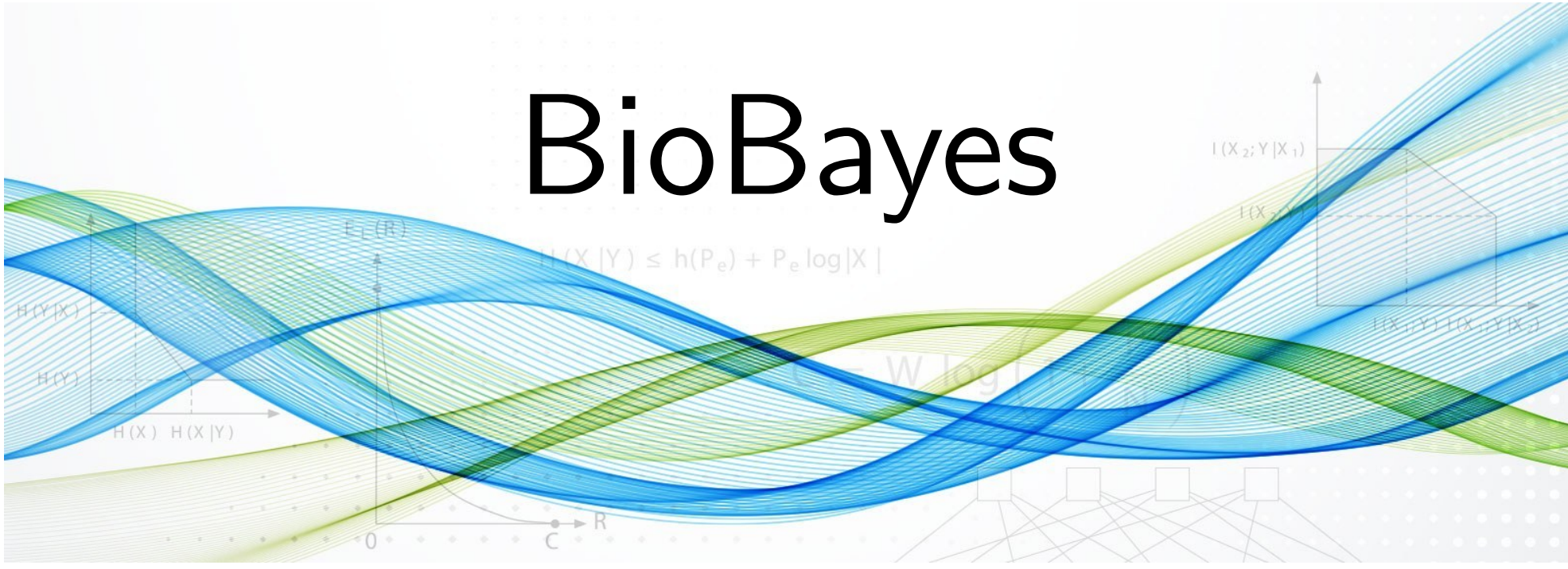


BioBayes

Crash-course on Bayesian statistics with R

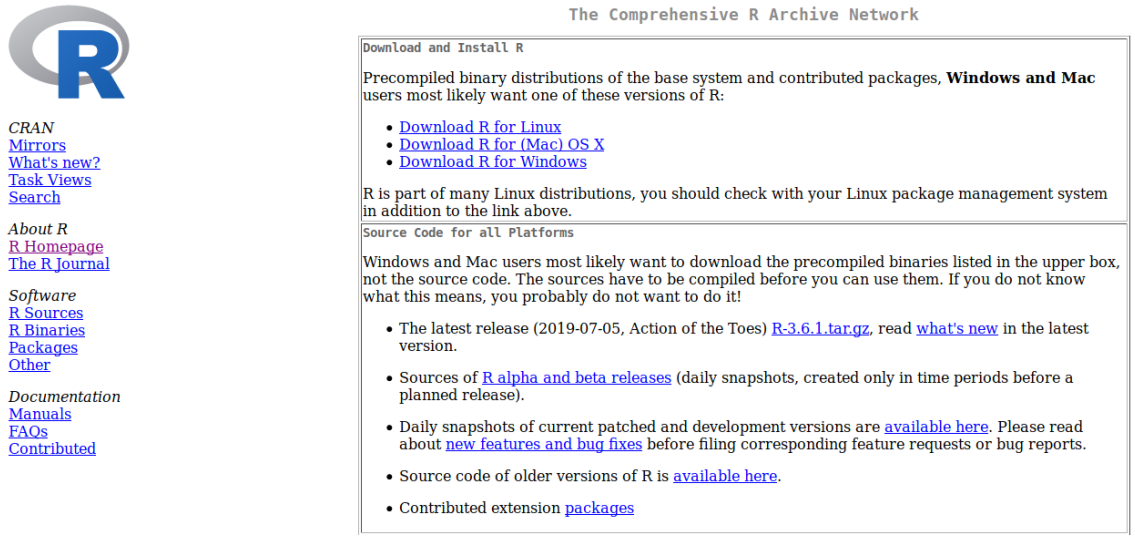
Model-guided data science, Como
5 Sep 2019

Plan

PART 1	Facts about Bayesian stats
PART 2	Practical examples with *RJags*

1. Download and install Rstudio (requires R):

R	https://cran.rstudio.com/
Rstudio	https://www.rstudio.com/products/rstudio/download/



The Comprehensive R Archive Network

CRAN
Mirrors
What's new?
Task Views
Search

About R
R Homepage
The R Journal

Software
R Sources
R Binaries
Packages
Other

Documentation
Manuals
FAQs
Contributed

Download and Install R

Precompiled binary distributions of the base system and contributed packages, **Windows and Mac** users most likely want one of these versions of R:

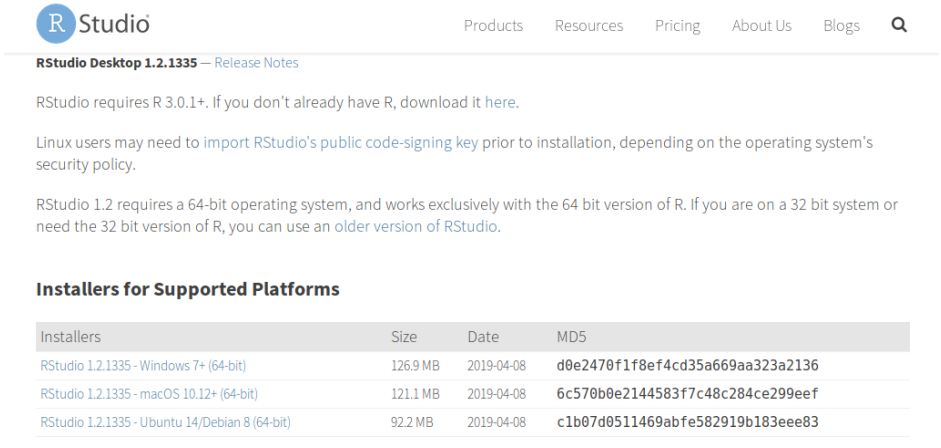
- Download R for Linux
- Download R for (Mac) OS X
- Download R for Windows

R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.

Source Code for all Platforms

Windows and Mac users most likely want to download the precompiled binaries listed in the upper box, not the source code. The sources have to be compiled before you can use them. If you do not know what this means, you probably do not want to do it!

- The latest release (2019-07-05, Action of the Toes) R-3.6.1.tar.gz, read what's new in the latest version.
- Sources of R alpha and beta releases (daily snapshots, created only in time periods before a planned release).
- Daily snapshots of current patched and development versions are available here. Please read about new features and bug fixes before filing corresponding feature requests or bug reports.
- Source code of older versions of R is available here.
- Contributed extension packages



RStudio Products Resources Pricing About Us Blogs

RStudio Desktop 1.2.1335 — Release Notes

RStudio requires R 3.0.1+. If you don't already have R, download it here.

Linux users may need to import RStudio's public code-signing key prior to installation, depending on the operating system's security policy.

RStudio 1.2 requires a 64-bit operating system, and works exclusively with the 64 bit version of R. If you are on a 32 bit system or need the 32 bit version of R, you can use an older version of RStudio.

Installers for Supported Platforms

Installers	Size	Date	MD5
RStudio 1.2.1335 - Windows 7+ (64-bit)	126.9 MB	2019-04-08	d0e2470f1f8ef4cd35a669aa323a2136
RStudio 1.2.1335 - macOS 10.12+ (64-bit)	121.1 MB	2019-04-08	6c570b0e2144583f7c48c284ce299eef
RStudio 1.2.1335 - Ubuntu 14/Debian 8 (64-bit)	92.2 MB	2019-04-08	c1b07d0511469abfe582919b183eee83

2. Install the JAGS library.
   - **linux/ubuntu:**
     from package manager or source: http://mcmc-jags.sourceforge.net/
   - **windows:**
     https://sourceforge.net/projects/mcmc-jags/files/JAGS/4.x/Windows/

   for more infos about the jags library visit the webpage
   http://mcmc-jags.sourceforge.net/
3. Open Rstudio and install RJAGS by running
   $>$ install.packages("rjags")
4. Lecture notes and R scripts available on github
   https://github.com/giovannidiana/BioBayes

PART 1

Introduction

- Understanding data variability
- Statistical confidence and probability
- Building statistical models
- Inference of “unknown” features from “known” data

Statistical models

Statistical models provide a probabilistic description of how the data have been generated. We will consider three types of models

- **simple models:** data are modeled by a single known probability distribution (normal, exponential etc..)
- **mixed models:** data are distributed according to a linear combination of known distributions
- **hierarchical models:** the parameters are hierarchically organized

Statistical models can involve *latent* (unobserved), features which are part of the generative model.

Parameters and likelihood

Given a statistical model M characterized by a set of parameters $\boldsymbol{\theta}$ the data **likelihood**

$$L_M(\boldsymbol{\theta}) = \mathrm{Prob}(\mathit{data}|\boldsymbol{\theta}, M)$$

measures the probability to generate the data from model M using the given parameter set $\boldsymbol{\theta}$.
In the presence of latent variables **z** alongside the observed variables **x**, the likelihood is expresse as the sum over the unobserved variables

$$L_M(\boldsymbol{\theta}) = \sum_{\mathbf{z}} \mathrm{Prob}(\mathbf{x} = \mathit{data}, \mathbf{z}|\boldsymbol{\theta}, M)$$

Bayes theorem

Assume we have a set of data and a statistical model with parameters θ. We are interested in the "posterior" probability

$$\mathrm{Prob}(\theta|\mathrm{data}, \mathrm{Model})$$

Meaning: how does my uncertainty about the model parameters change after observing the data?

$$\begin{aligned} P(\theta, \mathrm{data}|\mathrm{Model}) &= P(\mathrm{data}|\mathrm{Model}) \cdot P(\theta|\mathrm{data}, \mathrm{Model}) \\ &= P(\theta|\mathrm{Model}) \cdot P(\mathrm{data}|\theta, \mathrm{Model}) \end{aligned}$$

Posterior distribution

$$P(\theta|\mathrm{data}, \mathrm{Model}) = \frac{\overbrace{P(\mathrm{data}|\theta, \mathrm{Model})}^{\text{likelihood}} \cdot \overbrace{P(\theta|\mathrm{Model})}^{\text{prior}}}{\underbrace{P(\mathrm{data}|\mathrm{Model})}_{\text{marginal likelihood}}}$$

Expectations

Given the posterior distribution $P(x)$ we can compute posterior averages of any quantity as usual

$$\mathbb{E}[x] = \langle x \rangle_P = \sum_{x \in \Omega} P(x) \cdot x$$

of any function of the model variables

$$\mathbb{E}[g(x)] = \langle g(x) \rangle_P = \sum_{x \in \Omega} P(x) \cdot g(x)$$

Sampling and Monte Carlo approximation

To calculate expectations from the formulas above we need to perform sums or integrals (often in many dimensions) $\rightarrow$ HARD!
A much easier way to compute expectations is to generate first a large "sample" of values of x drawn from its probability distribution $\{x_1, \cdots, x_N\}$ and use the **Monte Carlo approximation**

Monte Carlo approximation

$$x_i \sim P(x), \quad i = 1, \cdots, N$$

$$\mathbb{E}[g(x)] \approx \frac{1}{N} \sum_{i=1}^{N} g(x_i)$$

"$\sim$" $\rightarrow$ "drawn from"

Summary so far

- Bayesian inference corresponds to the computation of posterior probabilities.
- In general we need Monte Carlo algorithms to sample efficiently from the posterior distribution
- We can use the Monte Carlo approximation of the posterior distribution to estimate model variables of interest (means, variances, correlations...).
- We still need to focus on appropriate prior distributions to represent our prior knowledge on the model parameters.

PART 2

script_1.R: normal distribution

Estimation of mean and variance from univariate normally distributed data

1. Run script
2. Plot histograms of sampled mean and variance
3. Modify the prior distribution of the mean and resample
4. Compare the variance of the posterior distribution of the mean with the SEM estimated from the data
5. Calculate credible intervals for mean and variance using the function quantile

script_2.R: linear regression

Model:

$$y|x \sim \mathcal{N}(ax + b, \sigma = 1/\sqrt{t})$$
$$a \sim \mathcal{N}(1, 10^{-3})$$
$$b \sim \mathcal{N}(1, 10^{-3})$$
$$t \sim gamma(1, 1)$$

1. obtain posterior distributions of a and b
2. obtain posterior distribution of $\sigma = 1/\sqrt{t}$
3. Are a and b correlated? Why?
4. If we define $y^{pred}(x) \equiv ax + b$, what is the distribution of $y^{pred}(x = 13)$?
5. What is the difference between the previous calculation and $P(y|x = 13, data)$?

script_3.R: logistic regression

Assume age affects the onset of a desease and we want to quantify this effect. We can model the probability of developing the desease before a given age using the logistic distribution

$$P_{logistic}(A) = \frac{1}{1 + e^{\alpha \cdot A}}.$$

Our statistical model looks like

$$status \sim bernoulli(P_{logistic}(A))$$
$$\alpha \sim \mathcal{N}(0, 10^{-3})$$

1. plot most probable $P_{logistic}(A)$
2. what is the probability to develop the desease before the age of 50? With which confidence?

script_4.R: Genetic effect on a desease

As in exercise 6 from previous lecture, consider the four probabilities $p_{s,A}, p_{h,A}, p_{s,B}, p_{h,B}$ where A and B are variants of a gene and (s, h) stand for "sick" and "healthy" condition of the patient. We can use the model

$$\text{Model} : \{N_{s,A}, N_{h,A}, N_{s,B}, N_{h,B}\} \sim \mathcal{M}(\{p_{s,A}, p_{h,A}, p_{s,B}, p_{h,B}\}; N = 100)$$
$$\text{Prior} : \{p_{s,A}, p_{h,A}, p_{s,B}, p_{h,B}\} \sim \mathit{Dirichlet}(\vec{\alpha})$$

1. calculate the distributions of $P(s|A)$ and $P(s|B)$.
2. calculate credible intervals
3. With which statistical confidence can we claim that the variant of the gene play a role in the desease?

script_5.R: Hierarchical model

Consider measurements of cell counts. We want to model the fraction of small and big cells given data from three trials:

	trial 1	trial 2	trial 3
big	10	25	35
small	90	75	65

We can use a hierarchical model to describe variability across trial

$$
\begin{aligned}
n_k^{big} &\sim binomial(p_k, 100), \quad k = 1, 2, 3 \\
p_k &\sim Beta(\alpha, \beta) \\
\alpha, \beta &\sim exponential(10^{-3})
\end{aligned}
$$

1. compare the distribution of p^{big} obtained in this model with the distribution obtained from model1 (see the script). Are they different? why?

script_6.R: Hierarchical model of normals

Consider now a similar situation where we measured the expression of a specific gene from single cells. Suppose we have data from three independent trials. Let us consider the hierarchical model

$$
\begin{aligned}
g_i &\sim \mathcal{N}(\mu^T, t^T) \\
\mu^T &\sim \mathcal{N}(\mu_0, t_0) \\
t^T &\sim gamma(2, 10^{-2}) \\
\mu_0 &\sim \mathcal{N}(0, 10^{-3})
\end{aligned}
$$

1. plot distributions of μ and μ_0
2. calculate the standard deviation of μ_0 and compare with the naive SEM obtained from the data.

script_7.R: Mixture of Gaussians

We can model the angle distribution with a mixture of two normal distributions. We define a "cell type" variable (latent) $t = 1, 2$ so that we have

$$
\begin{aligned}
angle_i &\sim Normal(\mu_i, t_i) \\
t_i &\sim Binomial(1/2, N) \\
\mu_1, \mu_2 &\sim Normal(0, 0.001) \\
\sigma_1, \sigma_2 &\sim Gamma(2, 1)
\end{aligned}
$$

NOTE 1: R multi-dimensional array

RJAGS returns draws from the posterior distribution in the format of a multidimensional array, which is the R container for multivariate data.

- For single numeric parameters:

$$A[1, sample, chain]$$

- For vector parameters $\theta = \{\theta_1, \theta_2, \cdots, \theta_K\}$

$$A[i, sample, chain]$$

- Form matrix parameters $\boldsymbol{\sigma}$

$$A[i, j, sample, chain]$$

NOTE 2: Prediction from a model and propagation of parameter uncertainties

Given a model M with parameters θ and some data, making a prediction on a new data point x correspond to evaluate the probability

$$P(x|data, M) = \sum_{\theta} P(x, \theta|data, M)$$
$$= \sum_{\theta} \underbrace{P(\theta|data, M)}_{posterior} \cdot P(x|\theta, M)$$

where we used the fact that x is conditionally independent of the data given the set of parameter θ. Now we can use the Monte Carlo approximation to get

$$P(x|data, M) \approx \frac{1}{N} \sum_{i} P(x|\theta_i, M)$$

where the sum runs over all the sampled values of the parameters.

NOTE 3: Marginal likelihood

The marginal likelihood (model evidence) is

$$L_M(data) = \sum_{\theta} P(data, \theta | M) =$$

$$= \sum_{\theta} P(\theta | M) \cdot P(data | \theta, M)$$