Data Modelling & Parameter estimation

School on Statistics, Data Mining, and Machine Learning
Instituto de Astrofísica de Andalucía (CSIC)

Dr. Carmen Sánchez Gil

Tuesday, Nov 5th 2019

Universidad de Cádiz

Table of contents

1. Preliminaries
2. Single Parameter Models
3. Multi Parameter Models
4. Bayesian Computation
5. Hierarchical Models

Preliminaries

Data modelling & parameter estimation

- We perform experiments or make observations in order to learn about a phenomenon, which only can be partially observed.
- First step: describe the resulting data (plots, summaries, statistics, etc)
- To interpret the data we usually have to model them
- Data (measurements) are always noisy
- Inference is the process of making general statements about a phenomenon, via a model, using noisy and incomplete data.
- *Generative model* is the theoretical model that generates (simulates) the observable data from the model parameters (mathematical equation)
- *Measurement or noise model* describes how the measurement process affects our data. It describes a probability distribution over possible observations given the ideal (noise-free) data, i.e. the Likelihood.

Example

$x \sim \mathcal{N}(\mu, \sigma)$, where x is the measured data, $\mu = g(\theta)$ is the ideal data, and σ is the uncertainty in the measurement: $f(x) = \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{(x-g(\theta))^2}{2\sigma^2}\right)$

Data modelling & parameter estimation

The key to data modeling is to use the given data, together the generative and measurement models to make consistent, probabilistic inferences.

Given some data D, for a specified model M, with parameter(s) θ:

- Parameter estimation. To infer the *parameter posterior* pdf $p(\theta|D, M)$
- Model comparison. Given a set of different models $\{M_i\}$, find out which one is best supported by the data: *model posterior probability* $P(M_i|D)$, or *posterior odds ratio* of two models $P(D|M_i)/P(D|M_j)$
- Prediction. Predict some new data, $p(\tilde{x}|D, M)$

Probability Theory

Notation

- x denotes the studied random variable (*phenomenon*), or the given data $D \equiv x$ (*measurements*) ($\boldsymbol{x} = \{x_1, \ldots, x_n\}$ random sample)
- We use the terms 'distribution' and 'density' interchangeably, and with same notation: $p(x) \quad (P(x > 2) = \int_{x>2} p(x)dx)$
- $x \sim N(\mu, \sigma)$ or $p(x) = N(x|\mu, \sigma)$
- $E[x] = \int xp(x)dx$; $var(x) = \int (x - E[x])^2 p(x)dx$
- Given u, v: $p(u, v)$ is the *joint density function*, $p(u|v)$ the *condicional pdf*, and $p(u) = \int p(u, v)dv$ *marginal pdf* (and vice versa).
- The joint pdf can be *factorized* as the product of the marginal and conditional pdf's : $p(u, v) = p(u|v)p(v)$, or $p(u, v, w) = p(u|v, w)p(v|w)p(w)$, etc

Probability Theory

Bayesian Inference

Given a *model* M, with parameter(s) θ,

- Likelihood: $p(x|\theta, M) \equiv p(x|\theta)$, *sampling* or *data distribution*. Key function in data modeling, it describes both the *phenomenon* and the *measurements*.

Probability Theory

Bayesian Inference

Given a *model* M, with parameter(s) θ,

- Likelihood: $p(x|\theta, M) \equiv p(x|\theta)$, *sampling* or *data distribution*. Key function in data modeling, it describes both the *phenomenon* and the *measurements*.
- Prior: $p(\theta|M) \equiv p(\theta)$, pdf over the model params. θ. Information we have, independent of the data, about the possible values of θ.

Probability Theory

Bayesian Inference

Given a *model* M, with parameter(s) θ,

- Likelihood: $p(x|\theta, M) \equiv p(x|\theta)$, *sampling* or *data distribution*. Key function in data modeling, it describes both the *phenomenon* and the *measurements*.
- Prior: $p(\theta|M) \equiv p(\theta)$, pdf over the model params. θ. Information we have, independent of the data, about the possible values of θ.
- *Joint probability d.* for θ and x: $p(\theta, x) = p(\theta)p(x|\theta)$

Bayesian Inference

Given a *model* M, with parameter(s) θ,

- Likelihood: $p(x|\theta, M) \equiv p(x|\theta)$, *sampling* or *data distribution*. Key function in data modeling, it describes both the *phenomenon* and the *measurements*.
- Prior: $p(\theta|M) \equiv p(\theta)$, pdf over the model params. θ. Information we have, independent of the data, about the possible values of θ.
- *Joint probability d.* for θ and x: $p(\theta, x) = p(\theta)p(x|\theta)$
- Posterior: $p(\theta|x, M)$ pdf over the model params., given the data and the background inform. on M, is the answer to an inference problem

$$\boxed{\text{Bayes' rule}} \quad p(\theta|x) = \frac{p(\theta, x)}{p(x)} = \frac{p(\theta)p(x|\theta)}{p(x)}$$

$$\propto \underbrace{p(\theta)p(x|\theta) = p^*(\theta|x)}_{\text{unnormalized post}}$$

Probability Theory

Bayesian Inference

- Evidence or *marginal likelihood*: probability, assuming model M, of observing the data for any values of θ,

$$p(x|M) = \begin{cases} \sum_\theta p(\theta)p(x|\theta), & \text{discrete} \\ \int p(\theta)p(x|\theta)d\theta, & \text{continuous} \end{cases} \qquad \text{(normalization cte)}$$

Probability Theory

Bayesian Inference

- Evidence or *marginal likelihood*: probability, assuming model M, of observing the data for any values of θ,

$$p(x|M) = \begin{cases} \sum_\theta p(\theta)p(x|\theta), & \text{discrete} \\ \int p(\theta)p(x|\theta)d\theta, & \text{continuous} \end{cases} \qquad \text{(normalization cte)}$$

- Multiparameter models: $\theta = (\theta_1, \theta_2)$
  - Joint posterior density: $p(\theta_1, \theta_2|x) \propto p(\theta_1, \theta_2)p(x|\theta_1, \theta_2)$
  - Conditional posterior distributions: $p(\theta_1|x, \theta_2)$, and $p(\theta_2|x, \theta_1)$
  - Marginal posterior distribution of θ_1, by averaging or marginalizing over θ_2 (and vice versa):

$$p(\theta_1|x) = \int p(\theta_1, \theta_2|x)d\theta_2 = \int p(\theta_1|x, \theta_2)p(\theta_2|x)d\theta_2$$

Probability Theory

Bayesian Inference

- Prior predictive distribution or *marginal distribution of x*: before data are considered, the distr. of the unknown, observable x is

$$p(x) = \int p(\theta, x) d\theta = \int p(\theta) p(x|\theta) d\theta$$

- Posterior predictive distribution: after the data have been observed, we can predict an unknown observable $\tilde{x}$ from the same process

$$\begin{aligned} p(\tilde{x}|x) &= \int p(\tilde{x}, \theta|x) d\theta = \int p(\tilde{x}|\theta, x) p(\theta|x) d\theta \\ &\overset{*}{=} \int p(\tilde{x}|\theta) p(\theta|x) d\theta \end{aligned}$$

* Assumed conditional independence of x and $\tilde{x}$ given θ

Probability Theory

Mean and variance of conditional distributions

$$
\begin{aligned}
E(u) &= E(E(u|v)) \\
Var(u) &= E(Var(u|v)) + Var(E(u|v))
\end{aligned}
$$

Transformation of variables

$v = f(u)$, u and v has the same dimension, and f is a one-to-one function. If $p_u(u)$ is the density f. of the variable u, then

$$
p_v(v) = \begin{cases} p_u(f^{-1}(v)) & \text{discrete distribution} \\ p_u(f^{-1}(v)) \left| \frac{dv}{du} \right| & \text{continuous distribution} \end{cases}
$$

NOTE:

$$
\begin{aligned}
u \in (0, \infty) \quad &\Rightarrow \quad log(u) \in \mathcal{R} \\
u \in [0, 1] \quad &\Rightarrow \quad logit(u) = log\left(\frac{u}{1-u}\right) \in \mathcal{R}, \text{ where } logit^{-1}(v) = \frac{e^v}{1+e^v}
\end{aligned}
$$

Computation & simulations in Bayesian Inference

Summarizing the posterior distribution

1. Choose a *grid* of θ over an interval that covers the post. d.
2. Compute the product of the prior, $p(\theta)$, and the likelihood $\mathcal{L}(\theta) = f(\boldsymbol{x}|\theta)$ on the grid: $p(\theta_i|\boldsymbol{x}) \propto p(\theta_i)\mathcal{L}(\theta_i)$, $i = 1, \ldots, n$.
3. Normalize, to approximate the posterior density by a discrete probability distribution on the grid:

$$p(\theta_i|\boldsymbol{x}) \simeq \frac{p(\theta_i)\mathcal{L}(\theta_i)}{\sum_{j=1}^{n} p(\theta_j)\mathcal{L}(\theta_j)} \quad i = 1 \ldots, n$$

4. Take a sample with replacement from the discrete distribution $\{\theta_1, \ldots, \theta_m \mid \boldsymbol{x}\}$ (m=1000 adequate for estimating the P_{95} in this way)

$\Rightarrow$ Simulation forms a central part of the Bayesian analysis applications.

$\Rightarrow$ To draw easily approximate samples from post. d., even when the density function cannot be explicitly integrated.

Computation & simulations in Bayesian Inference

Sampling using the inverse cumulative distribution function

Cdf $F(a) = P(x \leq a) = \begin{cases} \sum_{x \leq a} p(x) & \text{discrete} \\ \int_{-\infty}^{a} p(x)dx & \text{continuous} \end{cases}$

1. Draw random sample from $u \sim \mathcal{U}(0,1)$: $\{u_1, \ldots, u_m\}$
2. Let $x = F^{-1}(u)$ (F not necessarily 1-to-1, but $F^{-1}(u)$ unique)
3. Then, $\{F^{-1}(u_1), \ldots, F^{-1}(u_m)\}$ will be a random draw from $p(x)$

Examples

- $x \sim Exp(\lambda)$, $F(x) = 1 - e^{-\lambda x} \to x = F^{-1}(u) = -\frac{log(1-u)}{\lambda}$. Draw $\{u_1, \ldots, u_m\}$ from $\sim U(0,1) \to \left\{-\frac{log(u_1)}{\lambda}, \ldots, -\frac{log(u_n)}{\lambda}\right\}$ sample from $Exp(\lambda)$
- $x_1 \leq x_2 \leq \ldots \leq x_k$, with probability mass function p_i ($\sum_{i=1}^{m} p_i = 1$), and let $F(x_j) = \sum_{i \leq j} p_i$. Given $u \sim U(0,1)$, then:
  $P(F(x_{j-1}) \leq u \leq F(x_j)) = F(x_j) - F(x_{j-1}) = p_j = P(x = x_j)$

Integration in Bayesian Inference

- Marginal parameter distributions

$$p(\theta_1|x) = \int p(\theta_1, \theta_2|x)d\theta_2$$

- Expectation values (parameter estimation)

$$E[\theta] = \int \theta p(\theta|x)d\theta$$

- Evidence (marg. likelihood - model comparison)

$$p(x|M) = \int p(x|\theta, M)p(\theta|M)d\theta$$

- Prediction (post. predictive d.)

$$p(\tilde{x}|x) = \int p(\tilde{x}|\theta)p(\theta|x)d\theta$$

Monte Carlo integration

A simple solution to integrating a function is to evaluate it over a dense, regular grid: $\{\theta_1, \ldots, \theta_n\}$,

$$\int_{\theta_{min}}^{\theta_{max}} f(\theta)d\theta \approx \sum_{i=1}^{n} f(\theta_i)\delta\theta = \frac{\theta_{max} - \theta_{min}}{n} \sum_{i=1}^{n} f(\theta_i)$$

Model Comparison & Bayesian Evidence

2 competing models: M_1 and M_2,

Posterior odds ratio

$$R = \frac{p(M_1|x)}{p(M_2|x)} = \frac{p(M_1)p(x|M_1)}{p(M_2)p(x|M_2)} = \frac{p(M_1)}{p(M_2)} BF_{1,2}$$

Bayes Factor

$$BF_{1,2} = \frac{p(x|M_1)}{p(x|M_2)} = \frac{\int p(x|\theta_1, M_1)p(\theta_1|M_1)d\theta_1}{\int p(x|\theta_1, M_1)p(\theta_1|M_1)d\theta_1}$$

The evidence as a marginal likelihood

$$p(x|M) = \int \underbrace{p(x|\theta, M)}_{likelihood} \underbrace{p(\theta|M)}_{prior} d\theta$$

Assigning priors

How do we assign a prior?

- posterior pdf depends on both the prior and the likelihood;
- As data become more informative, posterior dominated by the likelihood (narrower);
- When data are poor, prior plays a more dominant role.
- Prior should incorporate any relevant information we have, what you know/believe/understand about the problem, the parameter range, limits/bounds of measurement or observability (there is no rule)
- Often we adopt standard distributions; discrete priors using histogram (finite support), etc.
- **Non-informative priors**: No population basis, minimal role in the posterior distribution (uniform !!)
- Improper priors can lead to proper posterior distributions

Assigning priors: Non-informative priors

Location paramaters

- θ specifies the location of some quantity (mean), and we have no prior knowledge other than some limits/range
- Posterior should be independent of the origin coord. system $p(x-\theta|x) \propto p(\theta)p(x-\theta|\theta)$

$\Rightarrow$ prior invariant to linear transformation of θ, $p(\theta+c)d\theta = p(\theta)d\theta$: $p(\theta) \propto cte$

Scale paramaters

- θ size or scale of some quantity (std. dev), and we know nothing about it, other than it must be positive.

$\Rightarrow$ prior invariant with respect to being stretched, $p(\theta)d\theta = p(c\theta)cd\theta$: $p(\theta) \propto \frac{1}{\theta}$

- Equiv. : $p(log\theta) \propto 1$, or $p(\theta^2) \propto \frac{1}{\theta^2}$

Assigning priors: Non-informative priors

Jeffreys prior

- *Jeffreys' invariance principle*: an approach to define no-inform prior, based on $1-1$ transformations, $\phi = h(\theta) : p(\phi) = p(\theta)\left|\frac{d\theta}{d\phi}\right| = p(\theta)\left|h'(\theta)\right|^{-1}$

$\Rightarrow$ $\boxed{p(\theta) \propto \mathcal{I}(\theta)^{1/2}}$, where $\mathcal{I}(\theta) = -E\left[\frac{\partial^2 log\ p(x|\theta)}{\partial^2\theta}\right]$ *Fisher information*

- Invariant respect to parameterizations:

$$\mathcal{I}(\phi)^{1/2} = \left(-E\left[\frac{\partial^2 log\ p(x|\phi)}{\partial^2\phi}\right]\right)^{\frac{1}{2}} = \left(-E\left[\frac{\partial^2 log\ p(x|\theta{=}h^{-1}(\phi))}{\partial^2\theta}\left|\frac{d\theta}{d\phi}\right|^2\right]\right)^{\frac{1}{2}} = \mathcal{I}(\theta)^{1/2}\left|\frac{d\theta}{d\phi}\right|$$

Example: Jeffreys prior for binomial likelihood

- $\frac{\partial ln\ L(x,\theta)}{\partial\theta} = \frac{\partial\left(ln\binom{n}{x}+x ln(\theta)+(n-x)ln(1-\theta)\right)}{\partial\theta} = \frac{x}{\theta} - \left(\frac{n-x}{1-\theta}\right) \rightarrow \frac{\partial^2 L(x,\theta)}{\partial^2\theta} = -\frac{x}{\theta^2} - \left(\frac{n-x}{(1-\theta)^2}\right)$
- $\mathcal{I}(\theta) = -E\left[\frac{\partial^2 log\ p(x|\theta)}{\partial^2\theta}\right] = -E\left[-\frac{x}{\theta^2} - \left(\frac{n-x}{(1-\theta)^2}\right)\right] = \frac{n\theta}{\theta^2} - \left(\frac{n-\theta}{(1-\theta)^2}\right) = \frac{n}{\theta(1-\theta)}$
- $p(\theta) \propto \theta^{-1/2}(1-\theta)^{-1/2}$

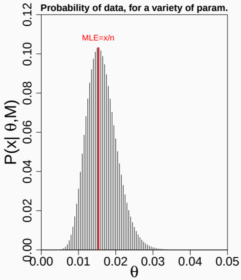
Single Parameter Models

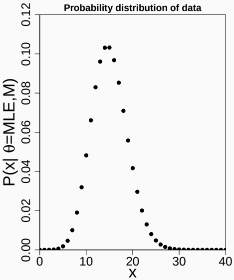
Single-parameter models: Binomial distribution

- $x =$ total number of successes in the n *Bernouilli* trials (0/1: failure/success).
- $p(x|\theta) = B(x|n,\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$, θ proportion of successes, or probability of success in each trial.

Example I: Estimate the AGN fraction in a galaxy sample

- It is observed a sample of 980 galaxies
- 15 of which are classified as AGN
- $0.001 \leq p \leq 0.015$ at low redshift (Bufanda et al. 2016)





$\Rightarrow$ To perform Bayesian inference, we must specify a prior distr. for θ

Single-parameter models: Binomial distribution

Non-informative prior

- Prior $\boxed{\theta \sim \mathcal{U}(0,1)}$: $p(\theta) = 1$
- Likelihood: $p(x|\theta) = \mathcal{L}(\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$
- Posterior distribution $p(\theta|x) \propto p(\theta)p(x|\theta)$:

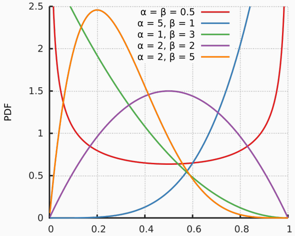
$$p(\theta|x) \propto \theta^x(1-\theta)^{n-x} \Rightarrow \theta|x \sim Beta(x+1, n-x+1)$$

Single-parameter models: Binomial distribution

Non-informative prior

- Prior $\boxed{\theta \sim \mathcal{U}(0,1)}$: $p(\theta) = 1$
- Likelihood: $p(x|\theta) = \mathcal{L}(\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$
- Posterior distribution $p(\theta|x) \propto p(\theta)p(x|\theta)$:

$$p(\theta|x) \propto \theta^x(1-\theta)^{n-x} \Rightarrow \theta|x \sim Beta(x+1, n-x+1)$$



Beta($\boldsymbol{\alpha}, \boldsymbol{\beta}$) distribution

$p(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}, \quad \theta \in [0,1]$
'Prior sample sizes' $\alpha > 0, \beta > 0$

$$E(\theta) = \frac{\alpha}{\alpha+\beta} \; ; \quad Var(\theta) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

$$Mo(\theta) = \frac{\alpha-1}{\alpha+\beta-2}$$

Single-parameter models: Binomial distribution

Posterior as a compromise between data & prior

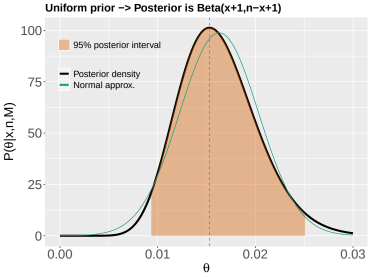
$$E(\theta) = E(E(\theta|x))$$
$$Var(\theta) = E(Var(\theta|x)) + Var(E(\theta|x))$$

Summarizing post. inference

- Plots: Post. d contains all the current inform. about param.
- Numerical summaries: mean, median, mode(s), std. dev, interquantile range, ...
- Post. uncertainty: post. quantiles and intervals central interv. of post prob $100(1-\alpha)\%$: $(P_{100(\alpha/2)\%}, P_{100(1-\alpha/2)\%})$

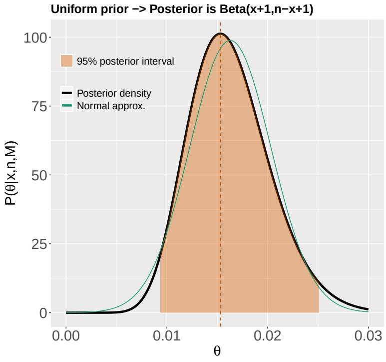
Non-informative prior

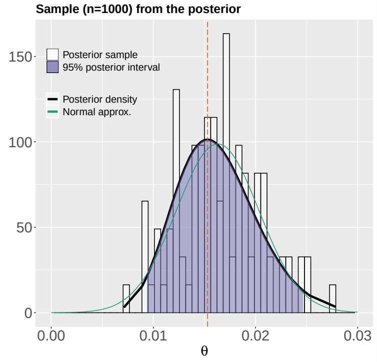
- $E(\theta|x) = \dfrac{x+1}{n+2} = \lambda \underbrace{\dfrac{1}{2}}_{E(\theta)} + (1-\lambda) \underbrace{\dfrac{x}{n}}_{\bar{\theta}}, \lambda \in [0,1]$
- $Var(\theta|x) = \frac{(x+1)(n-x+1)}{(n+2)^2(n+3)}$
- $M_o = \frac{x}{n}$



Example I: Estimate the AGN fraction in a galaxy sample

Summarizing post.	$P_{2.5}$	P_{50}	$P_{97.5}$	$E(\theta\|x)$	$Var(\theta\|x)$	M_o
Exact $p(\theta\|x)$	9.351e-3	1.597e-2	2.509e-2	1.629e-2	4.038e-3	1.531e-2
Normal approx.	8.379e-3	1.629e-2	2.421e-2	1.629e-2	4.038e-3	1.629e-2
post. sample n=1000	9.285e-3	1.560e-2	2.478e-2	1.606e-2	4.070e-3	1.533e-2





Single-parameter models: Binomial distribution

Informative prior: conjugated family

- Prior $\boxed{\theta \sim Beta(\alpha, \beta)}$: $p(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$ conjugate family for the binomial likelihood
- Likelihood: $p(x|\theta) = \mathcal{L}(\theta) = \binom{n}{x}\theta^{x}(1-\theta)^{n-x}$
- Posterior distribution

$$p(\theta|x) \propto \theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1} \Rightarrow \theta|x \sim Beta(\alpha + x, \beta + n - x)$$

Single-parameter models: Binomial distribution

Informative prior: conjugated family

- Prior $\boxed{\theta \sim Beta(\alpha, \beta)}$: $p(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$ `conjugate` family for the binomial likelihood
- Likelihood: $p(x|\theta) = \mathcal{L}(\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$
- Posterior distribution

$$p(\theta|x) \propto \theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1} \Rightarrow \theta|x \sim Beta(\alpha + x, \beta + n - x)$$

- $E(\theta|x) = \dfrac{\alpha + x}{\alpha + \beta + n} = \lambda \underbrace{\dfrac{\alpha}{\alpha + \beta}}_{E(\theta)} + (1 - \lambda) \underbrace{\dfrac{x}{n}}_{\bar{\theta}}, \lambda \in [0, 1]$
- $Var(\theta|x) = \frac{E(\theta|x)(1-E(\theta|x))}{\alpha+\beta+n+1}$

Informative prior: conjugated family

- Prior $\boxed{\theta \sim Beta(\alpha, \beta)}$: $p(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$ `conjugate` family for the binomial likelihood
- Likelihood: $p(x|\theta) = \mathcal{L}(\theta) = \binom{n}{x}\theta^{x}(1-\theta)^{n-x}$
- Posterior distribution

$$p(\theta|x) \propto \theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1} \Rightarrow \theta|x \sim Beta(\alpha + x, \beta + n - x)$$

- $E(\theta|x) = \dfrac{\alpha + x}{\alpha + \beta + n} = \lambda \underbrace{\dfrac{\alpha}{\alpha + \beta}}_{E(\theta)} + (1-\lambda) \underbrace{\dfrac{x}{n}}_{\bar{\theta}}, \lambda \in [0,1]$
- $Var(\theta|x) = \frac{E(\theta|x)(1-E(\theta|x))}{\alpha+\beta+n+1}$
- As $\uparrow x$, $\uparrow (n-x)$, α, β fixed: $E(\theta|x) \approx \frac{x}{n}$, $Var(\theta|x) \approx \frac{1}{n}\frac{x}{n}(1-\frac{x}{n})$

Single-parameter models: Binomial distribution

Informative prior: conjugated family

- Prior $\boxed{\theta \sim Beta(\alpha, \beta)}$: $p(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$ `conjugate` family for the binomial likelihood
- Likelihood: $p(x|\theta) = \mathcal{L}(\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$
- Posterior distribution

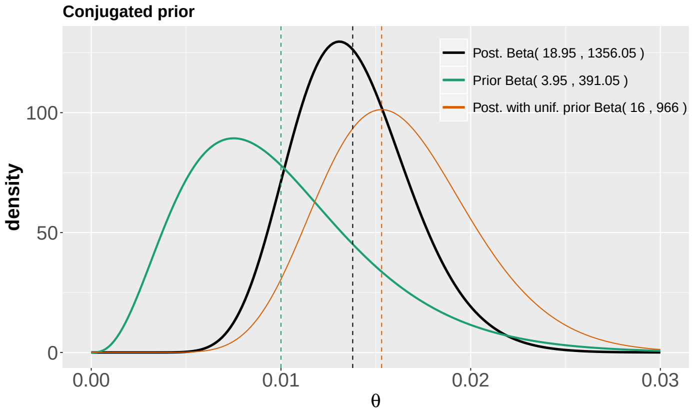
$$p(\theta|x) \propto \theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1} \Rightarrow \theta|x \sim Beta(\alpha+x, \beta+n-x)$$

- $E(\theta|x) = \dfrac{\alpha+x}{\alpha+\beta+n} = \lambda \underbrace{\dfrac{\alpha}{\alpha+\beta}}_{E(\theta)} + (1-\lambda) \underbrace{\dfrac{x}{n}}_{\bar{\theta}}, \lambda \in [0,1]$
- $Var(\theta|x) = \frac{E(\theta|x)(1-E(\theta|x))}{\alpha+\beta+n+1}$
- As $\uparrow x$, $\uparrow (n-x)$, α, β fixed: $E(\theta|x) \approx \frac{x}{n}$, $Var(\theta|x) \approx \frac{1}{n}\frac{x}{n}(1-\frac{x}{n})$
- $\left(\frac{\theta - E(\theta|x)}{DT(\theta|x)}\middle| x\right) \xrightarrow{CLT} N(0,1)$ $\quad$ (more accurate $\phi = logit(\theta) = \frac{\theta}{1-\theta}$)

Single-parameter models: Binomial distribution

Example I: Estimate the AGN fraction in a galaxy sample

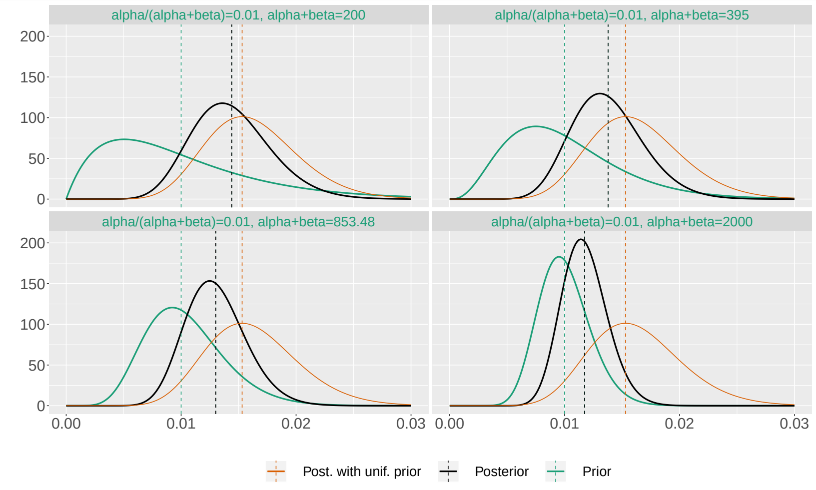
- Prior knowledge/assumption: $E(\theta) = 0.01$, $Var(\theta) = 2.5e-5$
- Data: $x = 15$ AGN in a sample of $n = 980$ galaxies



Single-parameter models: Binomial distribution

Example I: Estimate the AGN fraction in a galaxy sample

Illustrate the effect of priors



Single-parameter models: Binomial distribution

Informative non-conjugated prior

'Brute-force' numerical approximation method

1. Choose a *grid* $\{\theta_i\}$ of θ over an interval that covers the post. d.
2. Compute the product of the prior, $p(\theta)$, and the likelihood $\mathcal{L}(\theta) = f(\boldsymbol{x}|\theta)$ on the grid:

$$p(\theta_i|\boldsymbol{x}) \propto p(\theta_i)\mathcal{L}(\theta_i), \quad i = 1, \dots, n$$

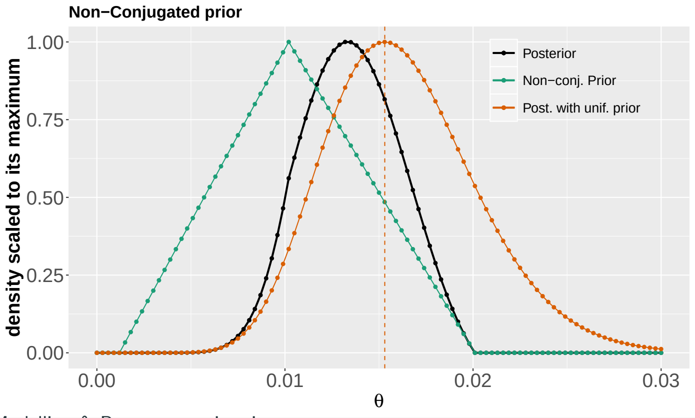
3. Normalize, to approximate the posterior density by a discrete probability distribution on the grid:

$$p(\theta_i|\boldsymbol{x}) \simeq \frac{p(\theta_i)\mathcal{L}(\theta_i)}{\sum_{j=1}^{n} p(\theta_j)\mathcal{L}(\theta_j)} \quad i = 1 \dots, n$$

Single-parameter models: Binomial distribution

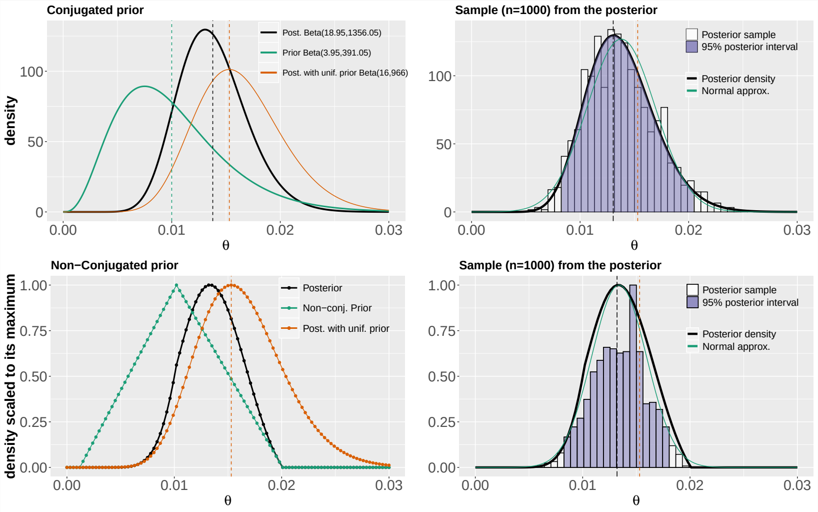
Example I: Estimate the AGN fraction in a galaxy sample

As an alternative to the conjugated beta family, we might prefer a prior distribution that is centered around 0.01 but is flat far away from this value to admit the possibility that the truth is far away (piecewise linear prior density)



Single-parameter models: Binomial distribution

Example I: Estimate the AGN fraction in a galaxy sample



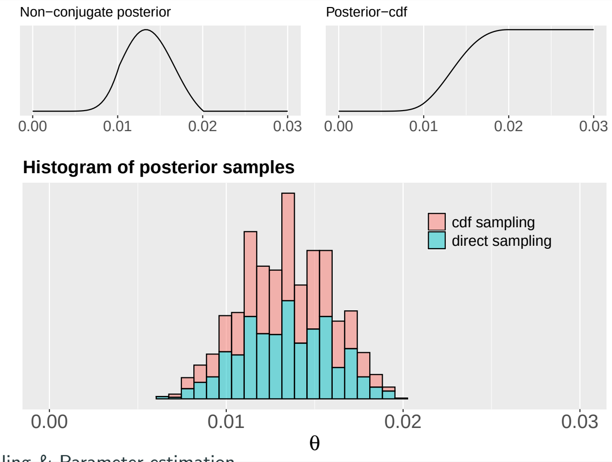
Example I: Estimate the AGN fraction in a galaxy sample

Prior	Summaries of the post. d.	$E(\theta\|x)$	$SD(\theta\|x)$	M_o	$P_{2.5}$	$P_{97.5}$
$U(0,1)$	Exact $p(\theta\|x)$	1.629e-02	4.038e-03	1.531e-02	9.351e-03	2.509e-02
	post. sample n=1000	1.629e-02	3.920e-03	3.421e-02	9.476e-03	2.463e-02
$Beta(3.95, 391.05)$	Exact $p(\theta\|x)$	1.378e-02	3.143e-03	1.307e-02	8.317e-03	2.058e-02
	post. sample n=1000	1.372e-02	3.126e-03	2.402e-02	8.598e-03	2.023e-02
$Beta(8.86, 844.62)$	Exact $p(\theta\|x)$	1.301e-02	2.646e-03	1.248e-02	8.348e-03	1.868e-02
$Beta(2, 198)$	Exact $p(\theta\|x)$	1.441e-02	3.467e-03	1.358e-02	8.421e-03	2.194e-02
$Beta(20, 1980)$	Exact $p(\theta\|x)$	1.174e-02	1.973e-03	1.142e-02	8.197e-03	1.591e-02
Non-conj.	discrete approx. $p(\theta_i\|x)$	1.350e-02	2.543e-03	1.320e-02	8.700e-03	1.830e-02
	post. sample n=1000	1.356e-02	2.533e-03	1.980e-02	9.000e-03	1.830e-02

Single-parameter models: Binomial distribution

Example I: Estimate the AGN fraction in a galaxy sample

Simulate samples from the resulting non-standard posterior distribution using inverse cdf using the discrete grid.



Single-parameter models: Binomial distribution

Bayes Factor

$$BF_{2,1} = \frac{p(x|M_2)}{p(x|M_1)} = \frac{\int p(x|\theta_2, M_2)p(\theta_2|M_2)d\theta_2}{\int p(x|\theta_1, M_1)p(\theta_1|M_1)d\theta_1}$$

Example I: Estimate the AGN fraction in a galaxy sample

- (A) Non-inform. prior $\theta \sim \mathcal{U}(0,1) = Beta(1,1)$: $p(\theta|M_A) = 1$
- (B) Inform. conj. prior $\theta \sim Beta(\alpha, \beta)$: $p(\theta|M_B) = B(\alpha,\beta)^{-1}\theta^{\alpha-1}(1-\theta)^{\beta-1}$
- Likelihood: $p(x|\theta) = \mathcal{L}(\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$

$\rightarrow$ $p(x|M_A) = \binom{n}{x}\int_0^1 \theta^x(1-\theta)^{n-x}d\theta = \binom{n}{x}B(x+1, n-x+1) = \boxed{\frac{1}{n+1}} = \frac{1}{981}$

$\rightarrow$ $p(x|M_B) = \frac{\binom{n}{x}}{B(\alpha,\beta)}\int_0^1 \theta^{x+\alpha-1}(1-\theta)^{n+\beta-x-1}d\theta = \boxed{\binom{n}{x}\frac{B(x+\alpha, n+\beta-x)}{B(\alpha,\beta)}} \simeq 0.034$

$\Rightarrow$ $BF_{A,B} \simeq 0.0299$ model B is favored over A

NOTE: Beta function $B(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}dx = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$, $a > 0, b > 0$

Multi Parameter Models

Multiparametric Models: Normal distribution

Non-informative prior

- Prior $\boxed{p(\mu,\sigma^2) = p(\mu)p(\sigma^2) \propto \sigma^{-2}}$, improper $(p(\mu, log\sigma) \propto 1)$
- Likelihood: $p(x|\mu,\sigma^2) \propto \sigma^{-n} exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right)$
- Posterior distribution :

(a) $p(\mu,\sigma^2|x) \propto p(\mu,\sigma^2)p(x|\mu,\sigma^2) \propto \sigma^{-n-2} exp\left(-\frac{1}{2\sigma^2}\left[(n-1)s_c^2 + n(\bar{x}-\mu)^2\right]\right)$

(b) $p(\mu,\sigma^2|x) = p(\mu|\sigma^2,x)p(\sigma^2|x)$, one of the few multipar. prob. simple enough to solve analytically

(b.1) Draw σ^2 from Marg. post. d.

$p(\sigma^2|x) = \int p(\mu,\sigma^2|x)d\mu \propto (\sigma^2)^{-\frac{n+1}{2}} exp\left(-\frac{(n-1)s_c^2}{2\sigma^2}\right) \sim \text{Inv-}\chi^2(n-1, s_c^2)$

i.e., $\sigma^2 = \frac{(n-1)s_c^2}{\chi^2_{n-1}}$

(b.2) Given σ^2, draw μ from Cond. post . d

$p(\mu|\sigma^2,x) \propto p(\mu)p(x|\mu,\sigma^2) \propto exp\left(-\frac{n(\mu-\bar{x})^2}{2\sigma^2}\right) \sim N(\bar{x}, \sigma^2/n)$

Multiparametric Models: Normal distribution

Non-informative prior

- Marginal post. d. of μ (analytically):

$$p(\mu|x) = \int_0^\infty p(\mu,\sigma^2|x)d\sigma^2 \propto \left(1 + \frac{n(\mu-\bar{x})^2}{(n-1)s_c^2}\right)^{-\frac{n}{2}} \sim t_{n-1}(\bar{x}, s^2/n)$$

- Predictive post. d. of

$$p(\tilde{x}|x) = \int\int p(\tilde{x}|\mu,\sigma^2,x)p(\mu,\sigma^2|x)d\mu d\sigma$$

(a) Analytically: $\tilde{x}|x \sim t_{n-1}\left(\bar{x}, \left(1+\frac{1}{n}\right)^{1/2} s\right)$

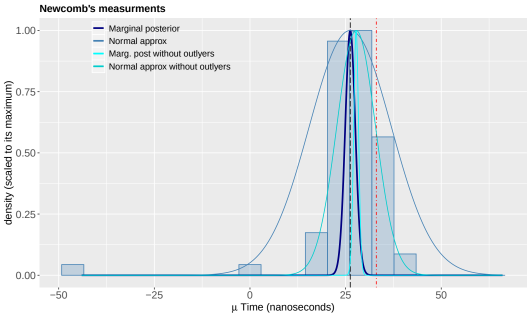
(b) Gral. sampling: (1) Draw μ, σ^2 from joint post. d; (2) Given (μ,σ^2), sample $\tilde{x}$ from $N(\mu,\sigma^2)$

(c) $p(\tilde{x}|\sigma^2,x) = \int p(\tilde{x}|\mu,\sigma^2,x)p(\mu|\sigma^2,x)d\mu \sim N\left(\bar{x},(1+\frac{1}{n})\sigma^2\right)$

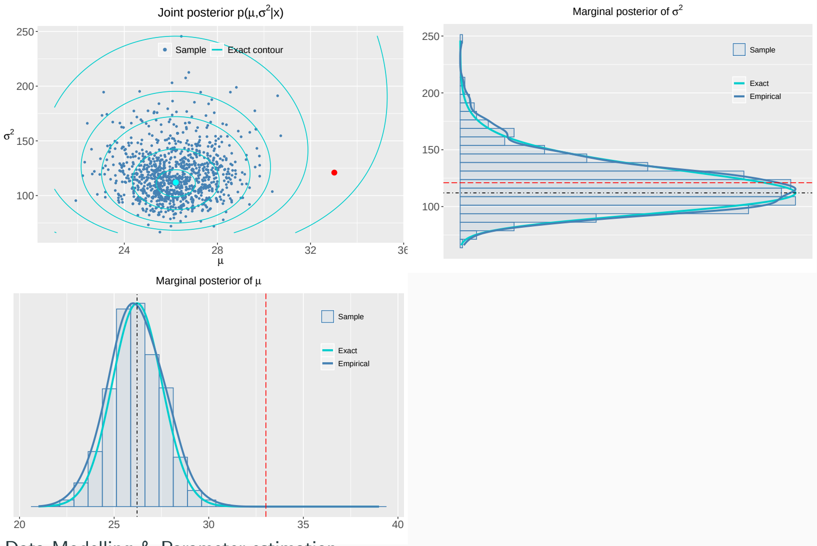
Multiparametric Models: Normal distribution

Example II: Estimating the speed of light

- Simon Newcomb, 1882. Experiment to measure the speed of light. 66 measurements of the time required for light to travel a distance of 7442 m. There are two inusual low measurements.
- We assume a Normal distribution (no the best choice), and indep. measurements: $x_i \sim N(\mu, \sigma^2), \quad i = 1, \ldots, 66$



Example II: Estimating the speed of light

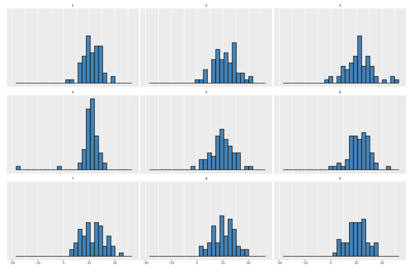


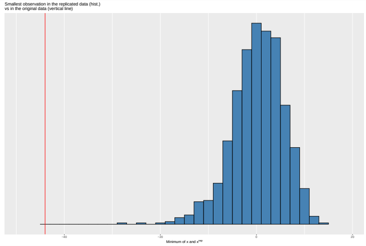
Multiparametric Models: Normal distribution

Posterior predictive checking

- **Self-consistency check**: If the model fits, then replicated data generated under the model ($x^{rep} \sim p(\tilde{x}|x)$) should look similar to observed data x. I.e., obs. data should look plausible under posterior predictive distribution.
- Discrepancy can be due to model misfit or chance. Any systematic differences indicate potencial failings of the model

Example II: Estimating the speed of light

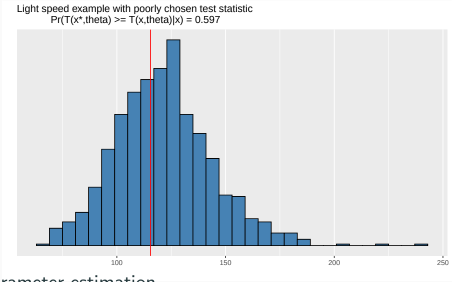




Multiparametric Models: Normal distribution

Posterior predictive checking

- **Test quantities** or **discrepancy measure**: $T(x, \theta)$ measures the discrepancy between the model and data
- **Tail probabilities**. Lack of fit of the data with respect to post. predictive d. can be measured by the tail-area probability, p-value, of the test quantity. $P(T(x^{rep}, \theta) \geq T(x, \theta) \mid x)$ (simulation)
- In practice, we usually compute the post. pred. d by simulation. And p-value is approx. by the proportion of these N simulations s.t. $T(x^{rep,i}, \theta^i) \geq T(x, \theta^i), i = 1, \ldots, N$



Multiparametric Models: Normal distribution

Informative prior

- Likelihood: $p(x|\mu,\sigma^2) \propto \sigma^{-n} exp\left(-\frac{1}{2\sigma^2}\left[(n-1)s_c^2 + n(\bar{x}-\mu)^2\right]\right)$
- Conjugated Prior: $p(\mu,\sigma^2) = p(\sigma^2)p(\mu|\sigma^2)$, where $\sigma^2 \sim \chi^2(\nu_0,\sigma_0^2)$ and $\mu|\sigma^2 \sim N(\mu_0, \sigma^2/\kappa_0)$ $\Rightarrow p(\mu,\sigma^2) \propto \sigma^{-1}(\sigma^2)^{-\nu_0/2+1} exp\left(-\frac{1}{2\sigma^2}\left[\nu_0\sigma_0^2 + \kappa_0(\mu_0-\mu)^2\right]\right)$
  $(\mu,\sigma^2) \sim \text{N-Inv-}\chi^2\left(\overbrace{\underbrace{\mu_0}_{local.}, \underbrace{\sigma_0^2/\kappa_0}_{scale}}^{\mu}; \overbrace{\underbrace{\nu_0}_{dof}, \underbrace{\sigma_0^2}_{scale}}^{\sigma}\right)$
- Joint posterior distribution :

$$p(\mu,\sigma^2|x) \propto \sigma^{-1}(\sigma^2)^{-\nu_n/2+1} exp\left(-\frac{1}{2\sigma^2}\left[\nu_n\sigma_n^2 + \kappa_n(\mu_n-\mu)^2\right]\right)$$

where

$$\begin{array}{lllllll} \mu_n & = & \frac{\kappa_0}{\kappa_0+n}\mu_0 + \frac{n}{\kappa_0+n}\bar{x} & & \kappa_n & = & \kappa_0 + n \\ \nu_n & = & \nu_0 + n & & \nu_n\sigma_n^2 & = & \nu_0\sigma_0^2 + (n-1)s^2 + \frac{\kappa_0 n}{\kappa_0+n}(\bar{x}+n)^2 \end{array}$$

$$(\mu,\sigma^2|x) \sim \text{N-Inv-}\chi^2\left(\mu_n, \sigma_n^2/\kappa_n; \nu_n, \sigma_n^2\right)$$

Multiparametric Models: Normal distribution

Informative prior

- Marginal posteriors
  $p(\sigma^2|x) = \int p(\mu, \sigma^2|x) d\mu \propto (\sigma^2)^{-\nu_n/2+1} exp\left(-\frac{\nu_n \sigma_n^2}{2\sigma^2}\right) \sim \text{Inv-}\chi^2(\nu_n, \sigma_n^2)$
  $p(\mu|x) = \int p(\mu, \sigma^2|x) d\sigma^2 \propto exp\left(1 + \frac{\kappa_n(\mu_n - \mu)^2}{\nu_n \sigma_n^2}\right)^{-(\nu_n+1)/2} \sim t_{\nu_n}\left(\mu_n, \sigma_n^2/\kappa_n\right)$
- Condicional post. d of μ, given σ^2

$$p(\mu|\sigma^2, x) \quad \propto \quad exp\left(-\frac{1}{2\sigma^2}\kappa_n(\mu_n - \mu)^2\right) \sim N\left(\mu_n, \sigma^2/\kappa_n\right)$$

- Sampling from the joint posterior distribution

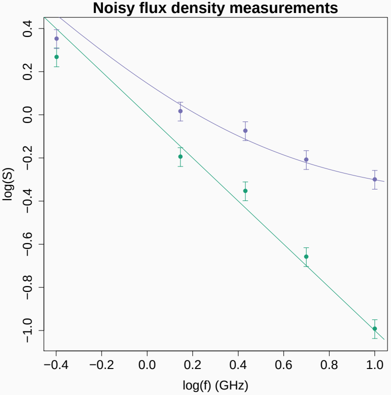
$$p(\mu, \sigma^2|x) = p(\mu|\sigma^2, x) p(\sigma^2|x) \quad \begin{cases} 1) & \sigma^2|x \sim \text{Inv-}\chi^2(\nu_n, \sigma_n^2) \\ 2) & \mu|\sigma^2, x \sim N\left(\mu_n, \sigma^2/\kappa_n\right) \end{cases}$$

Multiparametric Models

Example III: Radio-source spectra

We have noisy flux density measurements, S_i, at different frequencies f_i (green). Assume these follows a power law of slope -1, but have a $\epsilon = 10\%$ Gaussian noise. In purple, same data but with an offset error of 0.4 units

- Model A: $S = \kappa f^{-\gamma}$
- Model B: $S = \beta + \kappa f^{-\gamma}$
- Data: $\boldsymbol{x} \equiv \{x_1, \ldots, x_n\}$, where $x_i = (f_i, S_i)$



Multiparametric Models

Example: Radio-source spectra, Model A

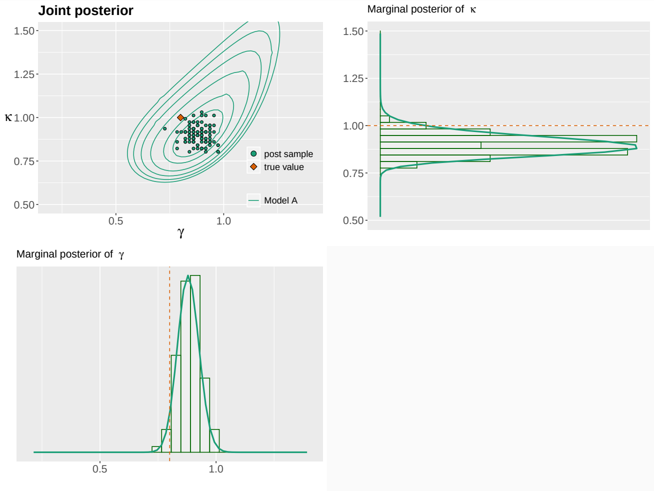
- Model A: $f(x_i|\kappa,\gamma) = \frac{1}{\sqrt{2\pi}\epsilon\kappa f_i^{-\gamma}} exp\left(-\frac{(S_i-\kappa f_i^{-\gamma})^2}{2(\epsilon\kappa f_i^{-\gamma})^2}\right)$
- Prior: $p(\kappa,\gamma) \propto 1$
- Likelihood: $\mathcal{L}(\kappa,\gamma) = f(\boldsymbol{x}|\kappa,\gamma) = \prod_{i=1}^{n} f(x_i|\kappa,\gamma)$
- Joint posterior:

$$p(\kappa,\gamma|\boldsymbol{x}) \propto (\sqrt{2\pi}\epsilon\kappa)^{-n}\left(\prod_{i=1}^{n} f_i\right)^{\gamma} exp\left(-\sum_{i=1}^{n}\frac{(S_i-\kappa f_i^{-\gamma})^2}{2(\epsilon\kappa f_i^{-\gamma})^2}\right)$$

- Marginal posteriors : marginalize the nuisance params. out
  - $p(\kappa|\boldsymbol{x}) = \int p(\kappa,\gamma|\boldsymbol{x})d\gamma$
  - $p(\gamma|\boldsymbol{x}) = \int p(\kappa,\gamma|\boldsymbol{x})d\kappa$

Multiparametric Models

Example: Radio-source spectra, Model A



Multiparametric Models

Example: Radio-source spectra, Model B

- Model B: $f(xi|\kappa,\gamma,\beta) = \frac{1}{\sqrt{2\pi}\epsilon\kappa f_i^{-\gamma}} exp\left(-\frac{(S_i-\beta-\kappa f_i^{-\gamma})^2}{2(\epsilon\kappa f_i^{-\gamma})^2}\right)$
- Prior: $p(\kappa,\gamma,\beta) \propto \frac{1}{\sqrt{2\pi}\epsilon} exp\left(-\frac{(\beta-\mu_\beta)^2}{2\epsilon^2}\right)$, known μ_β and ϵ
- Likelihood: $L(\kappa,\gamma,\beta) = (\sqrt{2\pi}\epsilon\kappa)^{-n}\left(\prod_{i=1}^n f_i\right)^\gamma exp\left(-\sum_{i=1}^n \frac{(S_i-\kappa f_i^{-\gamma}-\beta)^2}{2(\epsilon\kappa f_i^{-\gamma})^2}\right)$
- Joint posterior:

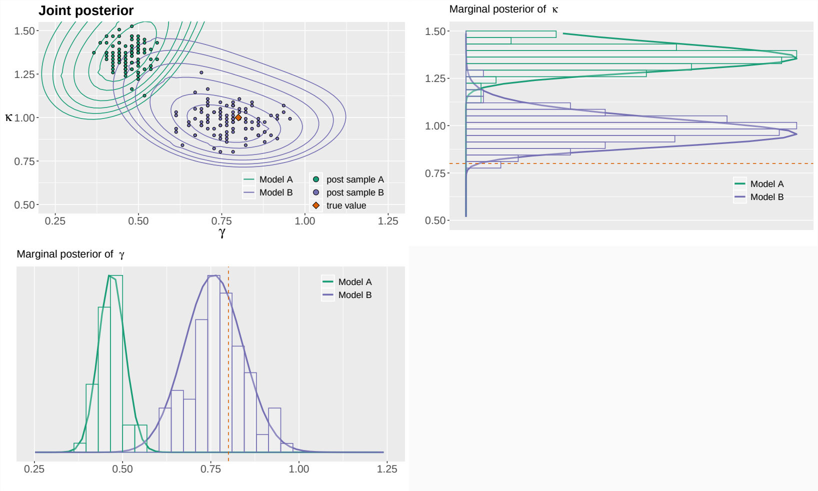
$$p(\kappa,\gamma,\beta|D) \propto (\sqrt{2\pi}\epsilon)^{-n-1}\kappa^{-n}\left(\prod_{i=1}^n f_i\right)^\gamma exp\underbrace{\left(-\frac{(\beta-\mu_\beta)^2}{2\epsilon^2}-\sum_{i=1}^n \frac{(S_i-\kappa f_i^{-\gamma}-\beta)^2}{2(\epsilon\kappa f_i^{-\gamma})^2}\right)}_{-A\beta^2+B\beta+C}$$

$$A = \frac{1}{2\epsilon}+\sum_{i=1}^n \frac{\kappa^{-2}f_i^{2\gamma}}{2\epsilon^2};\quad B = \frac{\mu_\beta}{\epsilon^2}+\sum_{i=1}^n \frac{S_i-\kappa f_i^{-\gamma}}{\epsilon^2\kappa^{-2}f_i^{-2\gamma}};\quad C = -\frac{\mu_\beta^2}{2\epsilon^2}-\sum_{i=1}^n \frac{(S_i-\kappa f_i^{-\gamma})^2}{\epsilon^2\kappa^{-2}f_i^{-2\gamma}}$$

- Marginal posterior

$$p(\kappa,\gamma|D) = \int_{-\infty}^{\infty} p(\kappa,\gamma,\beta|D)d\beta \propto (\sqrt{2\pi}\epsilon)^{-n-1}\kappa^{-n}\left(\prod_{i=1}^n f_i\right)^\gamma \sqrt{\frac{\pi}{A}}e^{\frac{B^2}{4A}+C}$$

Example: Radio-source spectra, Model B



Bayesian Computation

Bayesian Computation

- The quintessential objective of Bayesian analysis is the posterior distribution of the parameters,

$$p(\theta|x) \propto p(\theta)p(x|\theta) \qquad log\ p(\theta|x) \propto log\ p(\theta) + \sum_{i=1}^{n} log\ p(x_i|\theta)$$

  $\theta = (\theta_1, \ldots, \theta_k)$ unknown, with prior d $p(\theta)$
  Data $x = (x_1, \ldots, x_n)$, iid $p(x|\theta) = \prod_{i=1}^{n} p(x_i|\theta)$

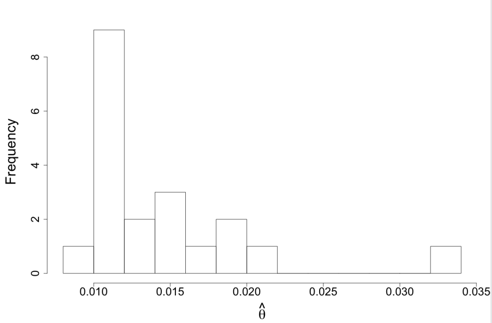
- In general, the main problems are
  1. **Draw samples from the posterior d.** : Rejection sampling, MCMC algorithms (Gibbs & Metropolis-Hastings)
  2. **Compute integrals with respect to posterior d** : Monte Carlo integration, Importance sampling

Bayesian Computation

Example I(bis): Estimate the AGN fraction

It is observed $m = 20$ sample of n_i galaxies, where $x_i =$ No. of AGN

Data: $\boldsymbol{x} = \{(x_i, n_i)\}_{i=1,\ldots,20}$



$$\left.\begin{array}{l} x_i \sim B(n_i, \theta), \\ \theta \sim Beta(\alpha, \beta), \end{array}\right\} \quad \theta|x \sim Beta\left(\alpha + \sum x_i, \beta + \sum(n_i - x_i)\right)$$

Bayesian Computation

Beta-binomial Model

$$\left.\begin{array}{l} \bullet\ x_i \sim B(n_i, \theta), \\ \quad p(x|\theta) = \prod_{i=1}^m \binom{n}{x} \theta^{x_i}(1-\theta)^{n_i - x_i} \\ \bullet\ \theta \sim Beta(\alpha, \beta), \\ \quad p(\theta|\alpha,\beta) = \underbrace{\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}}_{B(\alpha,\beta)^{-1}} \theta^{\alpha-1}(1-\theta)^{\beta-1} \end{array}\right\} \theta|x \sim Beta\left(\alpha + \sum x_i, \beta + \sum(n_i - x_i)\right)$$

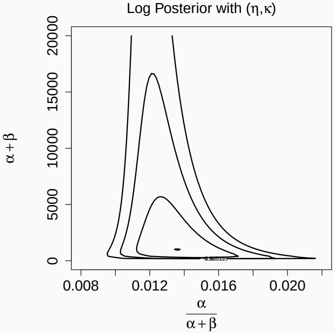
- $x_i \sim Beta - bin(n, \alpha, \beta)$, i.e. $p(x_i|\alpha,\beta) = \int_0^1 p(x_i|\theta)p(\theta)d\theta$

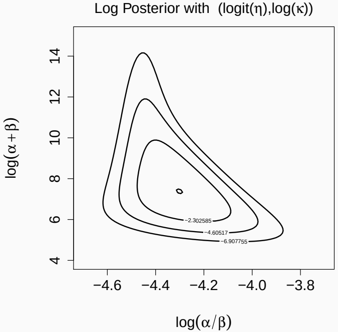
  $$p(x_i|\alpha,\beta) = \frac{\binom{n_i}{x_i}}{B(\alpha,\beta)} \int_0^1 \theta^{x_i+\alpha-1}(1-\theta)^{n_i - x_i + \beta - 1} d\theta = \binom{n_i}{x_i} \frac{B(x_i+\alpha, n_i - x_i + \beta)}{B(\alpha,\beta)}$$

  where $\eta = \frac{\alpha}{\alpha+\beta} = E(\theta) \in (0,1)$, and $\kappa = \alpha + \beta > 0$ 'prior sample size'
- $E_{B\text{-}bin}(x) = n\frac{\alpha}{\alpha+\beta} = n\eta = E_{bin}(x)$
- $Var_{B\text{-}bin}(x) = \frac{n\alpha\beta(\alpha+\beta+n)}{(\alpha+\beta)^2(\alpha+\beta+1)} = n\eta(1-\eta)\frac{\kappa+n}{\kappa+1} > Var_{bin}(x)$ (over dispersion)

Example I(bis): Estimate the AGN fraction

- Non-inform uniform prior d. to prior mean and variance* :
  $p(\eta,\kappa) \propto \frac{1}{\eta(1-\eta)}\frac{1}{(1+\kappa)^2}$ $\quad \left({}^{*}p(logit(\eta), \frac{1}{\kappa+1}) \propto 1\right)$
- $p(\eta,\kappa|x) \propto \frac{1}{\eta(1-\eta)}\frac{1}{(1+\kappa)^2}\prod_{i=1}^{20}\frac{B(\kappa\eta+x_i,\kappa(1-\eta)+n_i-x_i)}{B(\kappa\eta,\kappa(1-\eta))}$ $\quad$ (*proper post. d.)
- $(\phi_1,\phi_2) = (logit(\eta), log(\kappa)) : p(\phi_1,\phi_2|x) \propto p_{\eta,\kappa}\left(\frac{e^{\phi_1}}{1+e^{\phi_1}}, e^{\phi_2}\middle| x\right)\frac{e^{\phi_1+\phi_2}}{(1+e^{\phi_1})^2}$





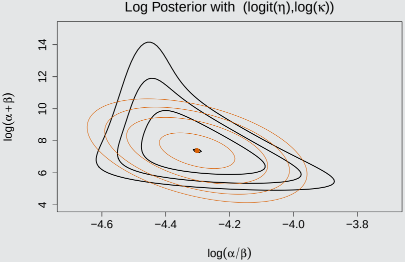
Bayesian Computation

Approximations based on Posterior Modes

- Method of summarizing a multivariate post. d. $p(\theta|x)$, based on behavior of density about its mode
- Let $h(\theta) = log\,(p(\theta)p(x|\theta))$, and $\hat{\theta} = M_o(\theta|x)$. 2nd order Taylor's series: $h(\theta) \approx h(\hat{\theta}) + \frac{1}{2}(\theta - \hat{\theta})^T h''(\hat{\theta})(\theta - \hat{\theta}) \rightarrow \theta|x \sim N(\hat{\theta}, (-h''(\hat{\theta}))^{-1})$
- To find $\hat{\theta}$: Newton's Method, Nelder-Mead's Algorithm (`laplace`)

Example I(bis): Estimate the AGN fraction

- $\phi^0 = (-4.3, 7.3)$: $\hat{\phi} = (-4.30, 7.38)$,
  $\Sigma = \begin{pmatrix} 0.008 & -0.032 \\ -0.032 & 0.671 \end{pmatrix}$
- $PI_{90\%}(logit\eta) = (-4.45, -4.16)$,
  $PI_{90\%}(log\kappa) = (6.03, 8.73)$
- $\hat{\eta} = E(\theta) = 0.01336$;
  $(\hat{\alpha}, \hat{\beta}) = (21.46, 1584.41)$



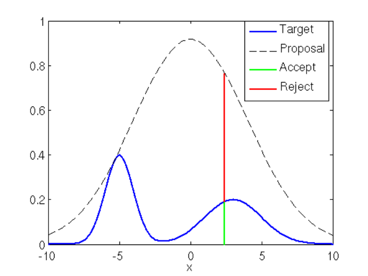
Bayesian Computation

Produce simulated samples from a given post. d $\mathbf{p}(\theta|\mathbf{x})$ (unfamiliar func. form), where the normalizing cte. may not be known

Rejection Sampling

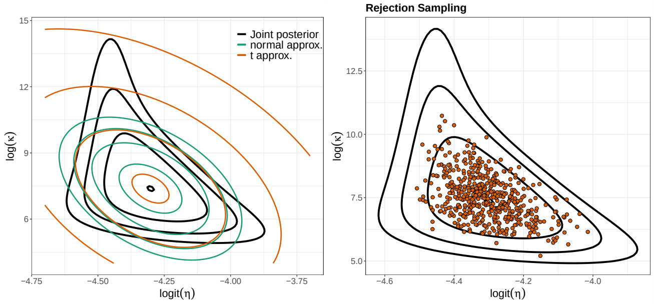
- To find a proposal d. $\tilde{p}(\theta)$ s.t.: $\begin{cases} \text{easy-to-sample PDF} \\ \text{resembles the post. d. (location and spread)} \\ \exists c : p(\theta|x) \leq c\tilde{p}(\theta) \quad \forall\theta \end{cases}$
- Obtain draws from $p(\theta|x)$ using the following accept/reject algorithm:

1. Independently simulate: $u \sim U[0,1]$, and θ_i from $\tilde{p}(\theta)$.
2. If $u \leq \frac{p(\theta|x)}{c\tilde{p}(\theta)}$ accept θ_i; otherwise reject it
3. Repeat 1-2 until suff. sample size is reached: $\{\theta_1, \ldots, \theta_n\}$



Example I(bis): Estimate the AGN fraction

- Proposal distribution on $(\phi_1, \phi_2) = (logit(\eta), log(\kappa))$:
  $\tilde{p}(\phi) = t_{\nu=4}\left(\phi \middle| \mu = \hat{\phi} = (-4.30, 7.38), S = 2\Sigma\right)$
- $p(\phi|x) \leq c\tilde{p}(\phi), \forall\phi \Leftrightarrow log(c) \approx \max_\phi logp(\phi|x) - log\,\tilde{p}(\phi)$
- $E(\phi|x) \simeq (-4.3051, 7.513) \pm (0.0039, 0.0388)$ (by Monte Carlo approx.)
- $\hat{\eta} = E(\theta) = 0.0133$; $(\hat{\alpha}, \hat{\beta}) = (24.4, 1807.39)$



Bayesian Computation

Produce simulated samples from a given post. d $\mathbf{p}(\theta|\mathbf{x})$ (unfamiliar func. form), where the normalizing cte. may not be known

Importance Sampling

Given some function of the params, $h(\theta)$ (i.e. post. mean),

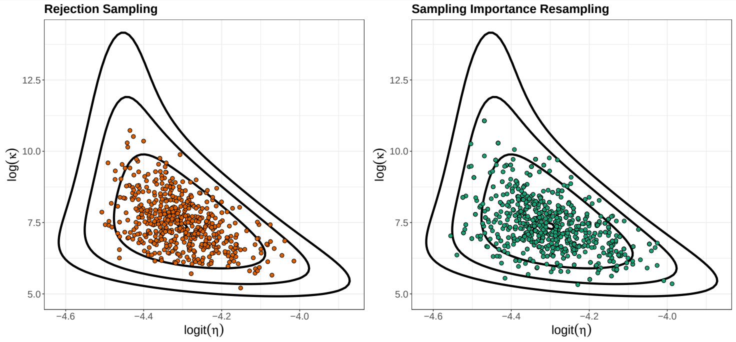
$$E(h(\theta)|x) = \int h(\theta)p(\theta|x)d\theta = \frac{\int h(\theta)p(\theta)p(x|\theta)d\theta}{\int p(\theta)p(x|\theta)d\theta} = \frac{\int h(\theta)\omega(\theta)\tilde{p}(\theta)d\theta}{\int \omega(\theta)\tilde{p}(\theta)d\theta}$$

- Simulate $\{\theta_k\}_{k=1,\ldots,m}$ from $p(\theta|x)$, $\boxed{\bar{h} = \frac{\sum_{k=1}^m h(\theta_k)}{m}}$ $\pm se_{\bar{h}} = \sqrt{\frac{\sum_{k=1}^m (h(\theta_k)-\bar{h})^2}{(m-1)m}}$
- To find proposal $\tilde{p}(\theta)$ $\begin{cases} \text{easy-to-sample PDF} \\ \text{resembles the post.} \\ \text{relatively flat tails} \end{cases}$, $\omega(\theta) = \frac{p(\theta)p(x|\theta)}{\tilde{p}(\theta)}$ weight f.

  1. Simulate $\{\theta_k\}_{k=1,\ldots,m}$ from $\tilde{p}(\theta)$
  2. Imp. Sam. Estimate $\boxed{\bar{h}_{SI} = \frac{\sum_{k=1}^m h(\theta_k)\omega(\theta_k)}{\sum_{k=1}^m \omega(\theta_k)}}$ $\pm se_{\bar{h}_{SI}} = \frac{\sqrt{\sum_{k=1}^m (h(\theta_k)-\bar{h}_{SI})^2\omega(\theta_k)}}{\sum_{k=1}^m \omega(\theta_k)}$
  3. Sampling Importance Resampling: Take new $\{\theta_j^*\}_j$ from discrete distr. over $\{\theta_k\}_k$, with resp. prob. $p_k = \frac{\omega(\theta_k)}{\sum_{k=1}^m \omega(\theta_k)}$ $\left(\{\theta_j^*\}_j \approx p(\theta|x)\right)$

Example I(bis): Estimate the AGN fraction

- Proposal distribution on $(\phi_1, \phi_2) = (logit(\eta), log(\kappa))$:
  $\tilde{p}(\phi) = t_{\nu=4}\left(\phi \,\middle|\, \mu = \hat{\phi} = (-4.30, 7.38), S = 2\Sigma\right)$
- $E(\phi|x) \simeq (-4.3048, 7.4420) \pm (0.00297, 0.0304)$ (by Monte Carlo approx.)
- $\hat{\eta} = E(\theta) = 0.0133$; $(\hat{\alpha}, \hat{\beta}) = (22.73, 1683.47)$



Bayesian Computation: Markov Chain Monte Carlo (MCMC)

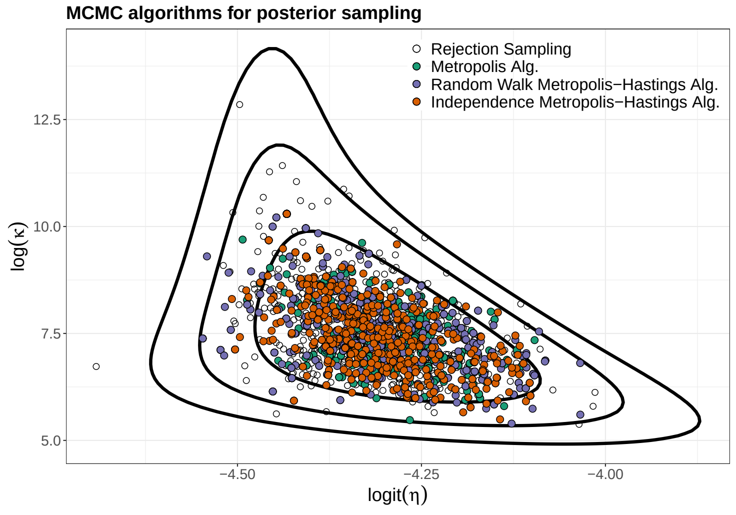
- Algorithms for summarizing the posterior distribution
- RS, IS, SIR algs. are general-purpose methods for simulating an arbitrary post. d. Requires the construction of a suitable proposal density, that may be difficult to find for high-dim problems.
- MCMC algs. are attractive: easy to set up and program, and little prior input from the user
- Sampling strategy sets up an irreducible, aperiodic Markov Chain [sequence of random vars. $\{\theta^t\}_{t=1,2,\ldots}$, s.t. $p(\theta^t|\theta^1,\ldots,\theta^{t-1}) = p(\theta^t|\theta^{t-1}), \forall t$] for which the stationary distribution equals the posterior d.
- Basic Markov Chain simulation methods: **Metropolis-Hastings** & **Gibbs sampling**

Bayesian Computation: Markov Chain Monte Carlo (MCMC)

Metropolis-Hastings Algorithm

- Given $\tilde{p}(\theta)$, *proposal*, *jumping* or *jumping* distribution, (easy-to-sample pdf, and approx. the target d.), and starting point θ^0 (crude approx. estimate), for $t = 1, 2, \ldots$
  1. Sample θ^* from $\tilde{p}(\theta^*|\theta^{t-1})$ (*transition kernel*)
  2. Compute the ratio $R = \frac{p(\theta^*|x)\tilde{p}(\theta^{t-1}|\theta^*)}{p(\theta^{t-1}|x)\tilde{p}(\theta^*|\theta^{t-1})}$
  3. Set $\theta^t = \begin{cases} \theta^* & \text{with prob. } P = \min\{R, 1\} \\ \theta^{t-1} & \text{otherwise} \end{cases}$
  4. Repeat steps $1-3$, up to desired sample size. Eliminate the first simulations to make the result independent of the choice of θ^0
- Metropolis Alg.: $\tilde{p}(\theta)$ symmetric, $\tilde{p}(\theta^t|\theta^{t-1}) = \tilde{p}(\theta^{t-1}|\theta^t); \rightarrow R = \frac{p(\theta^*|x)}{p(\theta^{t-1}|x)}$
- Independence Chain: $\tilde{p}(\theta^*|\theta^{t-1}) = \tilde{p}(\theta^*)$
- Random Walk Chain: $\tilde{p}(\theta^*|\theta^{t-1}) = h(\theta^* - \theta^{t-1})$, h-symmetric d. about the origin

Example I(bis): Estimate the AGN fraction



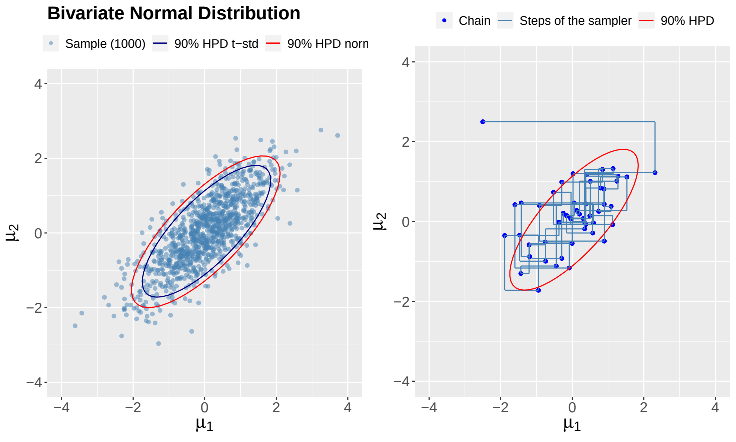
Gibss Sampling

- Given the param. vector $\theta = (\theta_1, \ldots, \theta_p)$, $p(\theta|x)$ may be of high-dimension and difficult to summarize.
- We can set up a Markov-Chain simulation alg. for the joint post d by succesfully simulating individual params. from the set of p-cond. distr.
- Given an initial param., $\theta^0 = (\theta_1^0, \ldots, \theta_p^0)$,
    - for $t = 1, 2, \ldots$

$$\left.\begin{array}{l} \theta_1^t \sim p(\theta_1|x, \theta_2^{t-1}, \ldots, \theta_p^{t-1}) \\ \vdots \\ \theta_j^t \sim p(\theta_1|x, \theta_1^t, \ldots, \theta_{j-1}^t, \theta_{j+1}^{t-1}, \ldots, \theta_p^{t-1}) \\ \vdots \\ \theta_p^t \sim p(\theta_1|x, \theta_1^t, \ldots, \theta_{p-1}^t) \end{array}\right\} \text{one-cycle of Gibbs sampling}$$

    - Eliminate first simulations to make indp. the results from the initial choice of the params.
- Metropolis within Gibbs: when it is not convenient/possible to sample directly from the cond. distr., one can use a Metropolis Alg. to simulate

Example: Gibbs sampling



Hierarchical Models

Hierarchical Models

- Multi-parameters models, related or connected in some way by the estructure of the problem $\rightarrow$ joint prob.d. should reflect their dependence
- θ_j's viewed as a sample from a common population distrb.:

$$\underset{hyper-param}{\Phi} \longrightarrow \underset{param}{\Theta} \longrightarrow \underset{obs}{x}$$

- ϕ unknown, and thus has its own prior d., **hyperprior distr.** $p(\phi)$
- Exchangeability: No information, other than data, available to distinguish between θ_j's, and no ordering or grouping

$$p(\theta \mid \phi) = \prod_{j=1}^{J} p(\theta_j \mid \phi) \xrightarrow[\text{as } J\rightarrow\infty]{\text{de Finetti's,}} p(\theta) = \int \left(\prod_{j=1}^{J} p(\theta_j \mid \phi)\right) p(\phi) d\phi$$

- **Joint prior** d: $p(\theta, \phi) = p(\phi) p(\theta \mid \phi)$
- **Joint posterior** d:

$$p(\theta, \phi \mid x) \propto p(\theta, \phi) p(x \mid \theta, \phi) = p(\phi) p(\theta \mid \phi) p(x \mid \theta)$$

Hierarchical Models

Drawing simulations from the joint posterior distribution

1. Draw ϕ from its **marginal post.** d.,

$$\begin{aligned} p(\phi \mid x) &= \int p(\phi, \theta \mid x) d\theta, \quad \text{integrating over } \theta \\ &= \frac{p(\phi, \theta \mid x)}{p(\theta \mid \phi, x)}, \quad \text{or algebraically (conjugated HM)} \end{aligned}$$

2. Draw θ from its **conditional post.** d., given the drawn value ϕ, for fixed obs. x (analiticaly or MCMC):

$$p(\theta \mid \phi, x) = \prod_{j=1}^{J} p(\theta_j \mid \phi, x) \longrightarrow \theta_j \sim p(\theta_j \mid \phi, x)$$

3. If desired, draw predictive values $\tilde{x}$ from **posterior predictive d.**, corresponding to an existing θ_j, or a future $\tilde{\theta}_j$ drawn from the same super population.

Example I(bis): Estimate the AGN fraction

- It is observed 20 samples of n_i galaxies, where $x_i =$ No. of AGN, and θ_i prob. of being an AGN galaxy, $i = 1, \ldots, 20$.
- $\left.\begin{array}{l} x_i \sim B(n_i, \theta_i) \text{ iid,} \\ \theta_i \sim Beta(\alpha, \beta) \text{ iid, unknown } (\alpha, \beta) \end{array}\right\} \rightarrow \theta_i | x_i \sim Beta(\alpha + x_i, \beta + n_i - x_i)$

  (i) Puntual estimate, $\hat{\alpha}, \hat{\beta}$: $E(\theta) = \frac{\alpha}{\alpha+\beta} \simeq \bar{\theta}$; $\quad Var(\theta) = \frac{E(\theta)(1-E(\theta))}{\alpha+\beta+1} \simeq S^2_\theta$

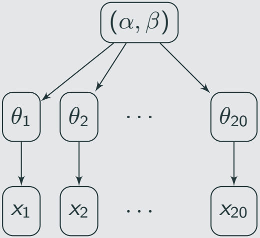
  (ii) Full Bayesian treatment of the Hierarchical model:

  → Non informative hyper-prior, $p(\alpha, \beta)$

  → Joint post. d, $p(\theta, \alpha, \beta \mid x)$

  → Marginal post. d, $p(\alpha, \beta \mid x)$

  → Cond. post. d, $p(\theta \mid \alpha, \beta, x)$



Hierarchical Models

Example I(bis): Estimate the AGN fraction

- Joint posterior d.

$$
\begin{aligned}
p(\alpha,\beta,\theta \mid x) \quad &\propto \quad \underbrace{p(\alpha,\beta)}_{\text{hiperprior}}\underbrace{p(\theta \mid \alpha,\beta)}_{Beta(\alpha,\beta)}\underbrace{p(x \mid \theta,\alpha,\beta)}_{Bin(n,\theta)} \\
&= \quad p(\alpha,\beta)\prod_{i=1}^{20}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta_i^{\alpha-1}(1-\theta_i)^{\beta-1}\prod_{i=1}^{20}\binom{n_i}{x_i}\theta_i^{x_i}(1-\theta_i)^{n_i-x_i}
\end{aligned}
$$

- Hyperprior d. selection:
  - Re-parameterize to $\mathcal{R}$ scale, $\left(logit\left(\frac{\alpha}{\alpha+\beta}\right) = log\left(\frac{\alpha}{\beta}\right), log(\alpha+\beta)\right)$ But uniform prior d. for these param. leads to improper post. d * (!!)
  - $p\left(\frac{\alpha}{\alpha+\beta},(\alpha+\beta)^{-\frac{1}{2}}\right) \propto 1 \overset{**}{\Rightarrow} \begin{cases} p(\alpha,\beta) \propto (\alpha+\beta)^{-5/2} \text{ (improper)} \\ p\left(log\left(\frac{\alpha}{\beta}\right), log(\alpha+\beta)\right) \propto \alpha\beta(\alpha+\beta)^{-5/2} \end{cases}$

* *General problem in HM when uniform priors for the log of std. dev. of the exchangeable params, results in improper post. d. To avoid impropriety, assign unif. prior to std. dev. itself, rather than its log*
** *Transformation of variable: if* $v = f(u), \rightarrow p_u(f^{-1}(v))\left|\frac{dv}{du}\right|$

Hierarchical Models

Example I(bis): Estimate the AGN fraction

- **Conditional post.** d. of θ, given (α, β) and fixed obs. x:

$$p(\theta_i \mid \alpha, \beta, x) \propto p(\theta_i \mid \alpha, \beta) p(x \mid \theta_j) \propto \theta_i^{\alpha-1}(1-\theta_i)^{\beta-1}\theta_i^{x_i}(1-\theta_i)^{n_i-x_i} \sim Beta(\alpha + x_i, \beta + n_i - x_i)$$

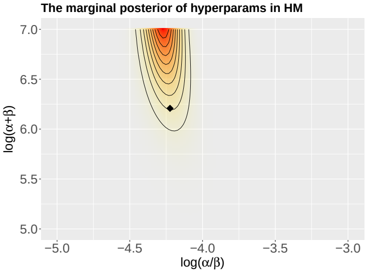
$$\mathbf{p}(\boldsymbol{\theta} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{x}) = \prod_{i=1}^{20} p(\theta_i \mid \alpha, \beta, x) = \prod_{i=1}^{20} \frac{\Gamma(\alpha+\beta+n_i)}{\Gamma(\alpha+x_i)\Gamma(\beta+n_i-x_i)} \theta_i^{\alpha+x_i-1}(1-\theta_i)^{\beta+n_i-x_i-1}$$

- **Marginal post. d.** of (α, β):

$$\mathbf{p}(\boldsymbol{\alpha}, \boldsymbol{\beta} \mid \mathbf{x}) = \frac{p(\alpha,\beta,\theta|x)}{p(\theta|\alpha,\beta,x)} \propto (\alpha+\beta)^{-5/2} \prod_{i=1}^{I} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+x_i)\Gamma(\beta+n_i-x_i)}{\Gamma(\alpha+\beta+n_i)}$$

Initial approx. $E(\theta) \simeq \bar{\theta}$, $Var(\theta) \simeq S_\theta^2$

$$\begin{cases} (\alpha_0, \beta_0) = (7.17, 489.58) \\ (log(\frac{\alpha_0}{\beta_0}), log(\alpha_0+\beta_0)) = (-4.22, 6.21) \pm 3dex \end{cases}$$

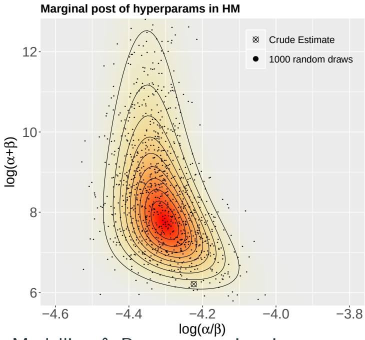


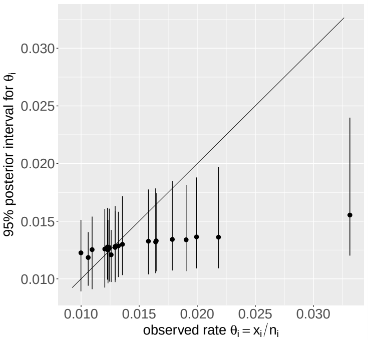
Example I(bis): Estimate the AGN fraction

Posterior moments,

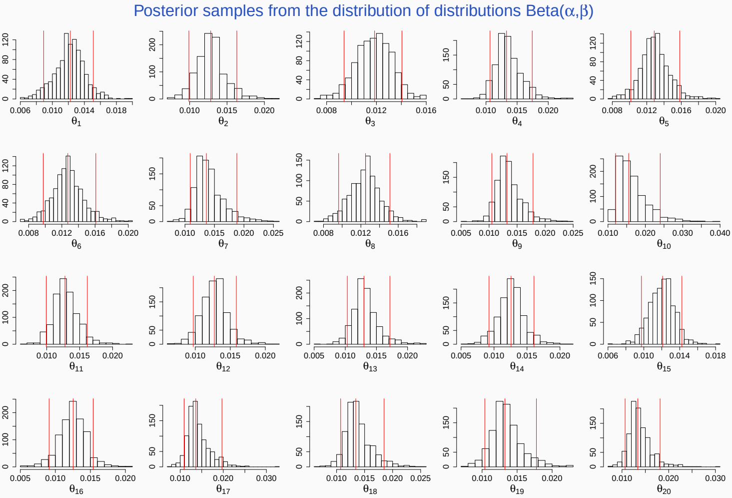
$$E(\alpha \mid x) \simeq \sum_{m,n} \alpha_m p\left(log\left(\frac{\alpha_m}{\beta_n}\right), log(\alpha_m + \beta_n) \mid x\right) = 2.4$$

$$E(\beta \mid x) \simeq \sum_{m,n} \beta_n p\left(log\left(\frac{\alpha_m}{\beta_n}\right), log(\alpha_m + \beta_n) \mid x\right) = 14.3$$

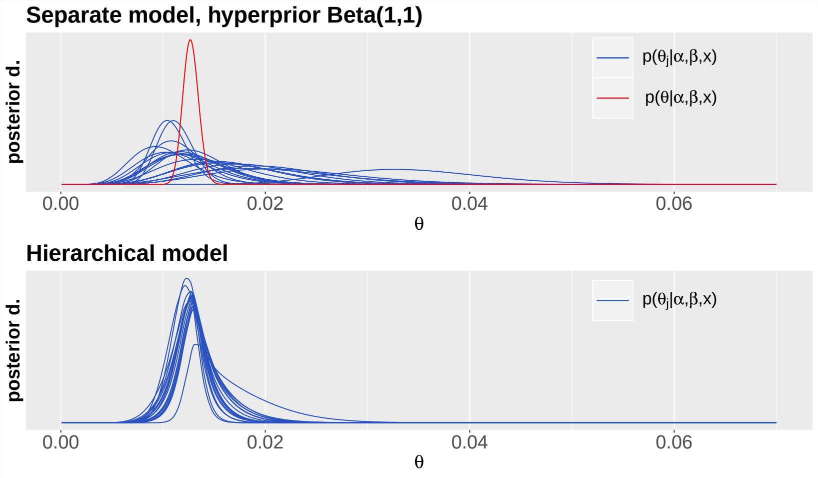




Example I(bis): Estimate the AGN fraction



Example I(bis): Estimate the AGN fraction



Hierarchical Models

Example: Radio-source spectra

- $D = \{(f_i, S_i)\}_{i=1,\dots,n}$, $f(f_i, S_i|\kappa, \gamma, \beta) = \frac{1}{\sqrt{2\pi}\epsilon\kappa f_i^{-\gamma}} exp\left(-\frac{(S_i - \kappa f_i^{-\gamma} - \beta)^2}{2(\epsilon\kappa f_i^{-\gamma})^2}\right)$
- Prior: $p(\kappa, \gamma, \beta) \propto \frac{1}{\sqrt{2\pi}\epsilon} exp\left(-\frac{(\beta - \mu_\beta)^2}{2\epsilon^2}\right)$, Hyper-prior: $p(\mu_\beta) \propto 1$
- Likelihood:

$$\mathcal{L}(\kappa, \gamma, \beta) = (\sqrt{2\pi}\epsilon\kappa)^{-n}\left(\prod_{i=1}^{n} f_i\right)^{\gamma} exp\underbrace{\left(-\sum_{i=1}^{n}\frac{(S_i - \kappa f_i^{-\gamma} - \beta)^2}{2(\epsilon\kappa f_i^{-\gamma})^2}\right)}_{-A\beta^2 + B\beta + C}$$

- Joint posterior:

$$p(\kappa, \gamma, \beta; \mu_\beta|D) = p(\kappa, \gamma, \beta)p(\mu_\beta)\mathcal{L}(\kappa, \gamma, \beta) \propto \frac{1}{\sqrt{2\pi}\epsilon} exp\left(-\frac{(\beta - \mu_\beta)^2}{2\epsilon^2}\right)\mathcal{L}(\kappa, \gamma, \beta)$$

- Marginal posteriors
  - $p(\kappa, \gamma, \beta|D) = \int_{-\infty}^{\infty} p(\kappa, \gamma, \beta; \mu_\beta|D) d\mu_\beta \propto \mathcal{L}(\kappa, \gamma, \beta)$
  - $p(\kappa, \gamma|D) = \int_{-\infty}^{\infty} p(\kappa, \gamma, \beta|D) d\beta \propto (\sqrt{2\pi}\epsilon\kappa)^{-n}\left(\prod_{i=1}^{n} f_i\right)^{\gamma}\sqrt{\frac{\pi}{A}} e^{\frac{B^2}{4A} + C}$

$$A = \sum^{n} \frac{\kappa^{-2} f_i^{2\gamma}}{2\epsilon^2}; \quad B = \sum^{n} \frac{S_i - \kappa f_i^{-\gamma}}{\epsilon^2\kappa^{-2} f_i^{-2\gamma}}; \quad C = -\sum^{n} \frac{(S_i - \kappa f_i^{-\gamma})^2}{\epsilon^2\kappa^{-2} f_i^{-2\gamma}}$$