Applied Bayesian Analysis and Numerical Methods

- The principles of Bayesian Analysis
- How the posterior distribution arises as a result of Bayes' rule
- Posterior inference
- Computation of the posterior distribution (analytically or more often numerically)
  - Asymptotic approximations
  - MCMC
- Relation to non-Bayesian methods

aug 25-12:12

What is Bayesian Inference

- Conclusions about unobservable parameters (or observable, but unobserved data) are made in terms of probability statements conditional on observed data: $p(\theta|y)$ (or $p(\tilde{y}|y)$)
  (condition implicitly on known values of any covariate information)
- This is a fundamental difference from classical (or 'frequentist') statistical inference
- Can make statements like "it is 95% probability that the unknown mean is in this interval". In contrast, a 95% confidence interval from classical inference means that in 95% of repeted experiments, the resulting confidence interval would cover the true, unknown mean.
- However, in many situations, classical and Bayesian inference give (for all practical purposes) similar conclusions.

aug 25-12:26

- One advantage of the Bayesian approach is that it easily allows building complex models with many parameters and layers, which is useful for modeling complex phenomena.
  - ↳ Over the last ~ 40 years, advances in computer power and algorithms have made such models possible to handle

---

Bayes theorem is fundamental

$$P(A \mid B) = \frac{P(A,B)}{P(B)} = \frac{\overbrace{P(A)}^{\text{Prior}} \cdot \overbrace{P(B \mid A)}^{\text{Likelihood}}}{\underbrace{P(B)}_{\text{Marginal for } B}}$$

A: Unobserved quantity, B: Observed data

aug 25-12:38

Parameters, data, explanatory variables

- Two kinds of unobserved quantities for which we would like to do statistical (Bayesian) inference:
  1. Parameters: unobservable, governing the process/model assumed to give the observed data, for example means, variances etc.
     ↓ θ-vector
  2. Potentially observable, for example future observations, that we wish to make predictive inference
     $\tilde{y}$-vector

- y: Vector of observed data (sometimes a matrix)
  Also called "outcomes". Considered as observed values of random quantities
  Data are most often collected for n units, which may be for example animals, countries, patients etc. If the data is a vector: $y = (y_1, \ldots, y_n)^T$
  If the data is a matrix, we write $y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$ where y_i is a vector

aug 25-12:46

· Explanatory variables (or covariates) X: Quantities we observe for each unit that we do not consider to be random. X denotes the set of covariates for all units, and is a matrix with n rows and k columns, where k is the number of covariates measured for each unit.

If we would like to consider covariates as random, they can be moved into the y category.

Exchangeability

A key starting assumption in most statistical models, either explicitly or implicitly, is that we regard the n values y_i as exchangeable. This means that we assume that the joint probability density $p(y_1,\dots,y_n|\theta)$ is invariant to permutations of the indexes. A simple example is that we model the outcomes y_i as iid given the parameter vector θ. If we have covariate information, a more natural assumption is to model y_i, $i=1,\dots,n$ as exchangeable given θ AND x_i. Hence $p(y_1,\dots,y_n|\theta)$ is invariant to permutations of the pairs (x_i,y_i). Then, for two units

aug 25-12:58

with the same x-value, the distributions for the outcomes will be the same.

If information relevant to the outcome were conveyed in the unit indexes rather than by explanatory variables, an exchangeable model is not appropriate.

Notation on distributions

- $p(\cdot|\cdot)$: a conditional probability density with arguments determined by the context
- $p(\cdot)$: a marginal distribution with argument given by the context
- The same notation is used for continuous density functions and discrete probability mass functions
- Distributions for different variables will all be denoted by p (and not indexed by the variable), for a compact presentation

aug 25-13:32

- Sometimes, $Pr(\cdot)$ will be used to denote the probability of an event, for example $Pr(\theta > 2) = \int_{\theta>2} p(\theta)\, d\theta$

- For standard, known, distributions, notation based on the name of the distribution is used. For example, if θ has a normal distribution with mean μ and variance σ^2, we write

$$\left.\begin{array}{l} \theta \sim N(\mu, \sigma^2) \\ \text{or } p(\theta) = N(\theta \mid \mu, \sigma^2) \\ \text{or } p(\theta \mid \mu, \sigma^2) = N(\theta \mid \mu, \sigma^2) \end{array}\right\} \text{Equivalent}$$

Notation such as "$\theta \sim N(\mu, \sigma^2)$" means the random variable (r.v.) θ has this distribution, while "$N(\theta \mid \mu, \sigma^2)$" denotes the actual density function.

APPENDIX A!

aug 25-13:40

Bayesian Inference

- The basis of Bayesian inference is the probability statement called the posterior distribution, which we get from Bayes' rule:
  - posterior (seeing/analyzing/conditioning on observed data

$$p(\theta \mid y) = \frac{p(\theta, y)}{p(y)} = \frac{p(\theta) \cdot p(y \mid \theta)}{p(y)}$$

where $p(y) = \begin{cases} \sum_{\theta} p(\theta) \cdot p(y \mid \theta), & \theta \text{ discrete} \\ \int p(\theta) p(y \mid \theta)\, d\theta, & \theta \text{ continuous} \end{cases}$

aug 25-13:46

Two core steps for Bayesian inference:

1. Formulate the full model which expresses the joint probability distribution of θ and y, namely

$$p(\theta, y) = p(\theta) \cdot p(y|\theta)$$

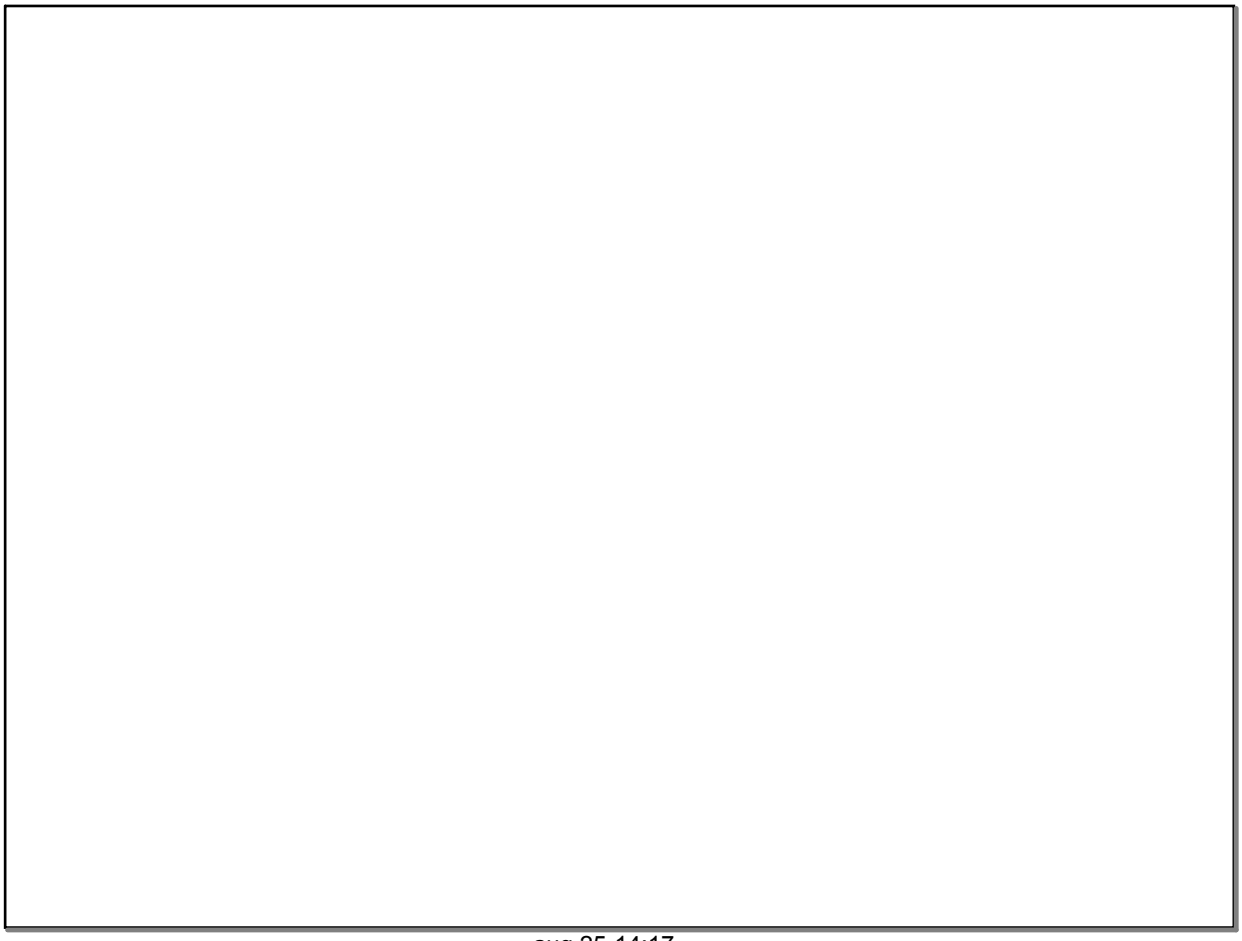
- $p(\theta)$ is called the prior distribution for θ, and should express our knowledge/beliefs regarding the distribution of θ prior to seeing/conditioning the observed data y.
- $p(y|\theta)$ is called the sampling (or data) distribution, and should reflect our knowledge and assumptions of the distribution of y given θ, before seeing the data.

For fixed, observed data, $p(y|\theta)$ is regarded as a function of θ and called the likelihood function.

aug 25-13:51

2. Perform computations, analytically or most frequently by approximative algorithms (such as MCMC), to summarize $p(\theta|y)$ in appropriate ways (e.g. the whole distribution, posterior intervals, some form of point estimates accompanied by an uncertainty statements).

aug 25-14:16



aug 25-14:17