Bayesian Statistics II: Techniques and Models

1 Statistical modeling and Monte Carlo estimation

1.1 Statistical modeling

1.1.1 Objectives

A statistical model is a mathematical model used to **imitate and approximate the data generating process**. It describes the relationship between variables while taking into account uncertainty in the data.

For what kind of problems may we use a statistical model? Here are four common objectives:

1. **Quantify uncertainty**
    - "We're 99% confident the probability of heads is between 0.48 and 0.52."
2. **Inference**
    - "Given poll results, what does that tell us about the rest of the population?"
3. **Measure support for hypothesis**
4. **Prediction**
    - "We have demographic information about a voter, but not about which candidate she supports. A statistical model can predict their voting decision."

Neural networks excel at prediction, which is almost always important. However, they are also black-boxes, and therefore do not necessarily deliver on the rest of the objectives. Statistical models attempt to balance all four objectives.

1.1.2 Modeling process

1. Understand the problem
2. Collect relevant data
    - For generalisable results, make sure samples are random and representative
3. Explore your data

- Inspect your data for possible errors
- Visualise your data to inform future model choice

4. Postulate model
    - May have to balance off model complexity and generalisability (known as the bias-variance trade off)
5. Fit model
    - Choose between Bayesian and Frequentest approaches
6. Check model
7. Iterate (go back to 4 - 6) if model is inadequate
8. Use model to arrive at conclusions

1.2 Bayesian modeling

1.2.1 Components of Bayesian models

Suppose we have data that consist of the heights of $n = 15$ adult men. We'll assume their heights are normally distributed as follows:

$$y_i = \mu + \epsilon_i \qquad \epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2) \qquad i = 1, ..., n$$

Or, more succinctly,

$$y_i \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$$

So far this model is the same for Frequentists and Bayesians. Obtaining uncertainty estimates is where they differ:

- Frequentists consider the parameters (μ and σ^2) as fixed but unknown. To fit the model, we estimate them with MLE. Uncertainty in these estimates are dependent on how much the MLE changes with many different samples of $n = 15$ men.
- Bayesians solve the problem of uncertainty estimates by placing probability distributions over parameters (i.e. priors).

There are three components to Bayesian models:

- The **likelihood**, written as $p(y|\theta)$, which is the probabilistic model for the data.
- The **prior**, written as $p(\theta)$, which characterises our uncertainty of the parameters.
- The **posterior**, written as $p(\theta|y)$, is fully determined by the likelihood and prior. Our choice of likelihood and prior determines how we update our beliefs about the parameters given the data.

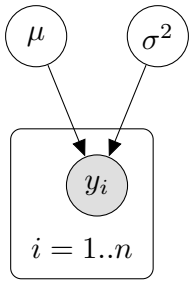
1.2.2 Model specification

We'll consider a hierarchical model, where priors are places on the parameters μ and σ^2. When specifying a model, we usually start with the likelihood.

$$\begin{aligned} y_i|\mu,\sigma^2 &\overset{iid}{\sim} \mathcal{N}(\mu,\sigma^2) \qquad i=1,...n \\ \mu &\sim \mathcal{N}(\mu_0,\sigma_0^2) \\ \sigma^2 &\sim \mathcal{IG}(\nu_0,\beta_0) \end{aligned}$$

The conjugate prior for μ when σ^2 is known is the normal distribution. The conjugate prior for σ^2 when μ is known is the inverse gamma distribution (or gamma if using precision). Hence the normal distribution for μ and inverse gamma for σ^2 is a sensible choice.

The corresponding graphical model is[1]:



We can now simulate data from this model. Simulating μ and σ^2 allows us to simulate $y_1,...,y_n$ further down the chain.

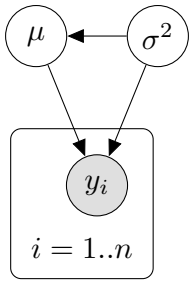
1.2.3 Posterior derivation

Lets consider a three layer hierarchical model, where μ now depends on σ^2.

$$\begin{aligned} y_i|\mu,\sigma^2 &\overset{iid}{\sim} \mathcal{N}(\mu,\sigma^2) \qquad i=1,...n \\ \mu|\sigma^2 &\sim \mathcal{N}(\mu_0,\frac{\sigma^2}{\omega_0}) \\ \sigma^2 &\sim \mathcal{IG}(\nu_0,\beta_0) \end{aligned}$$

The corresponding graphical model is as follows:

[1]We can put y_i in the plate since the random variables $\{y_1,...,y_n\}$ are in *exchangeable*, since they are iid. Intuitively, exchangeability means we can reorder variables in the sequence without changing their joint distribution.



To simulate data from this model, we'd have to first simulate σ^2 to simulate μ. We then use these simulations of σ^2 and μ to simulate $y_1, ..., y_n$.

Once we have a model specification, we can write down the full posterior over all parameters given the data. The joint distribution is as follows:

$$
\begin{aligned}
p(y_1, ..., y_n, \mu, \sigma^2) &= p(y_1, ..., y_n|\mu, \sigma)p(\mu|\sigma^2)p(\sigma^2), \quad \text{by the chain rule} && (1)\\
&= \prod_{i=1}^{n} \left[\mathcal{N}(y_i|\mu, \sigma^2)\right] \mathcal{N}(\mu|\mu_0, \frac{\sigma^2}{\omega_0})\mathcal{IG}(\sigma^2|\nu_0, \beta_0) && (2)\\
&\propto p(\mu, \sigma^2|y_1, ..., y_n) && (3)\\
& && (4)
\end{aligned}
$$

To see this, remember that the joint is simply the numerator in Bayes theorem:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{\int p(y|\theta)p(\theta)d\theta} \propto p(y|\theta)p(\theta)$$

Since the denominator is a function of y, which are known values, it is just a constant.

The only thing missing in (2) is a normalising constant. **If we can recognise that this expression is proportional to a distribution, our work is done and we know what our posterior distribution is**. This is the motivation behind choosing conjugate priors. If we do not use conjugate priors, or if the models are more complicated, then the posterior distribution will not have a standard form that we can recognise.

1.2.4 Non-conjugate models

Lets consider models whose posteriors aren't recognisable distributions. For example,

$$
\begin{aligned}
y_i|\mu &\overset{iid}{\sim} \mathcal{N}(\mu, 1) \qquad i = 1, ...n\\
\mu &\sim \mathcal{T}(0, 1, 1)
\end{aligned}
$$

The posterior is

$$\begin{aligned} p(\mu|y_1,...,y_n) &\propto \prod_{i=1}^{n}\left[\frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}(y_i-\mu)^2)\right]\frac{1}{\pi(1+\mu^2)} \\ &\propto \exp(-\frac{1}{2}\sum_{i=1}^{n}(y_i-\mu)^2)\cdot\frac{1}{1+\mu^2} \\ &\propto \exp(-\frac{1}{2}\sum_{i=1}^{n}(y_i-2y_i\mu+\mu^2))\cdot\frac{1}{1+\mu^2} \\ &\propto \frac{\exp\left(n\left(\bar{y}\mu-\frac{\mu}{2}\right)\right)}{1+\mu^2} \end{aligned}$$

This doesn't follow any standard distribution (it's close to being proportional to a normal distribution, but $1+\mu^2$ appears in the denominator). We have the posterior distribution up to a normalizing constant, but we are unable to integrate it to obtain important quantities, such as the posterior mean or probability intervals. In low dimensional problems with only a few parameters, we can resort to numerical methods for integration, but this solution only works for a narrow set of models. Computational methods invented in the 1950s allow us to simulate the posterior of such models, which is the topic of the next section.

1.3 Monte Carlo estimation

Monte Carlo estimation refers to simulating hypothetical draws from a probability distribution in order to calculate important quantities of that distribution. Some of these quantities might include the mean, the variance, the probability of some event, or the quantiles of the distribution. All of these calculations involve integration, which, except in the simplest cases, can be very difficult or even impossible to compute analytically.

Suppose

$$\theta \sim \text{Gamma}(a,b)$$

We can calculate the expected value analytically:

$$\mathbb{E}\left[\theta\right] = \int_0^\infty \theta p(\theta) = \frac{a}{b}$$

We can also estimate this integral with Monte Carlo estimation. To do so we draw a large number of simulations θ_i^* for $i = 1,...,m$. The sample mean $\bar{\theta} = \frac{1}{m}\sum_{i=1}^{m}\theta_i^*$ provides an unbiased estimate of the theoretical mean $\mathbb{E}(\theta)$. That is,

$$\mathbb{E}\left[\theta\right] \approx \bar{\theta} = \frac{1}{m}\sum_{i=1}^{m}\theta_i^* \qquad \theta_i^* \overset{iid}{\sim} \text{Gamma}(a,b)$$

In general, Monte Carlo estimation can approximate

$$\mathbb{E}\left[h(\theta)\right] \approx \frac{1}{m}\sum_{i=1}^{m} h(\theta_i^*) \qquad \theta_i^* \overset{iid}{\sim} \text{Gamma}(a, b)$$

Example. Let $h(\theta) = I_{\theta<5}(\theta)$.

$$\begin{aligned}
\mathbb{E}\left[h(\theta)\right] &= \int_0^\infty h(\theta)p(\theta)d\theta \\
&= \int_0^\infty I_{\theta<5}(\theta)p(\theta)d\theta \\
&= \int_0^5 p(\theta)d\theta \\
&= \mathbb{P}(0 < \theta < 5) \\
&\approx \frac{1}{m}\sum_{i=1}^{m} I_{\theta^*<5}(\theta_i^*) \qquad \theta_i^* \overset{iid}{\sim} \text{Gamma}(a, b)
\end{aligned}$$

Example. To estimate the 90th percentile of the distribution, we may sample and sort (in ascending order) a large number of θ_i^*. We then pick the smallest θ_i^* that's greater than 90% of the others.

1.4 Monte Carlo error and marginalisation

TODO: finish section