3. Generate fake data that resemble the true data to a reasonable degree
4. Fit the model on fake data
    a. Check that the true values are recovered
    b. Check the fit of the model
5. Fit model on real data
6. (Optional) Specify and integrate the utility function over posterior distribution

In what follows we will use $\mathcal{M}$ to denote the model, θ for the model parameters, and $\mathcal{D}$ for the data. The model is usually proposed by the researcher as a function of parameters, but the true parameter values are typically unknown before the analysis.

Scope of problem

Bayesian modeling is one method of statistical inference. But why do we need statistical inference in the first place? We need it to answer our questions about the world. Usually our questions refer to an implicit or an explicit parameter θ in a statistical model, such as

- what values of θ are most consistent with the data?
- does the data support a certain condition, e.g. $\theta > 0$?
- what ranges of values for θ are consistent with the data?

To proceed further with the statistical analysis we need to define clearly the model we will be using. The choice of model is usually tied up to the exact research questions we are interested in. We can choose to start with a postulated data generation process and then decide how to interpret the parameters in relation to the research question. Alternatively, it is equally valid to start from the research question and design the model in such a way so that the model parameters are directly connected to the the specific questions we wish to answer. In the case study we provide an example to illustrate this how a model is designed to answer a specific research question.

Specify likelihood and priors

Once we have defined the scope of the problem we need to finalize the design of the model and define it precisely, which is captured in the likelihood function. This function ties together the parameters to the data and allows us to learn the former from the latter.

The second ingredient of Bayesian inference is the prior distribution. Priors are inescapably part of the Bayesian approach and hence have to be considered carefully. The goal of Bayesian inference is to combine the prior knowledge with the evidence from the data to come up with the posterior distribution. It is difficult to predict how sensitive the final results will be to a change in the priors. However, as a general rule, the priors have a bigger impact when the data has less information.

The ideal scenario for the Bayesian approach is when prior knowledge is available and can be captured in the prior distributions of the parameters. But sometimes we might want to avoid expressing prior knowledge, especially when such knowledge is not available. Researchers often, in an effort to be as objective as possible, are interested to express minimal or no prior knowledge. Priors that express very little or no prior knowledge are called vague or uninformative priors. In fact, Bayesian inference works even when the prior is technically not a proper distribution but a function that assumes all values are equally likely, referred to as *improper prior*. However, it is generally advisable to avoid improper priors and if no prior knowledge is available it is still better to use a very vague normal distribution with big variance. Improper or vague priors can have unexpected results especially when it comes to model selection.

Another incentive to avoid vague priors is the additional benefits that priors can have. For example sometimes priors can also be used to guard against overfitting by pulling the parameters away from improbable values. Bayesian critics often see priors as a weakness, whereas in reality priors are an opportunity. Priors are an opportunity to use our knowledge to guide the inference in the absence of evidence from the data. Also, it is important to remember that we rarely have absolutely no prior knowledge and that we consider any parameter value equally likely.

Generate fake data

Once we have agreed on a generative process, a model $\mathcal{M}$, we can use it to simulate data $\mathcal{D}'$. Note how this is a hypothesized process and comes with necessary assumptions and simplifications. It's highly unlikely that the real world follows exactly model $\mathcal{M}$. However, if $\mathcal{M}$ is close enough to the real generative process, it can still be very useful to help us understand something about the world. All models are wrong, but some models are useful.

If simulating fake data using our generative process $\mathcal{M}$ is the forward direction, statistical inference is the reverse direction. Inference is the process by which we start with data and try to find what parameters that could have produced such data, under $\mathcal{M}$.

Fit the model to fake data

Statistical inference, also referred to as fitting the model to data or parameter estimation, is the process of finding θ using the data $\mathcal{D}$. The most popular statistical inference algorithm is maximum likelihood estimation (MLE) which finds the parameter values that maximize the likelihood given the observed data.

Under the Bayesian approach we treat the parameters θ as random variables and express our prior knowledge about θ with the prior probability distribution. Bayesian inference is the process of updating our beliefs about θ in light of the data $\mathcal{D}$. The update uses Bayes theorem and results in the conditional distribution $\pi(\theta|\mathcal{D})$, called the posterior distribution. Bayesian inference is generally a hard problem. In most cases we cannot find the exact distribution; instead we settle for an algorithm that returns samples from the posterior distributions.

a. Once we recover posterior samples for the model parameters we need to check them for correctness. A typical test is whether the confidence intervals of the parameters include the true parameter values that were used to generate the fake data.

b. More generally, this is a good time to check the model fit and decide if we need to change our model. This step is specific to each application. The purpose is to rule out mistakes and confirm that the inference procedure works, at least if the data generation process is true. Success at this stage is not sufficient guarantee that the model will fit well on the real data but it's a necessary condition for proceeding further.

Fit model on real data

Once we have finalized the design of our model and have tested it on fake data we are ready to fit it to the real data and get the results. Usually we are interested in a specific metric that is derived from the posterior samples. It is important to build confidence in the power of our inference algorithm before we proceed to interpreting the results.

Further reading

For a concise overview of statistical modeling and inference, including a high level comparison to the frequentist approach, see :cite: `Wood15` . For a more extended treatment of the Bayesian approach, including utility functions, see :cite: `robert2007bayesian` . For an accessible Bayesian modeling primer, especially for beginner Bayesians, see :cite: `McElreath15` and :cite: `Marin2006` .