Model identifiability in complex systems

XII GEFENOL Summer School on Statistical Physics of Complex Systems

Mario Castro

Madrid, July 1-12, 2024

Session 1: Into the Bayesian-verse

Summary of this session:

Instead of Data...Observed variable
Instead of Parameter...Unobserved variable
Instead of Posterior...Conditional probability (unobserved | observed)
Instead of Prior...Probability of unobserved variables
Instead of Likelihood...Probability of observed variables

Problem 1.1: Calibrate your priors

Fill the **lower** and **upper** bounds in the following table, and think about how to model this using normal, beta or uniform priors. Quantify your own uncertainty in the last column. That will also help you to define the prior.

#	Question	Lower	Upper	Confidence (%)
1	Engine Car speed record (km/h) in 1955			
2	Time (total mission time) it took Apollo 11 from Earth to Moon			
3	Length of a typical credit card (cm)			
4	How many parsecs tool Millenium falcon to do the Kessel Run			
5	Average time until HIV mutates			
6	Air distance between Madrid and Kiev (km)			
7	Percentage of a square covered by a circle of the same width			
8	How many cells in the body			
9	Days for the Moon to orbit Earth			
10	Date Star Wars first aired			
11	How many bacteria in our skin			
12	Average number of goals scored per match in La Liga			
13	Length of longest match played by R. Nadal			
14	Budget of European Union			
15	Number of deaths by Covid-19 during the 2020 first wave			

Problem 1.2: The hidden die

I throw a 6-sided die privately and I keep the result.

1. Now, I throw it again and I don't tell you the outcome, just that the outcome was greater than the first one. What is the posterior distribution for the outcome of the first throw? **Note:** Use a generative model and the Approximate Bayesian Computation (ABC) approach.

2. Now, I throw the die 6 more times, so the whole sequence of outcomes is: {greater, greater, equal, lower, greater, equal, equal} What is now the posterior of the first throw?

3. Imagine that I offer you 100€ if you correctly guess the first throw. How much money would you risk to accept the bet?

This problem is attributed to Thomas Bayes himself.

Problem 1.3: The prison window

Two men looked through prison bars; one saw stars, the other tried to infer where the window frame was. From their viewpoint, they look through a window and see stars at locations $\{(x_n, y_n)\}$. They can't see the window edges because it is totally dark apart from the stars. Assuming the window is rectangular and that the visible stars' locations are independently randomly distributed, what are the inferred values of $(x_{\min}, y_{\min}, x_{\max}, y_{\max})$?

Problem 1.4: The radioactive nucleus and the emitter

We have a radioactive source ...which is emitting particles of some sort. There is a rate p, in particles per second, at which a radioactive nucleus sends particles through our counter; and each particle passing through produces counts at the rate θ (note that the counter is not perfect, so not every fraction of the received particles produce secondary emission). From measuring the number c_1, c_2,...of counts in different seconds,

1. If θ is 10% and the counter registers $c = 15$ particles in a one second interval, what is the posterior distribution of n, the actual number of particles that hit the counter, and r, the average rate particles are emitted?

   Note: Use a generative model and the Approximate Bayesian Computation (ABC) approach. There are many ways to generate random emissions from the nuclei. Use a Poisson distribution.

2. What can we say about the rates p and θ?

Problem 1.5: The football model

Let p_{ij} the probability per minute that team i scores a goal against team j in a football match. We want to infer that probability from a record of football matches. Let us study different models. In all cases $p_{ij} \neq p_{ji}$. **Note:** Again, use ABC.

1. In the first model, we have only two matches: R. Madrid-Barcelona (2-1) and Barcelona-R. Madrid (3-0). We know that being the host of the match gives you a 10% increased probability (define $\gamma = 1.1$) of scoring a goal. Compute the posterior distribution for p. Test a couple of priors.

2. The problem with that model is that is does not help us learn from one match to another. So, in the second model, we want to infer the *intrinsic offensive ability* of each team, α_k, and the intrinsic defensive ability, β_k, so we can write

   $$p_{ij} = \alpha_i \beta_j$$

   Again, consider the same home effect, $\gamma = 1.1$. Use the same *dataset* with just two matches.

3. Finally, we want to learn the home effect parameter, γ. Try different priors for this parameter as well.
4. Now you can try your model using a large dataset of results from a season of La Liga. Using your model, compute the posterior distribution for the ranking of each team in that season.
5. **Exploting the posterior:** Simulate a new scoring system where teams scoring 4 or more goals get an extra point for that match.
6. **Exploting the posterior:** What's the probability that a team do not receive any goal in N matches in a row?

Session 2: Maths and Probabilistic programming

Summary of this session:

- Bayes rule

$$P(\text{Unobserved} \mid \text{Observed}, H) = \frac{P(\text{Observed} \mid \text{Unobserved}, H)P(\text{Unobserved}, H)}{\int P(\text{Observed} \mid \text{Unobserved}, H)P(\text{Unobserved}, H)d\text{Unobserved}}$$

- For n independent observations, $\{x_k\}$,

$$P(\text{Observed}, H) = \prod_{k=1}^{n} P(x_k|\text{Unobserved}, H)$$

- For instance, for a normal distribution with same variance for all observations

$$P(\text{Observed}, H) \sim e^{-\frac{1}{\sigma^2}\sum_{k=1}^{n}(x_k-\mu)^2}$$

- Probabilistic languages perform a stochastic integration of the denominator of Bayes rule. Typically they are expressive and compact. For example, a bayesian linear regression in `stan` will be simply

```
data {
  int<lower=0> N; // number of data points
  vector[N] x; // predictor variable
  vector[N] y; // response variable
}

parameters {
  real beta0; // intercept
  real beta1; // slope
  real<lower=0> sigma; // standard deviation (always positive)
}

model {
  // Priors (very, very vague....)
  beta0 ~ normal(0, 10);
  beta1 ~ normal(0, 10);
  sigma ~ normal(0, 10);

  // Likelihood (vectorized)
  y ~ normal(alpha + beta * x, sigma);
}
```

Listing 1: Bayesian Linear Regression in Stan

Problem 2.1: The (extended) football model

1. Re-do the problem with the intrinsic offensive/defensive abilities and the *home effect* using Stan, exclusively.
2. Try another model where the scoring probability per minute, p_{ij} is a function of time, of the form

$$p_{ij} = \frac{1}{1 + e^{\beta_0 + \beta_1 t}}.$$

Imagine that you have a dataset of the form (`-1` means that the final score was 0-0)

```
Match,Team_A,Team_B,Scoring_team,Time_minute
1,1,2,2,35
1,1,2,2,37
1,1,2,1,87
1,1,7,-1,-1
1,2,7,2,21
......
```

Write an Stan code to fit the whole football model and use the information from this dataset. Analyze the posterior of β_1 and discuss the practical implications for football.