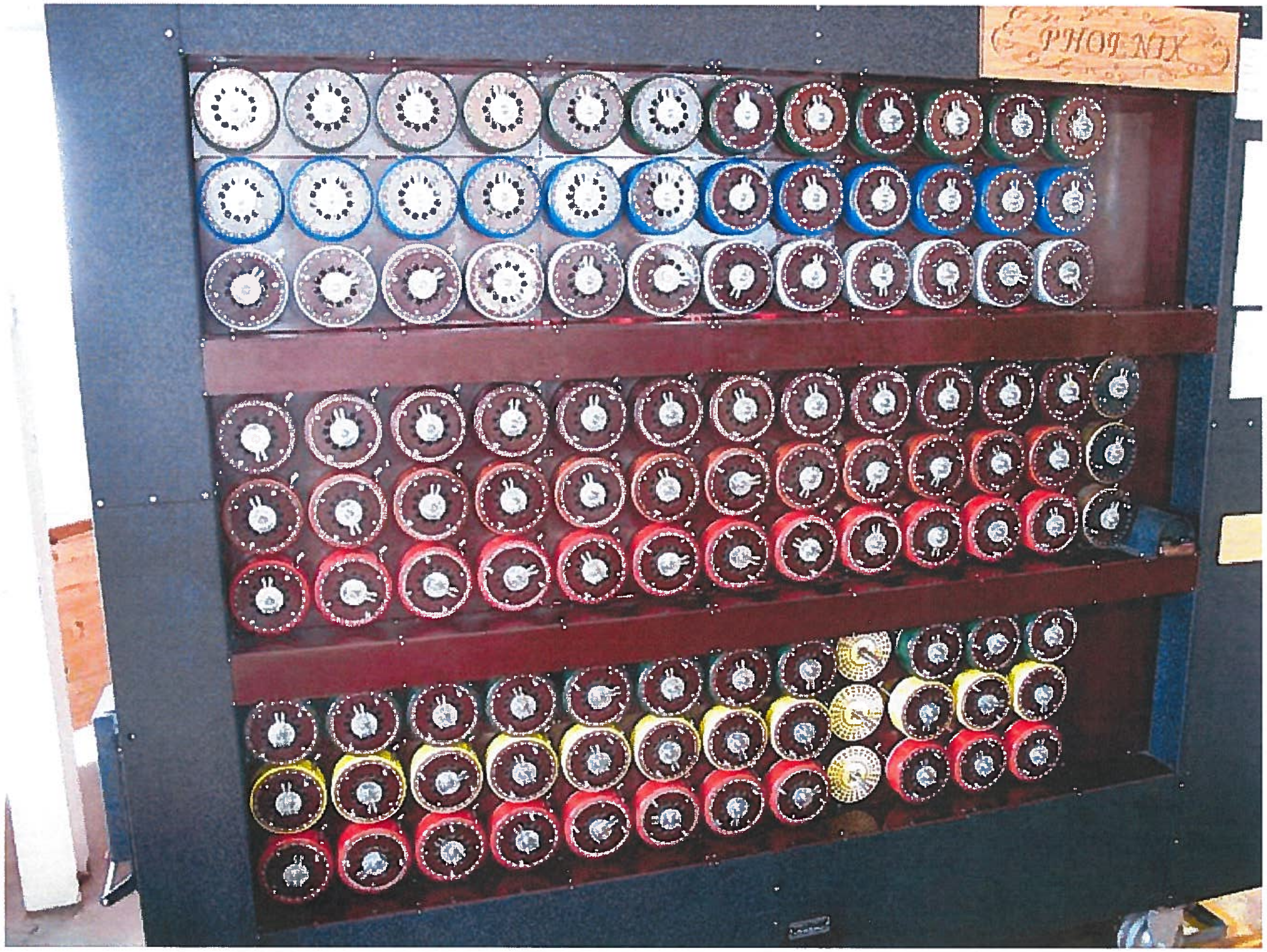
Introduction to Bayesian statistics

GMS-MT2018

Wellcome Centre for Human Genetics

Instructor: Dr Andre Python

UNIVERSITY OF OXFORD



A rebuilt replica of a 'bombe' machine used by cryptologists to crack the German enigma code, Wikimedia.

Picture of Enigma, E. Simpson, June 2010, Significance 2010

Bayesian inference is everywhere…

Science

Engineering

Philosophy

Medicine

Sport

Law

…

really everywhere…

Machine learning

Spam filters

Speech recognition

Bioinformatics

Economics

…

General schedule

Date: 8 and 9 November 2018 | Location: WCHG Room B | Time: 09:00-17:30

Detailed schedule

8 November

- Theory: 9:15-12:15 (15' break at 10h30)
- Practical/Theory: Introduction to R-INLA: 13:30-16:30 (15' break at 15h00)
- Additional questions (optional): 16:30-17:30

9 November

- Practical: Bayesian regression: 9:00-12:15 (15' break at 10h30)
- Practical: Personal work: 13:30-16:30 (15' break at 15h00)
- Summary/ further guidance/additional questions (optional): 16:30-17:30

Summary

THEORY (0.5 day)

1. Preliminaries
   1.1. Random variables
   1.2. Probability function for a random variable
2. Introduction to Bayesian statistics
   2.1. Comparison Bayesian vs frequentist statistics
   2.2. Bayes' Theorem
   2.3. Prior distributions
   2.4. Posterior summary
   2.5. Bayes' Theorem with multi-parameters
   2.6. Model selection
   2.7. Hypotheses testing

PRACTICAL (1.5 days)

3. Introduction to R-INLA (0.5 day)
   3.1 A brief overview of R-INLA
   3.2. Practical 1: Hubble's Law
   3.3. Practical 2: Theory of INLA
4. Bayesian regression (0.5 day)
   4.1. Practical 3: Bayesian regression
   4.2. Practical 4: Generalized linear models
   4.3. Practical 5: Spatial models
5. Personal work (0.5 day)
   5.1.a) Further work on examples in the course
   5.1.b) Personal work using new datasets
   5.2. Summary & further guidance

Main references

Wang, X., Yue Ryan, Y., Faraway, J. (2018). Bayesian Regression Modeling with INLA. New York: Chapman and Hall/CRC.

- Data: : https://github.com/julianfaraway/brinla

Blangiardo, M., & Cameletti, M. (2015). Spatial and spatio-temporal Bayesian models with R-INLA. John Wiley & Sons.

King R., Papathomas M., Thomas L. (2015). Course note MT4531. Bayesian Inference, University of St Andrews, School of Mathematics & Statistics.

THEORY

1. Preliminaries

1.1. Random variables

A random (or stochastic) variable is a variable whose possible values are numerical outcomes of a random (or stochastic) phenomenon. We usually denote random variables with capital letters: X (discrete) ,Y (continuous).

- Discrete random variable: takes only a countable number of distinct values
  - Ex: number of children in a family
- Continuous (or nondiscrete) random variable: takes an infinite number of possible values
  - Ex: temperature

Ex 2: toss a coin (head/tail)

Bernoulli distribution : $p(x) = p^x(1-p)^{1-x}$ with $x = \{0, 1\}$

Another way of writing it:

$$\begin{cases} p(X=0) = p^0(1-p)^1 = 1-p \\ p(X=1) = p^1(1-p)^0 = p \end{cases}$$

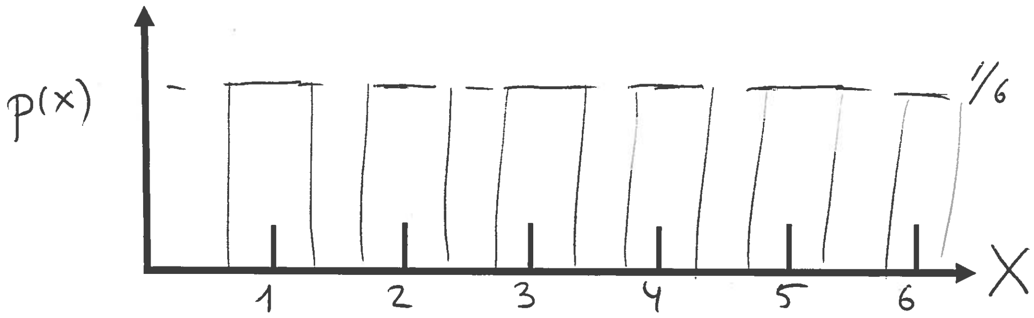
1.2 Probability function for a random variable

a) discrete: probability mass function (pmf): $p(x) = p(X = x)$

RV particular value

Ex 1: dice

pmf defined $\forall x \in$ range of X.



Properties:

$$0 \leq p(x) \leq 1$$

$$\sum_{X} p(x) = 1$$

b) continuous: probability density function (pdf): $f(y)$

$P(y_1 \leq Y \leq y_2) =$ ...... $\int_{y_1}^{y_2} f(y)\,dy$ ...... $\forall y \in$ range of Y ......

Properties:

$$f(y) \geq 0$$

$$\int_Y f(y)dy = 1$$

Discrete case

Mean of X (also called expected value of X)

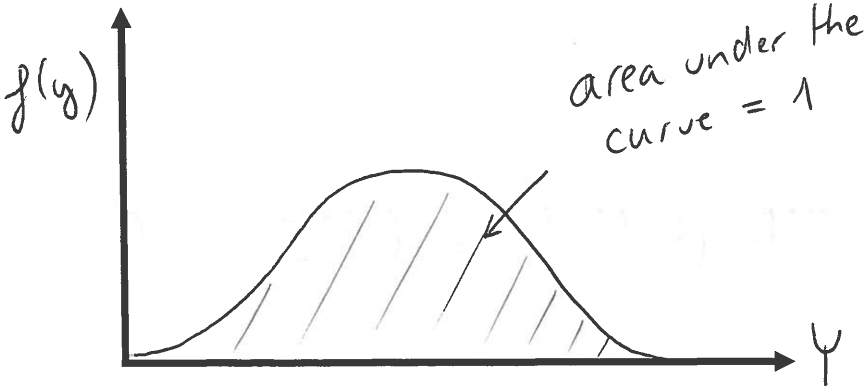
$$\mu = E(X) = \sum_{X} x\, p(x)$$

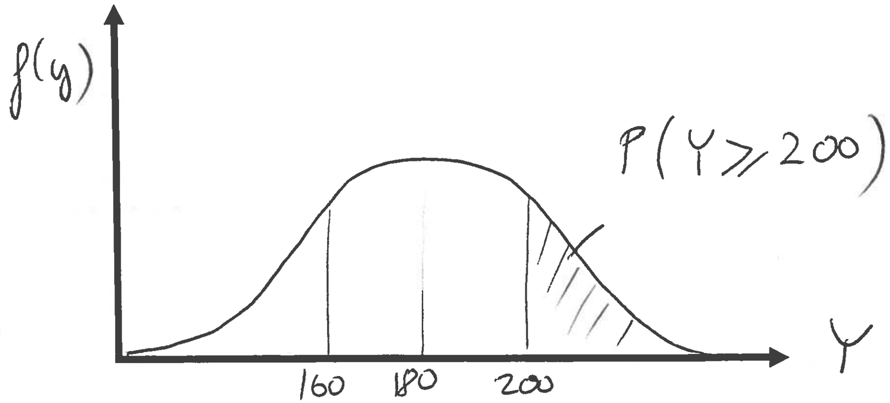
Variance of X:

$$\sigma^2 = V(X) = E[(X-\mu)^2] = \sum_{X} (x-\mu)^2 p(x)$$

Ex : Normal distribution

$$f(y|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$$





Continuous case

Mean of Y (also called expected value of Y)

$$\mu = E(Y) = \int_Y y\, f(y)\, dy$$

Variance of Y:

$$\sigma^2 = V(Y) = E[(Y-\mu)^2] = \int_Y (y-\mu)^2 f(y)\, dy$$

2. Introduction to Bayesian statistics

Statistical inference: process of analysing data to deduce properties of an underlying probability distribution. It is assumed that the dataset is sampled form a larger population.

2.1. Comparison Bayesian vs frequentist approaches

Bayesian statistics is an alternative complete theory of statistics different from the frequentist, also-called classical statistics.

Any statistical problem can be tackled by either theory.

- Probability theory: measure uncertainty in a coherent manner (e.g. probabilities should not contradict each other)
- Classical statistics: inference when looking at data, usually after considering a mathematical model. It uses probability theory and various procedures (confidence intervals, p-value, etc)
- Bayesian statistics: inference when looking at data after considering a mathematical model. Based solely on probability theory

Comparison table

Frequentist	Object	Bayesian
........NO..............	Prior information	........YES..........
........YES..............	Data	........YES..........
........YES..............	LSE, MLE, p-value	........NO..........
......fixed..........	Model parameter θ	.random variable...
likelihood....f(data|θ)	Inference	posterior...f(θ|data)

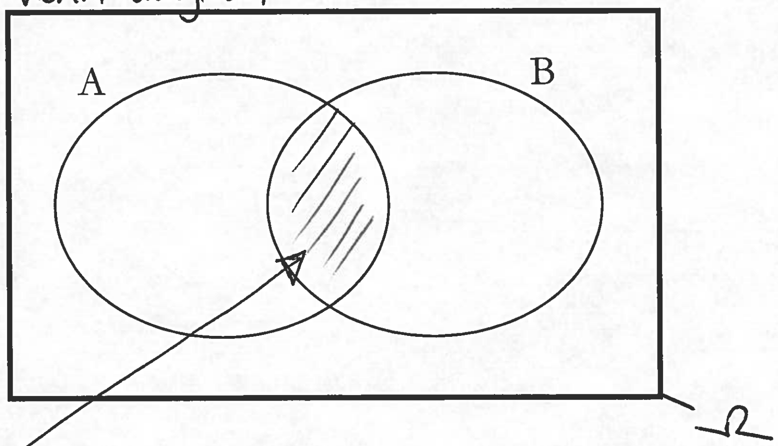
2.2. Bayes' Theorem

Thomas Bayes (1701-61)

Bayes' Theorem (BT) results of elementary probability theory

area ≡ probability

Venn diagram



a) Discrete case

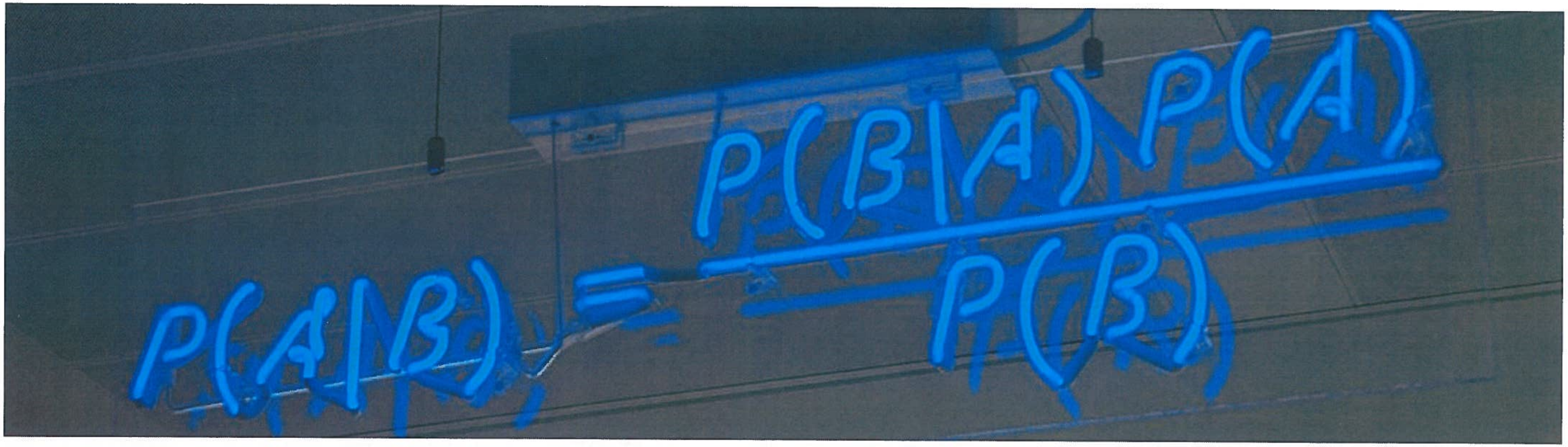
A, B: events with $P(B) > 0$

$P(\Omega) = 1$

BT: $P(A|B) = \dfrac{P(B|A)P(A)}{P(B)}$

Rewritten as:

$$\underbrace{P(A|B)\,P(B)}_{\text{Proportion of overlap area}} = \underbrace{P(B|A)\cdot P(A)}_{\text{Proportion of overlap area}}$$


P(A|B) = P(B|A)P(A) / P(B)

Exercise: testing malaria

A: an individual has malaria

B: the result of the test is positive

Known:

$P(B) = 0.1$ (prevalence)

$P(B|A) = 0.95$ (true positive: efficiency of the test if the individual tested has malaria)

$\rightarrow P(B^c|A)=$……0.05

$P(B^c|A^c) = 0.8$ (true negative: efficiency of the test if the individual tested hasn't malaria)

$\rightarrow P(B|A^c)=$……0.2

If $A_1, A_2, \ldots, A_n$ are mutually exclusive and exhaustive events, i.e. one of the event is certain to occur but two can't occur together:

→ Law of total probability: $P(B) = \sum_{i=1}^{n} P(B|A_i)P(A_i)$

BT: $P(A_j|B) = \dfrac{P(B|A_j)\,P(A_j)}{\sum_{i=1}^{n} P(B|A_i)\,P(A_i)}$ with A_j : specific event

Special case with a binary variable $\{A, A^c\}$

BT: $P(A|B) = \dfrac{P(B|A) \cdot P(A)}{P(B|A)\,P(A) + P(B|A^c)\,P(A^c)}$

Exercise: test malaria (10-15')

Question 1: P("false negative")

$$P(A|B^c) = \frac{P(B^c|A)\,P(A)}{P(B^c)} = \frac{P(B^c|A)\,P(A)}{P(B^c|A)\,P(A) + P(B^c|A^c)\,P(A^c)} = \frac{0.05 \times 0.1}{(0.05 \times 0.1) + (0.8 \times 0.9)} =$$

$\underline{\underline{0.69\,\%}}$

Question 2: P("false positive")

$$P(A^c|B) = \frac{P(B|A^c) \cdot P(A^c)}{P(B)} = \frac{P(B|A^c) \cdot P(A^c)}{P(B|A)\,P(A) + P(B|A^c)\,P(A^c)} = \frac{0.2 \times 0.9}{(0.95 \times 0.1) + (0.2 \times 0.9)}$$

$= \underline{\underline{65.45\,\%}}$

Break 10 minutes

a) Continuous case

We want to make inference on a parameter $\theta \in \Theta$, with observed data $\boldsymbol{y} = \{y_1, y_2, \dots y_n\}$, from some known probability distribution $f(\boldsymbol{y}|\theta)$, function of the parameter θ.

prior : initial belief prior any data being observed

BT: $\pi(\theta|\boldsymbol{y}) = \dfrac{f(\boldsymbol{y}|\theta)\, p(\theta)}{f(\boldsymbol{y})}$

$= \int_{\Theta} f(\boldsymbol{y}|\theta)\, p(\theta)\, d\theta$

posterior : update of the belief

likelihood : information contained in the data on θ

makes integral of $\pi(\theta|\boldsymbol{y}) = 1$

$f^{-1}(\boldsymbol{y})$: normalising constant

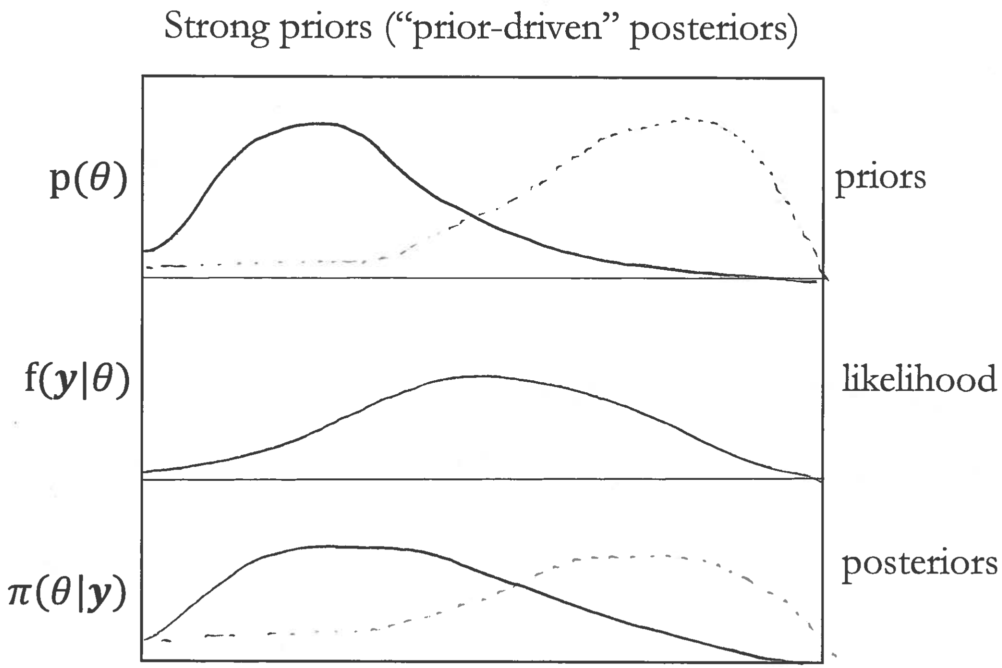
BT : $\pi(\theta|\boldsymbol{y}) \propto f(\boldsymbol{y}|\theta)\, p(\theta)$

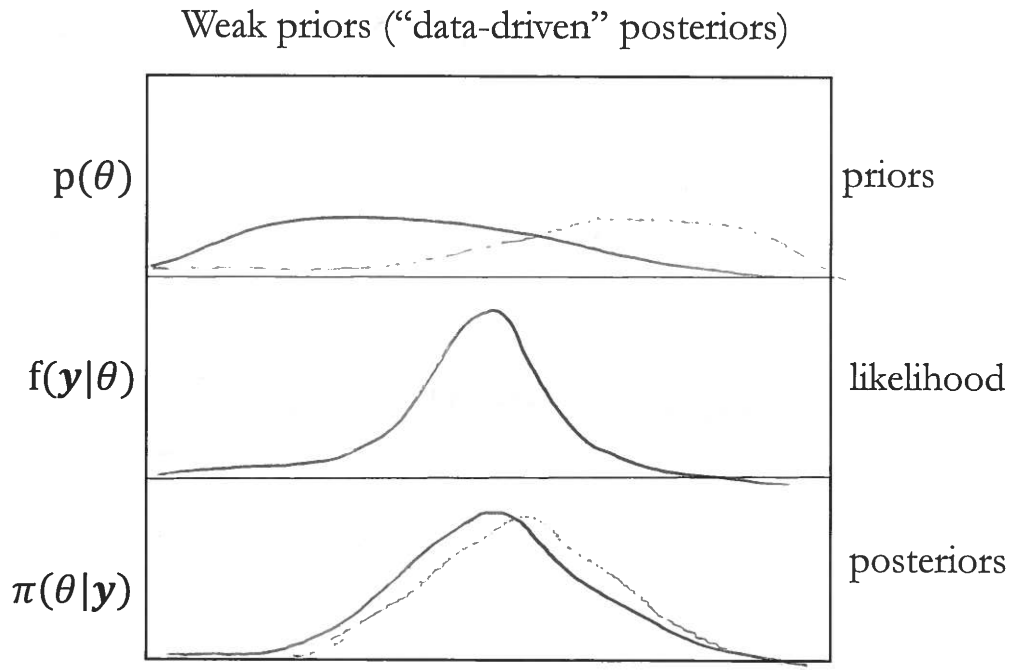
posterior $\propto$ likelihood $\times$ prior

2.3. Prior distributions

-specification of a prior on the unknown parameter $p(\theta)$, before observing the data is controversial but BT is not controversial

-there is no "correct choice" of $p(\theta)$ for a given problem





Without any specific prior information on a parameter θ , we often use non-informative/vague priors.

One example is the Uniform prior with $p(\theta) = c$, with $c \in \mathbb{R}$ (! transformation)

$\rightarrow \pi(\theta|\boldsymbol{y}) = f(y|\theta) \cdot c \propto f(y|\theta)$

$\rightarrow \pi(\theta|\boldsymbol{y}) \propto f(y|\theta)$

The shape of the posterior = the shape of the likelihood function

Note that it is common practice to use different priors and check the results on the posterior $\pi(\theta|\boldsymbol{y})$. This refers to prior sensitivity analysis.

2.4. Posterior summary

All information on the parameter θ is included in the posterior distribution $\pi(\theta|\boldsymbol{y})$ but we often want to provide some summary (posterior means, variances, quantiles, etc. e.g. reports read by a general audience).

Mean: $E_\pi(\theta) = \int_\Theta \theta \pi(\theta|\mathbf{y}) d\theta$

Credible intervals (CI): $\int_a^b \pi(\theta|\mathbf{y}) d\theta = 1 - \alpha$ with $0 \leq \alpha \leq 1$

If $\alpha = 0.05 \rightarrow$ 95%CI contains 95% of the posterior distribution of θ.

Confidence intervals (classical): repeated data collection → long-run 5% of confidence intervals do not contain the fixed unknown parameter.

2.5. Bayes' Theorem: multivariate

BT can be easily generalised to multi-parameters cases. We want to make inference on a set of parameters $\boldsymbol{\theta}$ (bold) and observed data $\boldsymbol{y}$.

(joint) posterior distribution $\pi(\boldsymbol{\theta}|\boldsymbol{y}) = \dots \dfrac{f(\boldsymbol{y}|\boldsymbol{\theta})\, p(\boldsymbol{\theta})}{f(\boldsymbol{y})} \dots$

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

marginal posterior distribution $\pi(\theta_1|\boldsymbol{y}) = \dots \int \pi(\boldsymbol{\theta}|\boldsymbol{y})\, d\theta_2, \dots, d\theta_n \dots$

Integration is usually too complex in practice and require MCMC or other approaches (e.g. INLA) to get summaries of posterior distributions.

2.6. Model selection

How to choose the most suitable model for a given dataset?

Non-Bayesian: one often uses the Akaike information criterion (AIC)

MLE estimation

AIC = ... $-2\log p(y \mid \hat{\theta}_{MLE}) + 2k$ ... number of parameters

It is problematic to count the number of parameters e.g. hierarchical models.

Bayesian: deviance information criterion (DIC)

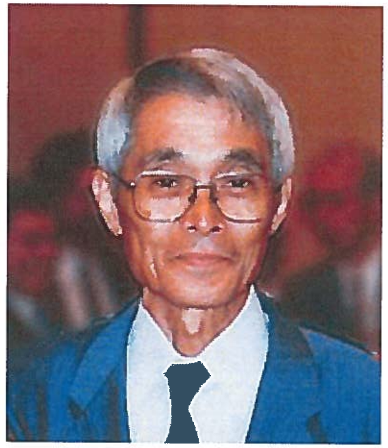
Deviance $D(\theta)$= ... $-2\log p(y \mid \theta)$ ...

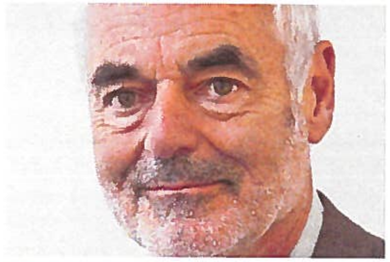
Effective number of parameters P_D

= ... $E[D(\theta)] - D(E(\theta)) = \bar{D} - D(\bar{\theta})$

DIC = ... $\bar{D} + P_D$ ...

Other approaches: Watanabe information criterion (WAIC), etc.





Rule of thumb to interpret the Bayesian factor (BF) (Kass & Raftery, 1995)

BF	Interpretation
<3	no evidence of H_0 over H_1
>3	positive evidence for H_0
>20	strong evidence for H_0
>100	very strong evidence for H_0

2.7. Hypothesis testing

Classical: often, the null hypothesis H_0 is a single point, and the alternative, H_1 represents everything else. The hypothesis testing process can be summarised as:

1. select a suitable test
2. $P(T \geq t|H_0) \rightarrow$ p-value

Bayesian: H_0: $\theta \in \Theta_0$ and H_1: $\theta \in \Theta_1$, where Θ_0 and Θ_1 are disjoints and exhaustive subsets of Θ.

$$\underbrace{\frac{p(H_0|\boldsymbol{y})}{p(H_1|\boldsymbol{y})}}_{\text{posterior odds}} = \underbrace{\frac{p(\boldsymbol{y}|H_0)}{p(\boldsymbol{y}|H_1)}}_{\substack{\text{Bayesian factor} \\ \text{(BF)}}} x \underbrace{\frac{p(H_0)}{p(H_1)}}_{\text{prior odds}}$$

If posterior odds >1 $\rightarrow$ favour H_0

If too difficult to specify prior odds $\rightarrow$ report Bayesian factor only.

End of the theory

PRACTICAL (1.5 days)

3. Introduction to R-INLA (0.5 day)
   - 3.1 A brief overview of R-INLA
   - 3.2. Practical 1: Hubble's Law
   - 3.3. Practical 2: Theory of INLA
4. Bayesian regression (0.5 day)
   - 4.1. Practical 3: Bayesian regression
   - 4.2. Practical 4: Generalized linear models
   - 4.3. Practical 5: Spatial models
5. Personal work (0.5 day)
   - 5.1.a) Further work on examples in the course
   - 5.1.b) Personal work using new datasets
   - 5.2. Summary & further guidance

First practical

Time: morning and afternoon November 8

Materials: please use the following materials:

- Reference provided by the teacher: Chapter 1: Introduction. From Wang, X., Yue Ryan, Y., Faraway, J. (2018). Bayesian Regression Modeling with INLA. New York: Chapman and Hall/CRC.
- Data: provided by the teacher. Also available (slightly amended) here: https://github.com/julianfaraway/brinla/tree/master/docs/scripts