MSA101/MVE187 2020 Lecture 13.1
Bayesian modeling

Petter Mostad

Chalmers University

October 6, 2020

Bayesian modelling

- When using Bayesian statistics, you generally start with:
  - A context.
  - A question or goal within that context.
  - A dataset.
- There are then generally two steps:
  1. Going from the above to **a set of variables consisting of variables you want to predict, variables representing the data, and often additional variables** and **a joint probability model for these variables**.
  2. Doing the computations to make the probabilistic predictions.
- So far this course has focused on the second step above. In this lecture, we focus on the first step.
- The lecture is a bit of a "cookbook", but we also look at some mathematical tools.

Bayesian modelling

The modelling step can often be divided into

1. Understanding the context and the question, and *exploring the data* (e.g., visualizing it).
2. Based on the context, decide on the set of variables you want in your model, and an *independence structure* between them.
3. Based on the context, come up with one or a handful of possible stochastic models relating the variables.
4. If necessary and if possible: Compare the possible candidate models by looking at the probability of observing the actual data under each model. Decide on one or a combination of models. **We return to this question in Lecture14.1.**
5. After computations for your chosen model has been made, do *posterior predictive checks*: If there are unreasonable results, go back to step 3.

Simple example: Braking distance of cars

Go back to "braking distance of cars" example used earlier. The steps become:

1. Question? (Breaking distance at some given speed?) Plotting!
2.
   - We decide on denoting the observed speeds as $x_1, \ldots, x_n$, denoting the observed breaking distance as $y_1, \ldots, y_n$, variables for prediction x_{new} and y_{new}, and a parameter (vector) θ.
   - We assume that

$$\begin{aligned} & \pi(y_1, \ldots, y_n, y_{new}, \theta \mid x_1, \ldots, x_n, x_{new}) \\ = \quad & \pi(\theta)\pi(y_{new} \mid x_{new}, \theta) \prod_{i=1}^{n} \pi(y_i \mid x_i, \theta) \end{aligned}$$

- Note: A lot of assumptions are made already here! We also assume that all the density functions $\pi(y_i \mid x_i, \theta)$ are the same.
- Causality can be a good *guide* for setting up model assumptions!

Simple example continued: Braking distance of cars

3. Some possible likelihood models are

$$y_i \mid x_i, \theta_1, \theta_2, \theta_3 \sim \text{Normal}(\theta_2 + \theta_3 x_i, \theta_1^{-1})$$

or

$$y_i \mid x_i, \theta_1, \theta_2, \theta_3, \theta_4 \sim \text{Normal}(\theta_2 + \theta_3 x_i + \theta_4 x_i^2, \theta_1^{-1}).$$

For the prior $\pi(\theta)$ some possibilities are

$$\pi(\theta_1, \theta_2, \dots) \propto 1 \qquad \text{(flat prior)}$$

or

$$\pi(\theta_1, \theta_2, \dots) \propto 1/\theta_1 \qquad \text{(uninformative prior)}$$

or an *informative prior* based on previous observations about braking distances for cars.

4. With our particular question and a fair amount of data: Why not use the most general model and the uninformative prior. **We return to discussing priors in part 3 of this lecture**.
5. In this simple model, it should be enough to check that, e.g., predictions at reasonable speeds are not whacky (i.e., that the data contains enough info to inform the distribution of θ).

Help in step 5: Posterior model checking

- The posterior model (conditioned on data) can be used to make many types of predictions, not only of the variables of interest.
- If some such predictions seem very unreasonable, then maybe a better model should be found.
- Toy example:
  - Data, 4.33, 4.32, 4.35, 4.30.
  - Model: $y_i \sim \text{Normal}(\mu, \sigma^2)$.
  - If the prior is $\mu \sim \text{Normal}(0, 100)$, $\sigma^2 = 1$, simulations from the posterior predictive will have too much spread in the data.
  - If the prior is $\mu = 0$, $\pi(\sigma^2) \propto 1/\sigma^2$, simulations from the posterior predictive will have both wrong mean and wrong spread.
- Generally: The prior may indicate that some "features" of the model can be "informed" by the data, while other "features" are fixed. Are there "features" that are fixed that need to be informed by the data? This can be investigated by comparing simulations from the posterior predictive with the actual data. Are there systematic differences?

Model checking with cross validation

- Assume you have observations $y_1, y_2, \ldots, y_n$ that are *independent* given the parameter θ.
- Then for $i = 1, \ldots, n$:
  - Remove y_i and find the posterior for θ using only the remaining y_j.
  - Compare the value of y_i with the y_i^* predicted using this posterior for θ.
- Are there any systematic differences between y_i and y_i^* when $i = 1, \ldots, n$?
- An example in part 2 of this lecture.

Model checking with posterior predictive p-values

Assume a model $\pi(y, \theta)$ for data y and parameters θ has been defined.

- We also assume a *discrepancy function* $D(x, \theta)$ is given, measuring a way in which data x is "unlikely" to be the result when θ is the parameter.
- Simulate a sample $\theta_1, \ldots, \theta_N$ from the posterior $\theta \mid y$, and simulate corresponding data $x_1, \ldots, x_N$ using the likelihood $\pi(x \mid \theta)$.
- The proportion of i's for which $D(x_i, \theta_i) \geq D(y, \theta_i)$ is an estimate for the posterior predictive p-value.
- If this value is "very small" and your intention had been to use a prior without much information, "reject" the model.

MSA101/MVE187 2020 Lecture 13.2
Example: Heart transplant mortality

Petter Mostad

Chalmers University

October 8, 2020

Heart transplant example (Sections 7.3 - 7.10 of Albert): Context, question, and model structure

- Objective: Learn about differences and similarities in success rates for heart transplants at 94 US hospitals.
- Concrete questions: Predict future mortalities at specific hospitals. And: When can we reliably say that one hospital has a lower mortality rate than another?
- Data (using some fixed time period): For $i = 1, \ldots, 94$:
  - y_i: The number of deaths within 30 days of receiving a heart transplant at hospital i.
  - e_i: A number ("exposure") measuring the number of heart transplants performed, *and* how sick the patients were, at hospital i.
- Other variables: $\lambda_1, \ldots, \lambda_{94}$ representing the mortality rate of each hospital, and a parameter (vector) θ.
- Structure of model: For $i = 1, \ldots, 94$:

$$y_i \mid \lambda_i, e_i \sim \text{Poisson}(e_i \lambda_i) \quad \text{with (possibly)} \quad \pi(\lambda_i \mid \theta)$$

and a prior $\pi(\theta)$.

Step 3: Some possible models

1. All λ_i are independent. (No θ in the model)
2. All λ_i are equal. (No θ in the model).
3. A hierarchical model: All λ_i are from the same density $\pi(\lambda \mid \theta)$, and there is a prior on θ.

The actual densities are discussed below.

We will go through each of these models, and use posterior predictive checks to select our model.

Model 1: All λ_i are independent

- If we don't have much other information about this type of data, we would want to use non-informative priors $\pi(\lambda_i) \propto 1/\lambda_i$.
- However, for the hospitals where $y_i = 0$, we then also get improper posteriors for λ_i.
- So we would need to use a proper prior, but what if we have no other data than this?
- Using fixed estimates $\hat{\lambda}_i = y_i/e_i$ does not help.

Model 2: All λ_i are equal

We write $\lambda_i = \lambda$.

- We can now easily use the non-informative prior $\pi(\lambda) \propto 1/\lambda$.
- Conjugate computations show that

$$\lambda \mid y_1, \ldots, y_{94} \sim \text{Gamma}\,(d, f)\,.$$

  where $d = \sum_{i=1}^{94} y_i$ and $f = \sum_{i=1}^{94} e_i$.
- Note: When $y_{new} \sim \text{Poisson}(e_{new}\lambda)$ and $\lambda \sim \text{Gamma}(\alpha, \beta)$, we get the predictive distribution

$$y_{new} \sim \text{Neg-Biomial}\left(\alpha, \frac{\beta}{\beta + e_{new}}\right)$$

  where we we write $x \sim \text{Neg-Binomial}(\alpha, p)$ for the density

$$\pi(x \mid \alpha, p) = \frac{\Gamma(\alpha + x)}{\Gamma(\alpha)x!} p^{\alpha}(1-p)^{x}$$

- We do a posterior predictive check using cross-validation.

Model 2: Posterior predictive check using cross-validation

- We get $\lambda \mid y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_{94} \sim \text{Gamma}(d - y_i, f - e_i)$.
- With $y_i^* \sim \text{Poisson}(e_i \lambda)$ and $\lambda \sim \text{Gamma}(d - y_i, f - e_i)$, the predictive density becomes

$$y_i^* \mid y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_{94} \sim \text{Neg-Binomial}\,(d - y_i, e_i/f)$$

- For each i we can now compute the quantile of y_i in the distribution for y_i^*.
- As the y_i are discrete counts, we compute the adjusted quantiles

$$\sum_{j=0}^{y_i - 1} \text{Neg-Binomial}(j; d - y_i, e_i/f) + \frac{1}{2} \text{Neg-Binomial}(y_i; d - y_i, e_i/f).$$

- If the model fits perfectly, the computed quantiles will be (approximately) uniformly distributed.
- R computations show that the y_i are more spread out than they should be according to the posterior predictive check, so we reject Model 2.

Model 3: A hierarchical model

- There is some similarity, but also some variation among the λ_i: We build a model where we infer from data the amount of similarity.
- We use the model

$$y_i \sim \text{Poisson}(\lambda_i e_i) \text{ and } \lambda_i \sim \text{Gamma}\left(\alpha, \frac{\alpha}{\mu}\right),$$

$$\pi(\alpha) \propto \frac{1}{\alpha} \text{ and } \pi(\mu) \propto_\mu \frac{1}{\mu}$$

- Note: With this parametrization, the expectation of the Gamma distribution is μ and its standard deviation is $\mu/\sqrt{\alpha}$, so these parameters can be more easily interpreted than standard ones.
- The posterior distribution on α will tell us how similar the λ_i actually are.

Computations for model 3

- The model above has $94 + 2$ unknown variables. For more easy computation, note that the distribution of $y_1, \ldots, y_{94}$, α, and μ is equivalent in the following marginalized model:

$$y_i \sim \text{Neg-Binomial}\left(\alpha, \frac{\alpha}{\alpha + \mu e_i}\right), \ \pi(\alpha) \propto_\alpha \frac{1}{\alpha} \text{ and } \pi(\mu) \propto_\mu \frac{1}{\mu}$$

- As we now only have 2 unknown variables, we can do inference for μ and α for example with discretization (see below).
- If we then want the posterior density for some particular λ_j, note that

$$\lambda_j \mid \alpha, \mu, \text{data} \sim \text{Gamma}\left(\alpha + y_j, \frac{\alpha}{\mu} + e_j\right).$$

- See computations in R for how we can now answer questions such as
  - What is the probability of no deaths in hospital 24 given a new exposure of 1000?
  - What is the probability that hospital 90 is really better than hospital 9, i.e., that $\lambda_{90} < \lambda_9$?

Computations in model 3

- For the posterior $\pi(\alpha, \mu \mid \text{data})$

$$\pi(\alpha, \mu \mid \text{data}) \propto_{\alpha,\mu} \frac{1}{\alpha\mu} \prod_{i=1}^{94} \text{Neg-Binomial}\left(y_i; \alpha, \frac{\alpha}{\alpha + \mu e_i}\right)$$

$$\propto_{\alpha,\mu} \frac{1}{\alpha\mu} \prod_{i=1}^{94} \frac{\Gamma(y_i + \alpha)}{\Gamma(\alpha)} \left(\frac{\alpha}{\alpha + \mu e_i}\right)^{\alpha} \left(\frac{\mu e_i}{\alpha + \mu e_i}\right)^{y_i}.$$

- To make the posterior more symmetrical (this improves numerical properties), we do the *reparametrization* $\theta_1 = \log(\alpha)$ and $\theta_2 = \log(\mu)$, i.e., $\alpha = e^{\theta_1}$ and $\mu = e^{\theta_2}$.

Reparametrizations

- Expressing the posterior density in terms of other parameters (reparametrization) can be a good idea, in particular if it makes the posterior more similar to a multivariatel normal (or at least more symmetric) and less correlated.
- If you set $\alpha = f(\theta)$ where α is the old and θ is the new parameter, and these are continuous parameters, you must also multiply with $|f'(\theta)|$ (so the integral is unchanged).
- If α and θ are multivariate parameters, then $|f'(\theta)|$ is the determinant of the Jacobian matrix.
- Example: If you substitute $\alpha = e^{\theta_1}$, then you must also multiply with the extra factor e^{θ_1} as this is the derivative of the reparametrization function.
- For "scale" parameters, a reparametrization like $\theta_1 = \log(\alpha)$ (i.e., moving to a log scale) is often a good idea.
- Reparametrization can improve numerical accuracy. An example is improved convergence for Gibbs sampling, if reparametrization makes the parameters less correlated.

More on computations for model 3

- With the reparametrization above, the posterior for $\pi(\theta_1, \theta_2 \mid \text{data})$ is proportional to

$$\prod_{i=1}^{94} \frac{\Gamma(y_i + e^{\theta_1})}{\Gamma(e^{\theta_1})} \left(\frac{e^{\theta_1}}{e^{\theta_1} + e^{\theta_2} e_i} \right)^{e^{\theta_1}} \left(\frac{e^{\theta_2} e_i}{e^{\theta_1} + e^{\theta_2} e_i} \right)^{y_i}.$$

- Remember that for numerical reasons we prefer to compute the logged posterior:

$$\begin{aligned} L(\theta_1, \theta_2) \quad = \quad & \sum_{1=1}^{94} \log(\Gamma(y_i + e^{\theta_1})) - \log(\Gamma(e^{\theta_1})) + \theta_1 e^{\theta_1} + y\theta_2 \log(e_i) \\ & - y_i e^{\theta_1} \log(e^{\theta_1} + e^{\theta_2} e_i) \end{aligned}$$

- When discretizing this in 2D, it's good to know for approximately what values you expect it to be large: Note that $\sum_i y_i / \sum_i e_i \approx 0.001$ so $\mu \approx 0.001$ and $\theta_2 \approx -7$. Furthermore, from our interpretation of α, we see that it is probably greater than 1, so $\theta_1 > 0$. See R computations.

MSA101/MVE187 2020 Lecture 13.3
Choosing priors

Petter Mostad

Chalmers University

October 8, 2020

Using improper priors

- It is quite useful to use improper priors: Completely OK as long as the posterior becomes proper.
- Proving that the posterior is proper may be difficult and may unfortunately be forgotten about.
- The output of a Metropolis-Hastings or Gibbs algorithm applied to an improper distribution will often look like some kind of random walk. However, it may not be direcly obvious to spot the problem from the output!
- Examples 7.18, 7.19 in RC

Modelling priors

For the prior, choose between

- Using an *uninformative* prior.
- Compute a prior based on other, previous data.
- *Elicit* a subjective prior from experts. (Example: Use `beta.select` in LearnBayes package).

Uninformative priors

- We want an "uninformative prior" on a parameter θ to represent "no knowledge". Unfortutately, it is not mathematically clear how this should be best defined.
- Such uninformative priors may or may not be *improper*.
- When using an improper prior, make sure your *posterior* is proper!!
- We have often used "flat" priors; however, a flat prior may not stay flat if θ is re-parametrized.
- If μ is a "location" parameter, you might use $\pi(\mu) \propto 1$; if λ is a "scale" parameter, a good alternative may be $\pi(\lambda) \propto 1/\lambda$.
- A number of theories have been developed. Some aim for maximized entropy.

Robustness

- Another approach to the choice of prior: Check if switching between different choices matters for the final result.
- NOTE: For any posterior, there exists a prior that will give this posterior (assuming nonzero densities).
- Revised question: Do *reasonable* changes in the prior affect the result much?
- If not, the prior is called *robust* for this likelihood.
- Example: See Albert 3.4

Comparing the *prior* predictive with context knowledge

- The prior model should represent "prior knowedge": A way to check that it does this correctly is to simulate new data from the prior predictive and check if they look like what you expect a priori.
- Examples
  - Simulate from the prior of a stochastic model for tree growth.
  - Simulate from the prior of a stochastic model for geological faults.
  - Simulate from the prior of a stochastic model for image noise.
- Example: If one believes some unobserved quantities should follow some distribution, one may compute or simulate their quantiles in this distribution: They should then be uniformly distributed. (Example: Prior predictive p-values).