Introduction

Bayesian inference

See [1] for a full treatment on all aspects of Bayesian data analysis.

In the following $\boldsymbol{\theta}$ denotes a vector of model parameters, and $\boldsymbol{x}$ is a vector of observed data. Our aim is to make inferences about $\boldsymbol{\theta}$ that are consistent with the observed data. The process of data analysis can be viewed as consisting of three steps:

1. Setting up a full probability model. This means formulating a joint probability model for $\boldsymbol{\theta}$ and $\boldsymbol{x}$: $p(\boldsymbol{\theta}|\boldsymbol{x})$.
2. Conditioning on the observed data. This entails looking at an posterior distribution, explained below.
3. Evaluating the model. If the model does not fit the observed data, it can be changed, and the process starts from step 1 again.

In step 2 above, we need to make inferences about $\boldsymbol{\theta}$, which are the variables we want to estimate, given the observed data $\boldsymbol{x}$. If we just condition on the observed data, we can make use of Bayes' rule and write the conditional distribution as:

$$p(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{x})}$$

The factor $p(\boldsymbol{x})$ does not depend on $\boldsymbol{\theta}$, and since the data is given, that is; $\boldsymbol{x}$ does not change, $p(\boldsymbol{\theta}|\boldsymbol{x})$ is a function of $\boldsymbol{\theta}$ only, we can view $p(\boldsymbol{x})$ as a constant. We can thus write:

$$p(\boldsymbol{\theta}|\boldsymbol{x}) \propto p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})$$

There are expressions for the two terms on the right side. $p(\boldsymbol{x}|\boldsymbol{\theta})$ is termed the likelihood, and noting since $\boldsymbol{x}$ is constant, it is a function of $\boldsymbol{\theta}$. $p(\boldsymbol{\theta})$ is termed the prior distribution.

Approximate Bayesian Computation

The likelihood function may be difficult to evaluate, for instance it may be computationally intractable to do so, or the function does not have an analytic form. In such settings what are termed likelihood-free methods can be employed to draw inferences. The name refers to the fact that the methods do not explicitly evaluate the likelihood function, but instead approximate it using simulations. A requirement is then that we must at least be able to generate samples from the likelihood.

Approximate Bayesian Computation is one likelihood-free method.

To understand how likelihood-free methods work, let us look at the posterior again, and introduce an auxillary variable, $\boldsymbol{x}$:

Algorithm 1 Likelihood-free

```
counter ← 0
samples ← empty array
while counter < N do
    Sample θ from π(θ)
    Sample x from π(x|θ)
    if x = y then
        samples[counter] ← θ
    end if
end while
```

$$p(\boldsymbol{\theta}, \boldsymbol{x}|\boldsymbol{y}) \propto p(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{y})p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})$$

The posterior can be retrieved by integrating out $\boldsymbol{x}$:

$$p(\boldsymbol{\theta}|\boldsymbol{y}) \propto p(\boldsymbol{\theta}) \int_Y p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})p(\boldsymbol{x}|\boldsymbol{\theta})d\boldsymbol{x}$$

It is assumed that we have an expression for the prior, $p(\boldsymbol{\theta})$, and are able to generate samples from $p(\boldsymbol{x}|\boldsymbol{\theta})$. If $p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})$ is defined to the Dirac-function with $\boldsymbol{x} = \boldsymbol{y}$ we get:

$$p(\boldsymbol{\theta}|\boldsymbol{y}) \propto p(\boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{\theta})$$

If we were able to draw samples from the likelihood function such that $\boldsymbol{x} = \boldsymbol{y}$, we could estimate the posterior probability. A simple algorithm for this is shown in 1. The algorithm will produce samples from the posterior, but there is a practical problem in the comparison $\boldsymbol{x} = \boldsymbol{y}$. For most continuous probability distributions, the probability that two samples are exactly the same. For discrete probability distributions the situation will also be bad in practice. So in practice the algorithm in 1 is not useful, since most samples will be discarded, and it will take a very long time to generate samples.

To make the algorithm more practically useful a couple of relaxations are made. The requirement that $\boldsymbol{x} = \boldsymbol{y}$ is relaxed, and instead samples that lie within some defined distance of $\boldsymbol{y}$ are accepeted as well: $d(\boldsymbol{x}, \boldsymbol{y}) \leq \rho$.

If the dimension of Y is large, it may be difficult to deal with $\boldsymbol{x}$ and $\boldsymbol{y}$ directly, in which case an alternative may be to look at some statistics, $S(\cdot)$, of the data instead.

Looking at an example from [2]

Examples

References

[1] A. Gelman, J.B. Carlin, H.S. Stern, D.B. Dunson, A. Vehtari, and D.B. Rubin. *Bayesian Data Analysis, Third Edition*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, 2013.

[2] Jean-Michel Marin, Pierre Pudlo, Christian P. Robert, and Robin J. Ryder. Approximate bayesian computational methods. *Statistics and Computing*, 22(6):1167–1180, 2012.