a.y. 2024/2025

“STATISTICAL METHODS FOR DATA SCIENCE & LABORATORY II”

Luca Tardella [1]

April 4, 2025

[1] Sapienza Università di Roma, Italy

SMDS-II

Instructor: Luca Tardella ⇐

Time Table:
Friday 9:00–12:00 (Lecture Room 201 – RM112)

Office hours:
Thursday 14:00–15:00 ⟹ (email: luca.tardella@uniroma1.it)
check last-minute updates on https://lucatardella.github.io/OfficeHour

on-line material: Moodle Site
`https://elearning.uniroma1.it/course/view.php?id=14819`
You can also find there a pdf booklet with all course info: office hour, tutor, contacts, syllabus, exam rules etc.

last update: April 4, 2025 alle ore 09:18

Final exam:

- For students regularly attending classes: 2 Homeworks + 2 Intermediate Tests [70%] + Final project with oral discussion on all homeworks/Test/Project [30%]
- For students not attending classes: Written Test [70%] + Final project + oral discussion on all the topics listed in the course content

Exam preparation materials:

- Lecture notes
- Reference textbooks:
- Office Hours: LT

Main topics

- Intro to the main ingredients and tools of Bayesian inference [review]
- From univariate conjugate to multivariate non-conjugate ...
- .... and the need to approximate!
- Random variables and i.i.d. simulation
- Monte Carlo as a simulation-based approximation device
- Accuracy and Efficiency in error control
- Basic Markov Chain theory
- Markov Chain simulation
- Monte Carlo Markov Chain methods
    - Algorithm for simulating ergodic Markov Chains
    - Basic theory of Markov Chain convergence
    - Efficiency and diagnostics for MCMC methods
- Extensions, variations and alternative approximation approaches
- Applications of Bayesian models and Bayesian inference on models for empirical data

Probability law, simulation and approximation

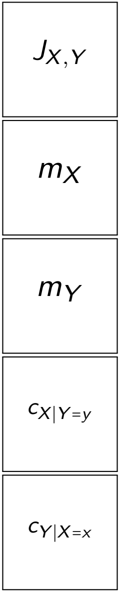
- Why should we care about randomness?
- PL - a mathematical tool to define/assign/represent the behaviour of a random outcome. We will deal with PL of a random variable X, or a random vector (Y, Z), of a stochastic process $(W_1, ..., W_n, ...)$
  can you tell what are specifically the alternative ways of assigning a probability law to the aforementioned random objects?
- LEARN - some way of inferring the unknown characteristic of interest of a phenomenon under investigation from (known) observed data
- SIM - an algorithmic and numerical recipe to pretend to produce observations from those ”random objects”
- APPR - some way of arriving ”close” to some desired target I, for instance

$$I = E[X]$$

Probability law, simulation and approximation

- Why should we care about randomness?
  - as a way of representing real phenomena before they are observed [conceptual framework for inferential statistics]
  - as a way of representing our state of knowledge about all is uncertain [conceptual framework for subjectivism]
  - as a framework for learning [conceptual framework Bayesian inference]
  - as a framework for approximating interesting quantities that can be regarded as expected values or, more generally, as estimable [consistently achievable] syntheses of a probability law

5-card "game"



$J_{X,Y}$

m_X

m_Y

$c_{X|Y=y}$

$c_{Y|X=x}$

[2]

[2] Indeed there are 5 card types, conditional distributions $c_{X|Y=y}(\cdot)$ and $c_{Y|X=x}(\cdot)$ have the same cardinality of the possible conditioning arguments $y \in \mathcal{Y}$ and $x \in \mathcal{X}$

Bayesian Inference

Use a **fully probabilistic** representation of the unknown phenomenon of interest: from the classical statistical model to the Bayesian statistical model

$$(\mathcal{Y} \times \Theta, J(y,\theta) = f(y|\theta)\pi(\theta))$$

1. $\mathcal{Y}$ Observation space
2. Θ Parameter space
3. $f(y|\theta)$ probabilistic representation of the observable data $Y \in \mathcal{Y}$ given (conditionally on) an unknown/uncertain parameter value θ
4. $\pi(\theta)$ prior distribution on the unknown(=uncertain) parameter value

Now we have a single probability space which involves jointly the observable data $Y = y$ and the parameter θ.

Main idea of Bayesian learning

Use probability to represent our state of uncertainty about unknown quantities (Y and θ)

Use the rule of probability to update our belief once some evidence is available (i.e. data $Y = y$ has been observed).

It is conceptually as easy as updating our state of knowledge when rolling a die! We have a (single) probability space which allows us to compute the probabilities of all events (e.g. $E = \{\text{the outcome is even}\}$ $U = \{\text{the outcome is 2}\}$)

When some partial information becomes available ... we change the state of knowledge and hence our beliefs are updated. A rational/coherent way of updating is the use of probability rules

$$P(U|E) = \frac{P(U \cap E)}{P(E)}$$

Bayes rule

Our Bayesian statistical model represents the state of uncertainty about all the unknown phenomena (Y and θ) and acts as the reference probability space.

1. prior to the observation of some data our beliefs on the unknown parameter of interest θ are represented in terms of the prior distribution $\pi(\theta)$
2. once we get some information from the knowledge of the outcome $Y = y$ our beliefs of the unknown parameter θ must be updated coherently though the rule of probability. In the absolutely continuous case under usual regularity conditions

$$\pi(\theta|y) = J(\theta|y) = \frac{J(y,\theta)}{J(y)} = \frac{f(y|\theta)\pi(\theta)}{m(y)}$$

Bayes main ingredients and recipe

1. we choose a statistical model that should reflect our beliefs about the conditional distribution of the observable data Y, $p(y|\theta)$
2. prior to the observation of any data our beliefs on the unknown parameter of interest θ are represented in terms of the prior distribution $\pi(\theta)$
3. once we get some information from the knowledge (certainty) of the outcome $Y = y$ our beliefs of the unknown parameter θ must be updated coherently though the rule of probability. In the absolutely continuous case under usual regularity conditions

$$\pi(\theta|y) = J(\theta|y) = \frac{J(y,\theta)}{J(y)} = \frac{f(y|\theta)\pi(\theta)}{m(y)}$$

Posterior distribution

The posterior distribution $\pi(\theta|y)$ contains all the updated information on the unknown quantity of interest which is still unknown (i.e. uncertain) but with a revised state of uncertainty which embodies both my prior information and the information derived from observing the data $Y = y$

Posterior combines Prior "&" Likelihood

"&" stands for Bayes rule (i.e. conditioning)
From the posterior distribution we can make inference using:

- summaries of $\pi(\theta|y)$ as point estimates
- quantile-based intervals for credibility intervals or HPD (Highest Posterior Density) intervals
- posterior probability statements for comparing two alternative hypothesis

Bayes rule

The denominator

$$m(y) = \int_{\Theta} J(y,\theta)d\theta = \int_{\Theta} f(y|\theta)\pi(\theta)d\theta$$

can be regarded from different perspectives:

1. normalizing constant
2. marginal likelihood
3. Bayesian evidence
4. (prior) predictive distribution

Denominator of the Bayes rule

1. $m(y)$ is the so-called normalizing constant of the posterior density and its reciprocal makes the posterior proportional to the product of the prior and the likelihood

$$\pi(\theta|y) \propto f(y|\theta)\pi(\theta)$$

2. it is also called marginal likelihood as it can be regarded as the likelihood $L_y(\theta) = f(y|\theta)$ integrated (hence marginalized) with respect to the prior distribution.
3. It will be a very important quantity of interest when more that one model is at stake. It is also called Bayesian evidence of a statistical model. In model comparison between two alternative models it allows us to compute the so called Bayes Factor.
4. Regarded as a function of y it represents a distribution. Technically it is a marginal distribution. It is also named prior predictive distribution.

Why Bayes?

... to B(ay)e(s) or not to B(ay)e(s) ... is a very debated issue. We will limit ourselves to understand it is an alternative framework for making inference of a phenomenon of interest and, hopefully we will learn pros and cons.

- Coherent framework which has always the same recipe and do not require ad hoc adjustments

Bayes pros and cons

Short list of pros

+ Coherent framework which has always the same recipe and do not require ad hoc adjustments
+ Allows to embody prior information
+ Simplicity and transparency: all the ingredients are declared in the probabilistic framework
+ It always works with small as well as for large sample sizes
+ Often it can have good frequentist properties for small sample sizes
+ For large sample sizes it approximates optimal frequentist solutions
+ It obeys the likelihood principle

Bayes pros and cons

Short list of cons

- prior distribution is a subjective ingredient hence it can be deemed not scientifically objective as it can bias the information provided by the data contained in the data
- even if there is genuine available prior information the prior distribution is difficult to elicit
- computational difficulties in deriving and summarizing the posterior distribution
- it requires somehow more skills to be understood and implemented

The role of multiplicative constants

- a multiplicative constant can turn a finite measure density into a probability measure density and viceversa
- the easiest way to handle the posterior distribution is recognizing its functional form up to a multiplicative constant

$$\pi(\theta|y) \propto L_y(\theta)\pi(\theta) \propto c \cdot L_y(\theta) \cdot \cdot \pi(\theta)$$

- a multiplicative constant multiplying the likelihood function does not affect the information on the unknown parameter provided by the data (likelihood principle)
- multiplicative constants play a crucial role in evaluating probability masses or densities. In particular in computing marginals such as $m(y)$ used for predictions as well as for Bayes factor and Bayesian evidence

Bayes-easy

Conjugate analysis: the most convenient way of approaching the derivation of the posterior distribution is to use a prior distribution which has, roughly speaking, the same functional form of the likelihood function. This leads to a so-called conjugate analysis. More precisely

Conjugate class of priors for a statistical model A parametric class $\mathcal{P} = \{\pi_h(\theta); h \in \mathcal{H}\}$ of prior distributions for the parameter $\theta \in \Theta$ is called conjugate for a sampling model $f(y|\theta)$ if

$$\pi_h(\cdot) \in \mathcal{P} \implies \exists h^* \in \mathcal{H} \text{ s.t. } \pi(\cdot|y) = \pi_{h^*(y)}(\cdot|y) \in \mathcal{P}$$

Bayes-easy: conjugate Bayesian analysis and updating formula

Updating formula is the transformation which for any prior hyperparameter $h_{prior} = h$ provides the corresponding posterior hyperparameter $h_{posterior} = h^*$ where the updated hyperparameter

$$h^* = u(h, y) = u_T(h, T(y))$$

can be expressed as a function of the prior hyperparameter and the observed data or, equivalently, as a function of the prior hyperparameter and a sufficient statistics $T(y)$.

Inferential Framework

1. INPUT: $J(y,\theta) = \pi(\theta) f(y|\theta)$
   - assumption on the choice of a parametric statistical model
   - assumption on the elicitation of a prior distribution on the unknown parameter (subjective or formal default choices)
   - assumptions sometimes reflect convenient simplifications and approximations
   - required technical skills in translating our (possibly partial) knowledge of an observable phenomenon into the INPUT ingredients
2. INFERENTIAL ENGINE: conditioning on observed (known) data $Y = y_{obs}$
3. OUTPUT: $\pi(\theta|y_{obs})$ posterior distribution on θ.
   - uncertainty as a whole distribution
   - point estimate (mode, median, expectation [when it exists], loss minimizer)
   - interval estimate (equal tail, HPD)
   - hypothesis testing (easy for composite hypotheses, some th. difficulties with point null hypothesis)

Conjugate analysis

Some examples for one parameter models:

- Binomial model + Beta prior
- Bernoulli model + Beta prior
- Poisson model + Gamma prior
- Exponential model + Gamma prior
- Normal model (known variance) + Normal prior

Conjugate analysis for a Binomial model

Binomial statistical model - Suppose that N is a known quantity and $Y|\theta$ is random and such that

$$Y|\theta \sim \mathrm{B}in(N, \theta)$$

i.e.

$$f(y|\theta) = L_y(\theta) = \binom{N}{y}\theta^y(1-\theta)^{N-y} \propto \theta^y(1-\theta)^{N-y}$$

Fotr this statistical/conditional/parametric model the conjugate family is $\mathcal{P} = \{\pi_h(\theta); h \in \mathcal{H}\}$ is the Beta family where $h = (a, b) \in (0, \infty) \times (0, \infty)$ and

$$\pi_{(a,b)}(\theta) = \frac{1}{B(a, b)}\theta^{a-1}(1-\theta)^{b-1}I_{(0,1)}(\theta)$$

We will consider here a single observed outcome $Y = y$

Beta distributions

For the Beta distribution

$$\pi_{(a,b)}(\theta) = \frac{1}{B(a,b)}\theta^{a-1}(1-\theta)^{b-1}I_{(0,1)}(\theta)$$

we should know few fundamental facts:

- $$\int_0^1 \theta^{a-1}(1-\theta)^{b-1}I_{(0,1)}(\theta)d\theta = B(a,b)$$
- $$B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$
- $E[\theta] = \frac{a}{a+b}$; $V[\theta] = \frac{ab}{(a+b)^2(a+b+1)}$
- for $a > 1$ and $b > 1$ the Beta density is unimodal and the mode is $\frac{a-1}{a+b-2}$

The family of Beta distributions and its reparametrizations

As usual in parametric distributions, we can reparameterize the Beta distribution [most often parameterized in terms of
$h = (a, b) = (\texttt{shape1}, \texttt{shape2})$]
in terms of a new parameter r using a transformation $r = g(h)$
where $g(\cdot)$ is one-to-one.

$$r = (\mu, \psi) = g(a, b) = \left(\mu = \frac{a}{a+b}, \psi = a + b \right)$$

where μ is easily recognized as the *mean* of the distribution while ψ can be thought of as a *concentration* parameter which can be also regarded as the strength of our prior belief around the mean.

Reparameterization

A parametric class of distributions

$$\mathcal{P} = \{\pi_h(\theta); h \in \mathcal{H}\}$$

where each distribution is indexed/parameterized in terms of $h \in \mathcal{H}$ can be reparameterized in many (arbitrary) ways. We only need a one-to-one function, say $b(\cdot)$ mapping from the original parameter space $\mathcal{H}$ onto a new parameter space $\mathcal{R}$ where $\mathcal{R} = \{r : r = b(h); h \in \mathcal{H}\}$. In such circumstance we have

$$\mathcal{P} = \{\pi_h(\theta); h \in \mathcal{H}\} = \{\tilde{\pi}_r(\theta); r \in \mathcal{R}\}$$

In what follows we might avoid stressing that $\pi_h(\theta)$ regarded as a function of the parameter index h has a functional form different from that of $\tilde{\pi}_r(\theta)$ as function of r and we will use the generic $\pi_r(\theta)$ notation.

Conjugate analysis for a Binomial model

For conjugacy one easily derive that if we elicit $\pi(\theta) \sim Beta(a,b)$ then the posterior distribution

$$\pi(\theta|y) \propto L_y(\theta)\pi(\theta) \propto \theta^y(1-\theta)^{N-y}\theta^{a-1}(1-\theta)^{b-1} = \theta^{a+y-1}(1-\theta)^{b+N-y-1}$$

hence from $h_{prior} = (a,b)$ to $h_{post} = (a+y, b+N-y)$, the posterior distribution is then a Beta distribution with parameter $h = h_{post} = (a+y, b+N)$, the prior-to-posterior hyperparameter update is understood as $h_{prior} = (a,b) \rightarrow h_{post} = (a+y, b+N-y)$, and derivation of the main posterior summaries is easily done for instance

$$E[\theta|y] = \frac{a+y}{a+b+N}$$

Conjugate analysis for a Binomial model

However, the reparameterized family makes the the prior-to-posterior hyperparameter update better understood as follows

$$
\begin{aligned}
r_{prior} &= (\mu_{prior}, \psi_{prior}) \rightarrow \\
r_{post} &= (\mu_{post} = w\mu + (1-w)y, \psi_{post} = \psi + N)
\end{aligned}
$$

where

$$w = \frac{\psi_{prior}}{\psi_{prior} + N} \in (0,1)$$

- Can you rephrase in words what happens to our prior beliefs after observing some data y?
- How this update is affected by the known N?

Now it should be easy to generalize this conjugate analysis from a single observed outcome $Y = y$ to many outcomes $Y_1, \ldots., Y_n$ possibly considering the special case [for $N = 1$] of multiple, say n, Bernoulli outcomes.

Conjugate analysis for the success probability of a Binomial trial

Let us go through Example in section 1.2.2 (PH) - Bayesian analysis on the prevalence of a rare disease in a small city

- formalize and understand the Bayesian model
- fine tune a suitable prior distribution (prior info in terms of expected prevalence 0.1, 0.7 probable interval [0.05,0.2])
- derive the posterior distribution
- visualize the whole posterior distribution
- provide basic posterior summaries
- compute HPD
- test whether or not the prevalence is larger than 10%
- predict results of a future sample of clinical test for the disease (and use it to plan your future experiment)

Inferential Framework

1. INPUT: $J(y,\theta) = \pi(\theta) f(y|\theta)$
   - assumption on the choice of a parametric statistical model
   - assumption on the elicitation of a prior distribution on the unknown parameter (subjective or formal default choices)
   - assumptions sometimes reflect convenient simplifications and approximations
   - required technical skills in translating our (possibly partial) knowledge of an observable phenomenon into the INPUT ingredients
2. INFERENTIAL ENGINE: conditioning on observed (known) data $Y = y_{obs}$
3. OUTPUT: $\pi(\theta|y_{obs})$ posterior distribution on θ.
   - uncertainty as a whole distribution
   - point estimate (mode, median, expectation [when it exists], loss minimizer)
   - interval estimate (equal tail, HPD)
   - hypothesis testing (easy for composite hypotheses, some theoretical difficulties with point null hypothesis)

Point estimation

In the Bayesian framework, the point estimate $\hat{\theta}$ for θ is a representative value of the posterior distribution.

Three main summaries of $\pi(\theta|y)$ are typically considered:

- **posterior mean** (if it exists!)

$$E(\theta|y) = \begin{cases} \int_\Theta \theta\pi(\theta|y)\,d\theta & \text{continuous case} \\ \sum_{\theta\in\Theta} \theta\pi(\theta|y)\,d\theta & \text{discrete case} \end{cases}$$

  it is frequently-used, but not necessarily always the best choice.

- **posterior median**

$$Median(\theta|y) = \theta_{.5}(y) = \inf\left\{\tilde{\theta} \in \Theta : \pi(\theta \le \tilde{\theta}|y) \ge 0.5\right\}$$

- **posterior mode** or *maximum a posteriori* (MAP)

$$Mo(\theta|y) = \underset{\theta\in\Theta}{arg\ max}\,\pi(\theta|y)$$

  that it is not necessary unique.

How to choose among the posterior summaries?

The appropriateness of a certain point estimate depends on the shape of the posterior distribution.

- Simmetry

$$E(\theta|y) = Median(\theta|y)$$

- Simmetry and unimodality

$$E(\theta|y) = Median(\theta|y) = Mo(\theta|y)$$

- Asimmetry

$$\begin{aligned} Mo(\theta|y) < Median(\theta|y) < E(\theta|y) &\qquad \text{positive skewness} \\ E(\theta|y) < Median(\theta|y) < Mo(\theta|y) &\qquad \text{negative skewness} \end{aligned}$$

the posterior mean may suffer from lack of robustness.

How to compute the posterior summaries?

- Exact method: when analytical formulas are available or we are able to derive them from the posterior (very limited number of cases!)
- Approximate methods: by using a simulation-based approximation of the posterior distribution and computing the summaries on the resulting sample.

Credible intervals

Limit of the posterior summaries: no quantification of estimation uncertainty.

Bayesian solution for set estimation → **credible sets/intervals**

- provide sets of plausible values for the unknown parameter θ on the basis of the information described by $\pi(\theta|y)$
- represent the Bayesian counterpart of frequentist confidence intervals
- enjoy a different and more straightforward interpretation than confidence intervals

Bayesian interval estimates

$1 - \alpha$ posterior probability credible interval estimate $[\theta_{L,(1-\alpha)}, \theta_{U,(1-\alpha)}]$ is an interval (in the univariate case) such that:

$$\int_{\theta_L}^{\theta_U} \pi(\theta | y_1^{obs}, \ldots, y_n^{obs}) d\theta = 1 - \alpha$$

There is no unique way of choosing θ_L and θ_U

- equal-tail interval with high posterior probability content (say 95% of all posterior probability)
- HPD (Highest Posterior Density) interval with high posterior probability content (say 95% of all posterior probability)

Credible intervals

A $100(1-\alpha)\%$ **credible set** for θ is a (observed) region $S(y) \subset \Theta$ such that

$$\pi(\theta \in S(y)|y) \geq 1-\alpha \qquad \alpha \in (0,1),$$

that is, conditionally on the observed sample, a set which includes the true value of the parameter with a predetermined posterior probability $1-\alpha$.

The value $(1-\alpha)$ is referred to as *Bayesian coverage* or *credibility level* and corresponds to the posterior probability mass over the region S(y).

$(1-\alpha) = 0.95$ is a typical value of high posterior probability content.

A comparison with confidence intervals

A $100(1-\alpha)\%$ **confidence interval** for θ is a (random) interval $[\ell_F(Y), u_F(Y)]$ such that

$$Pr(\ell_F(Y) \leq \theta^* \leq u_F(Y) | \theta^*) \geq 1-\alpha \qquad \alpha \in (0,1).$$

The value $(1-\alpha)$ is referred to as *frequentist coverage* or *confidence level* with respect to the data generating distribution conditionally on a fixed θ^*.

Note: the roles of the sample and parameter θ in the coverage probability are completely reversed in the two inferential approaches.

A comparison with confidence intervals

In confidence intervals, the randomness concerns the sample and follows the repeated sampling principle of hypothetical identical repetitions of the experiment under the true data generating mechanism (constant θ).

$$\downarrow$$

pre-experimental interval

In credible intervals, the randomness concerns the parameter θ and it is evaluated conditionally on the unique observed sample resulting from the experiment.

$$\downarrow$$

post-experimental interval

Equal-tails intervals

Note that there is no unique way of choosing $S(y)$ to guarantee a posterior coverage equal to $1-\alpha$. We will focus on

- equal-tails (ET)
- highest posterior density credible interval (HPD)

A $100(1-\alpha)\%$ **equal-tails credible interval** for θ is given by

$$ET_{1-\alpha}(y) = [\theta_{\alpha/2}(y), \theta_{1-\alpha/2}(y)]$$

where the two bounds correspond, respectively, to the posterior quantiles at level $\alpha/2$ and $1-\theta_{\alpha/2}$, i.e., in the regular continuous case

$$\pi(\theta \leq \theta_{\alpha/2}(y)|y) = \alpha/2$$
$$\pi(\theta \leq \theta_{1-\alpha/2}(y)|y) = 1-\alpha/2$$

Equal-tails intervals

In the discrete case

$$\pi(\theta \leq \theta_{\alpha/2}(y)|y) \geq \alpha/2$$
$$\pi(\theta \leq \theta_{1-\alpha/2}(y)|y) \geq 1 - \alpha/2$$

Features of ETs:

- the bounds are determined so that the posterior probabilities over the two tails coincide and is equal to $\alpha/2$
- always include the posterior median
- easy to compute, exactly or approximately
- are not optimal: they may include θ-values with posterior probability/density lower than that of θ-values outside the interval.

HPD intervals

A $100(1-\alpha)\%$ **highest posterior density credible interval** for θ is the smallest set $HPD_{1-\alpha}(y)$ such that

$$\begin{cases} \pi(\theta \in HPD_{1-\alpha}(y)|y) \geq 1-\alpha \\ \pi(\theta|y) > \pi(\theta'|y) \quad \forall \theta \in HPD_{1-\alpha}(y) \quad \forall \theta' \notin HPD_{1-\alpha}(y) \end{cases}$$

HPD intervals

Features of HPDs:

- are optimal, because they represent the shortest credible intervals for a given credibility level $1-\alpha$
- always include the posterior mode
- rarely available in closed-form. Several R contributed packages (`TeachingDemos`, `BayesTwin`, `coda`) allow to compute them approximately.
- are not always intervals, for example in the case of multimodality.

$\pi(\theta|y)$ is symmetric <u>and</u> unimodal $\Rightarrow$ ET=HPD

Bayes rule

The denominator

$$m(y) = \int_{\Theta} J(y,\theta)d\theta = \int_{\Theta} f(y|\theta)\pi(\theta)d\theta$$

can be regarded from different perspectives:

1. normalizing constant
2. marginal likelihood
3. Bayesian evidence
4. (prior) predictive distribution

Hypothesis testing for composite hypotheses

If you have set up two alternative composite hypothesis:

$$\begin{cases} H_0: & \theta \in \Theta_0 \\ H_1: & \theta \in \Theta_1 \end{cases}$$

then after observing $Y_1 = y_1^{obs}, ..., Y_n = y_n^{obs}$ one can easily exploit all the uptated uncertainty in $\pi(\theta | y_1^{obs}, ..., y_n^{obs})$ to compute

$$\pi(\Theta_0 | y_1^{obs}, ..., y_n^{obs}) = \begin{cases} \int_{\Theta_0} \pi(\theta | y_1^{obs}, ..., y_n^{obs}) d\theta & \text{(continuous case)} \\ \sum_{\theta_i \Theta_0} \pi(\theta_i | y_1^{obs}, ..., y_n^{obs}) d\theta & \text{(discrete case)} \end{cases}$$

and decide in favor of H_0 whenever $\pi(\Theta_0 | y_1^{obs}, ..., y_n^{obs}) > 0.5$ or a larger threshold if deemed appropriate

Hypothesis testing for point-null hypothesis

Not so straightforward to extend the "natural" previous approach to the case where $\Theta_0 = \{\theta_0\}$ i.e. the case of point null hypothesis

$$\begin{cases} H_0: & \theta = \theta_0 \\ H_1: & \theta \neq \theta_0 \end{cases}$$

especially when θ lives in a continuous space.

One way of solving this issue can be to adopt something similar to the frequentist rule:
decide in favor of H_0 if θ_0 is inside the 0.95 posterior probability credible interval estimate $[\theta_{L,0.95}, \theta_{U,0.95}]$. The huge difference w.r.t. the frequentist is the meaning of $1 - \alpha$

Bayes Factor i.e. Bayesian hypothesis testing discounting prior odds

The Bayes factor is the ratio between the posterior odds in favor of H_0 and the prior odds

$$BF_{01} = \frac{\frac{\pi(\Theta_0|y)}{\pi(\Theta_1|y)}}{\frac{\pi(\Theta_0)}{\pi(\Theta_1)}}$$

The Bayes factor is a measure of the evidence provided by the data in support of the hypothesis H_0. It is measured on a specific scale which should be suitably calibrated

Bayes factor calibration

You may find alternative guidelines on the scale interpretation. The following is one of the most cited[3]

BF_{01}	Interpretation
Under 1	Supports H_1
1-3	Weak support for H_0 - not worth more than a bare mention
3-20	Support for H_0
20-150	Strong support for H_0
Over 150	Very strong support in favor of H_0

Note that when the $BF_{01} < 1$ you can use $BF_{10} = 1/BF_{01} > 1$ to interpret the strength of support in favor of H_1

[3] Kass and Raftery (1995), *Bayes Factors*, JASA

Conjugate analysis for a Binomial trial

The prior predictive distribution is a Beta-Binomial distribution with the following features:

$$Y \sim m(y) = \binom{N}{y} \frac{Beta(a+y, b+N-y)}{Beta(a,b)}$$

$$\begin{aligned} E[Y] &= N\frac{a}{a+b} \\ V[Y] &= N\frac{ab(a+b+N)}{(a+b)^2(a+b+1)} \end{aligned}$$

Remark on predictive and conditional distribution

We can remark that if we fix a specific $\hat{\theta}$ even if we have

$$E[Y] = E[Y|\hat{\theta}]$$

we get

$$Var[Y] > Var[Y|\hat{\theta}]$$

This will have an impact in the (erroneously) reduced uncertainty in predicting an observable Y based on the conditional/statistical model for $Y|\hat{\theta} \sim f(y|\hat{\theta})$ rather than the marginal or predictive distribution for $Y \sim m(y)$

Prediction

what is the key understanding in Bayesian inference?

Prediction

Let us suppose that I want to predict something which is not known or, equivalently, which has not yet been observed. Suppose that this something is related to the observation Y.

what is the key understanding in Bayesian inference?

$\Longrightarrow$ Use the probability law of the random quantity Y!!

Prediction

$\Longrightarrow$ Use the probability law of the random quantity Y!!
Well, Y .....

- where does its randomness come from?
- how is that randomness framed?

Predict Y before observing it

It comes from our basic Bayesian modelling:

$$(\mathcal{Y}, \Theta)$$

$$f(y|\theta)\pi(\theta) \quad = \quad J(y,\theta) \quad \rightarrow \quad J(y) = m(y)$$

Hence we should use the so-called (prior or initial) predictive distribution

$$Y \sim m(\cdot) = \int_\Theta f(\cdot|\theta)\pi(\theta)d\theta$$

Prediction after observing the data $Y = y$

Let us suppose again that we want to predict something which is not known or, equivalently, which has not yet been observed. Suppose however that this something is in fact related to the observation Y but it is not the same observation $Y = y$ that has been observed. Perhaps I can enlarge my frame to include a possible future observation, say Y^{new}.

what should I do first?

what is the key understanding in Bayesian inference?

$\Longrightarrow$ Use the probability law of the random quantity Y^{new} ..... well .... yes but ..... which one?

Prediction after observing the data $Y = y$

what should I do first?

Let us understand what is the enlarged frame under the assumption of **conditionally i.i.d.** assumption that is that the vector $(Y, Y^{new})|\theta$ have (conditionally on θ) independent and identically distributed components

$$(\Theta, \mathcal{Y}, \mathcal{Y}^{new})$$

$$\pi(\theta)f(y|\theta)f(y^{new}|\theta) \quad = \quad J(\theta, y, y^{new}) \quad \rightarrow \quad \ldots.$$

Prediction after observing the data $Y = y$

what should we do when we observe some data?

what is the key understanding in Bayesian inference?

Prediction after observing the data $Y = y$

what should we do when we observe some data?

what is the key understanding in Bayesian inference?

$\Longrightarrow$ Use the **conditional** probability law of the random quantity $Y^{new}|Y = y$

Prediction after observing the data $Y = y$

What is left random after $Y = y$ in our new frame?

What random quantity are we really (only) interested in?

How can we derive the corresponding probability law?

Prediction after observing the data $Y = y$

Hence $\Longrightarrow$

- we should use the **conditional** probability law of the random quantity $Y^{new}|Y = y$
-

Prediction after observing the data $Y = y$

What is left random after $Y = y$ in our new frame?

$$(\theta, Y^{new} | Y = y) \sim J(\theta, y^{new} | y) = ?$$

What random quantity are we really (only) interested in?

$\Longrightarrow$ Use the **conditional** probability law of the random quantity $Y^{new} | Y = y$

How can I derive the (conditional) probability law of $Y^{new} | Y = y$ from $(\theta, Y^{new} | Y = y) \sim J(\theta, y^{new} | y)$?

$$(Y^{new} | Y = y) \sim m(y^{new} | y) = J(y^{new} | y) = ?$$

Updating by conditioning

The rule to update some joint distribution after observing some of the originally random components of the whole random vector is always the same

$$J(\theta, y^{new}|y) \propto J(\theta, y, y^{new})$$

more precisely

$$J(\theta, y^{new}|y) \propto J(\theta, y, y^{new}) \quad = \quad \pi(\theta) f(y|\theta) f(y^{new}|\theta)$$

Indeed the proportionality constant depends only on y and not on (y^{new}, θ) since its reciprocal is

$$\begin{aligned} J(y) &= \int_{\Theta \times \mathcal{Y}^{new}} \pi(\theta) f(y|\theta) f(y^{new}|\theta) dy^{new} d\theta \\ &= \int_{\Theta} \int_{\mathcal{Y}^{new}} f(y^{new}|\theta) f(y|\theta) \pi(\theta) dy^{new} d\theta \\ &= \int_{\Theta} \left[\int_{\mathcal{Y}^{new}} f(y^{new}|\theta) dy^{new} \right] f(y|\theta) \pi(\theta) d\theta = m(y) \end{aligned}$$

Updating by conditioning

$$\begin{aligned}
J(y^{new}, \theta|y) &= \frac{1}{m(y)} f(y^{new}|\theta) f(y|\theta)\pi(\theta) \\
&= \frac{f(y^{new}|\theta) f(y|\theta)\pi(\theta)}{m(y)} \\
&= f(y^{new}|\theta) \frac{f(y|\theta)\pi(\theta)}{m(y)} \\
&= f(y^{new}|\theta)\pi(\theta|y)
\end{aligned}$$

Let us provide some comment on this formula!! What does it resemble?

Prediction of a future observation in a Binomial model

If we remember the remarkable formula for the predictive distribution of Y when (Y,θ) is a Bayes conjugate model with a binomial $Y|\theta$ and a Beta distribution on θ we immediately derive that the posterior predictive distribution i.e. the conditional distribution of $Y^{new}|Y=y$ is a Beta-Binomial distribution with the following features:

$$Y^{new}|y \sim m(y^{new}|y) = \binom{N}{y^{new}} \frac{Beta((a+y)+y^{new},(b+N-y)+N-y^{new})}{Beta(a+y,b+N-y)}$$

$$\begin{aligned}
E[Y^{new}|Y=y] &= N\frac{a+y}{a+b+N} \\
V[Y^{new}|Y=y] &= N\frac{(a+y)(b+N-y)((a+y)+(b+N-y)+N)}{((a+y)+(b+N-y))^2((a+y)+(b+N-y)+\text{N+1})} \\
&= N\frac{(a+y)(b+N-y)((a+b+N)+N)}{(a+b+N)^2(a+b+N+\text{N+1})}
\end{aligned}$$

Prediction

Sometimes, once we have updated our prior beliefs in light of the data, there is another interesting distribution (uncertainty) which can be of interest: the predictive distribution for a future observation Y_{new}.
We can generalize our predictive distribution formula**s** (prior, posterior) to account for multiple observations.
Indeed, if I start from generalizing the prior predictive distribution for a single observation, $J(y) = m(y)$ to the case of repeated observations in the sample (under similar conditions i.e. conditional i.i.d. components) I will discover that the marginalized joint distribution $J(y_1, ..., y_n) = m(y_1, ..., y_n)$ has a special dependence structure ... (see later ...)

Joint posterior predictive

… for a single new observation and multiple observed outcomes. From the model, I have my initial belief

…

but once I observe the data $\boldsymbol{Y} = (Y_1, \ldots, Y_n) = \boldsymbol{y} = (y_1, \ldots, y_n)$ I should use

$$m(y_{new}|y_1, \ldots, y_n) = m(y_{new}|\boldsymbol{y}) = \int_\Theta f(y_{new}|\theta)\pi(\theta|\boldsymbol{y})d\theta$$

which is called posterior predictive distribution or conditional predictive distribution.
One can generalize the above formulas to account for multiple future observations, say k, i.e. $y_{new,1}, \ldots, y_{new,k}$

How many predictive distributions in Bayesian inference?

1. (initial/prior) predictive distribution[4] for a single observable:

$$Y \sim p_Y(y) = \int_\Theta f(y|\theta)\pi(\theta)d\theta$$

2. (initial/prior) predictive distribution for a many (exchangeable) observables:

$$(Y_1, ..., Y_n) \sim p(y_1, ..., y_n) = \int_\Theta \prod_{i=1}^{n} f(y_i|\theta)\pi(\theta)d\theta$$

[4] In the first slides we had denoted the prior predictive $p_Y(y)$ with $m(y)$. In that context we wanted to stress the role of $m(y_{obs})$ as a single positive number (names marginal likelihood or Bayesian evidence) in a post-experimental perspective

How many predictive distributions in Bayesian inference?

3. (final/posterior) predictive distribution for a new observation:

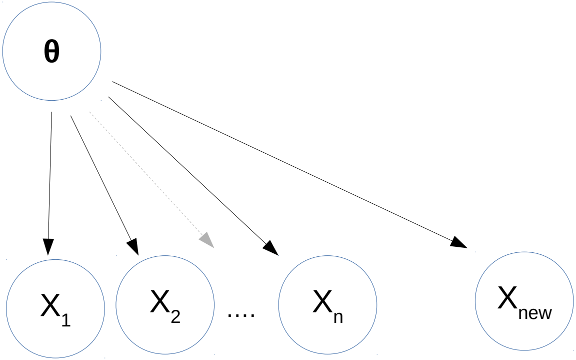
$$Y_{n+1}|Y_1 = y_1, \dots, Y_n = y_n \sim p(y_{n+1}|y_1, \dots, y_n)$$

$$\begin{aligned}
p(y_{n+1}|y_1, \dots, y_n) &= \frac{p(y_1, \dots, y_n, y_{n+1})}{p(y_1, \dots, y_n)} = \\
&= \int_\Theta f(y_{n+1}|\theta) \frac{\prod_{i=1}^n f(y_i|\theta)\pi(\theta)}{p(y_1, \dots, y_n)} d\theta = \\
&= \int_\Theta f(y_{n+1}|\theta) \frac{L_{y_1,\dots,y_n}(\theta)}{p(y_1, \dots, y_n)} \pi(\theta) d\theta = \\
&= \int_\Theta f(y_{n+1}|\theta)\pi(\theta|y_1, \dots, y_n) d\theta
\end{aligned}$$

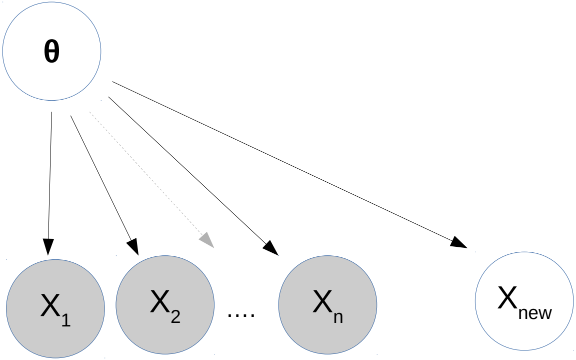
One can easily produce a simulation from $p(y_{n+1}|y_1, \dots, y_n)$ making the following two-step simulation:

- $\theta \sim \pi(\theta|y_1, \dots, y_n)$
- $Y_{n+1}|\theta \sim f(y_{n+1}|\theta)$

DAG before data observation



DAG after data observation



Conjugate analysis for a Poisson model

Poisson model - Suppose $Y = (Y_1, \ldots, Y_n)$ are such that

$$Y_i | \theta \overset{i.i.d.}{\sim} \mathrm{P}oisson(\theta)$$

$\Longrightarrow$ the likelihood function has the following functional form

$$L_{y_1, \ldots, y_n}(\theta) \propto e^{-n\theta} \theta^{\sum_{i=1}^{n} y_i} = e^{-b_{lik}\theta} \theta^{a_{lik}} = g(\theta, h_{lik} = (b_{lik}, a_{lik}))$$

We look for a family of probability densities with support $\Theta = (0, \infty)$ such that

$$\pi(\theta) \propto g(\theta, h_{lik} = (a_{prior}, b_{prior}))$$

and

$$\pi(\theta | y = (y_1, \ldots, y_n)) \propto g(\theta, h_{post} = (a_{post}, b_{post}))$$

Note that (a, b) are not necessarily the usual (most common) parameterization of the parametric family of densities

Conjugate analysis for a Poisson model

In fact, if $r = b$ and $s = a + 1$ we can recognize that

$$e^{-b\theta}\theta^{a} = e^{-r\theta}\theta^{s-1}$$

is the kernel of a Gamma density with *rate* parameter $r > 0$ and *shape* parameter $s > 0$ so that if we take
$\pi(\theta) \sim \text{Gamma}(h_{prior} = (rate = r_{prior}, shape = s_{prior}))$

$$\pi(\theta|y = (y_1, \ldots, y_n)) \propto \pi(\theta) L_{y_1,\ldots,y_n}(\theta) = e^{-r_{prior}\theta - n\theta}\theta^{s_{prior} + \sum_{i=1}^{n} y_i - 1}$$

which corresponds to a $\text{Gamma}(h_{post} = (rate = r_{post}, shape = s_{post}))$ with

$$\begin{aligned} r_{post} &= r_{prior} + n \\ s_{post} &= s_{prior} + \sum_{i=1}^{n} y_i \end{aligned}$$

Conjugate analysis for a Poisson model

Moreover, with a suitable reparameterization, we still have an updating formula for hyperparameters similar to the one already seen for the Binomial model

$$\begin{aligned} \mu_{post} &= w \cdot \mu_{prior} + (1 - w) \cdot \bar{y}_n \\ \nu_{post} &= \nu_{prior} + n \end{aligned}$$

In fact, considering the reparameterization

$$\begin{aligned} \mu_{prior} &= \frac{s_{prior}}{r_{prior}} \\ \nu_{prior} &= r_{prior} \end{aligned}$$

...

Conjugate analysis for a Poisson model

Again,

$$\begin{aligned}
\mu_{post} &= \frac{s_{prior} + \sum_{i=1}^{n} y_i}{r+n} = \frac{\frac{s_{prior}}{r_{prior}} r_{prior} + \frac{\sum_{i=1}^{n} y_i}{n} n}{r_{prior} + n} \\
&= w\mu_{prior} + (1-w)\bar{y}_n \\
&= w\mu_{prior} + (1-w)\hat{\theta}_{MLE}
\end{aligned}$$

Note that $n \to \infty \implies w \to 0$ and the data swamp the prior

Prior predictive distribution for a single observation

In a conjugate Bayesian model for (conditionally)Poisson counts $Y_i|\theta$ if we choose the prior for θ in the Gamma family i.e.

$$\theta \sim \pi(\theta) = \frac{r^s}{\Gamma(s)} e^{-r\theta}\theta^{s-1} \qquad \theta > 0$$

then

$$Y \sim m(y) = J(y) = Pr\{Y = y\} = \frac{\Gamma(s+y)}{\Gamma(s)} \frac{1}{y!} \left(\frac{r}{r+1}\right)^s \left(\frac{1}{r+1}\right)^y$$

which, setting $p = \frac{r}{r+1}$ as the probability success in an i.i.d. Bernoulli trial, can be interpreted as the probability of the number of failures to wait until one observes the s-th success (whenever s is a positive integer)

Prior predictive distribution for a single observation

For a positive integer s we have

$$\begin{aligned}
Pr\{Y = y\} &= \frac{(s+y-1)!}{(s-1)!}\frac{1}{y!}\left(\frac{r}{r+1}\right)^s\left(\frac{1}{r+1}\right)^y \\
&= \binom{s+y-1}{s-1}\left(\frac{r}{r+1}\right)^s\left(\frac{1}{r+1}\right)^y \\
&= \binom{s+y-1}{y}\left(\frac{r}{r+1}\right)^s\left(\frac{1}{r+1}\right)^y
\end{aligned}$$

This distribution is called Negative Binomial and the couple (s, p) can be used as (hyper-)parameter vector.
If $Y \sim NegBin(s, p)$, then

$$E[Y] = \frac{s(1-p)}{p} \qquad V[Y] = \frac{s(1-p)}{p^2}$$

Negative binomial

In a sequence of i.i.d. trials with success probability p one waits y failures before observing s successes if and only if the outcome of $(y+s)$ trials corresponds to exclusively one of any possible arrangements of binary $(y+s)$-tuples with 1 blocked success in the final trial

Prior predictive distribution for a single observation

Prediction of a single observation in a Poisson model.
A priori, before observing the data. We can derive now another example of a prior predictive distribution in a conjugate setting

$$Y \sim \text{NegBin}\left(p = \frac{r_{prior}}{r_{prior}+1}, \texttt{size} = s_{prior}\right)$$

$$\begin{aligned}
E\left[Y\right] &= E[E\left[Y|\theta\right]] = \frac{s_{prior}}{r_{prior}} \\
Var\left[Y\right] &= E[Var\left[Y|\theta\right]] + Var[E\left[Y|\theta\right]] \\
&= \frac{s_{prior}}{r_{prior}} + \frac{s_{prior}}{r^2_{prior}} = \\
&= \frac{s_{prior}}{r_{prior}} \cdot (1 + \frac{1}{r_{prior}}) \\
&= \frac{s_{prior}}{r^2_{prior}}(r_{prior}+1)
\end{aligned}$$

Prediction of a single future observation in a Poisson model

We can derive

$$Y_{new}|Y=(y_1,..,y_n) \sim \text{NegBin}\left(p^{post}=\frac{r_{prior}+n}{r_{prior}+n+1}, \texttt{size}^{post}=s_{prior}+\sum_{i=1}^{n} y_i\right)$$

$$\begin{aligned}
E\left[Y_{new}|Y=(y_1,..,y_n)\right] &= \frac{s_{prior}+\sum_{i=1}^{n} y_i}{r_{prior}+n} \\
Var\left[Y_{new}|Y=(y_1,..,y_n)\right] &= \frac{(s_{prior}+\sum_{i=1}^{n} y_i)(r_{prior}+n+1)}{(r_{prior}+n)^2} = \\
&= E\left[Y_{new}|Y=(y_1,..,y_n)\right]\cdot\frac{(r_{prior}+n+1)}{(r_{prior}+n)}
\end{aligned}$$

Prediction of a future observation in a Poisson model

Let us stress the difference in predicting the future observation using

$$Y^{new} \sim NegBin\left(p = \frac{n}{r_{prior} + n + 1}, m = s_{prior} + \sum_{i=1}^{n} y_i\right)$$

rather than with a *plug-in* formula

$$Y^{new} \sim Poisson(\hat{\theta})$$

as we could be tempted to argue as if we were working with a classical/frequentist approach rather than the Bayesian

Prediction of a future observation in a Poisson model

Let us stress the difference in predicting the future observation using

$$Y^{new} \sim NegBin\left(p = \frac{1}{r_{prior} + n + 1}, m = s_{prior} + \sum_{i=1}^{n} y_i\right)$$

rather than

$$Y^{new} \sim Poisson(\hat{\theta})$$

as we could be tempted to do classical/frequentist approach rather than the Bayesian

Prediction for a Poisson model

$$Y_{new}|Y = (y_1, .., y_n) \sim \text{NegBin}\left(p = \frac{1}{r_{prior} + n + 1}, m = s_{prior} + \sum_{i=1}^{n} y_i\right)$$

$$\begin{aligned} E\left[Y_{new}|Y = (y_1, .., y_n)\right] &= \frac{s_{prior} + \sum_{i=1}^{n} y_i}{r_{prior} + n} \\ Var\left[Y_{new}|Y = (y_1, .., y_n)\right] &= \frac{(s_{prior} + \sum_{i=1}^{n} y_i)(r_{prior} + n + 1)}{(r_{prior} + n)^2} \end{aligned}$$

There is an extra component of variability which makes our prediction more uncertain .....

Conjugate analysis

In a standard Bayesian modelling in the presence of conditionally i.i.d. observables once for the corresponding likelihood function a conjugate parametric class of prior distributions has been identified through a suitable parametrization one can easily make Bayesian inference exploiting the following nice circumstances:

- in order to understand the learning mechanism provided by observation of the data $y_1^{obs}, .., y_n^{obs}$ one can concentrate on the updating of the hyperparameters from h_{prior} to $h_{posterior}$
- if the conjugate class allows for an explicit identification of the prior predictive distribution depending on the prior hyperparameter h_{prior} this implies that the posterior predictive distribution has the same type of distribution with the prior hyperparameter h_{prior} replaced by the $h_{posterior}$

Conjugate analysis

A very general structure for conjugate analysis has been provided for a very broad, flexible and convenient class of statistical models: the exponential family:

- univariate exponential family
- multivariate exponential family

Univariate exponential family

Two alternative equivalent parameterizations:

$$f(x|\theta) \quad = \quad h(x) exp\left\{\eta(\theta) T(x) - B(\theta)\right\}$$

When we set $\psi = \eta(\theta)$, we call ψ the natural parameter and we can also highlight the role of $e^{-B(\theta)}$ as a normalizing factor (constant w.r.t. x) so that it can be rewritten in the natural parametrization as follows

$$f(x|\psi) \quad = \quad c(\psi) h(x) exp\left\{\psi T(x)\right\}$$

The family

$$\mathcal{F} = \left\{f(x|\psi); \psi \in \Psi\right\}$$

is called the canonical univariate exponential family where

$$\Psi = \left\{\psi : c(\psi) = \left(\int_{\mathcal{X}} h(x) exp\left\{\psi T(x)\right\} dx\right)^{-1} < \infty\right\}$$

Conjugate priors for univariate exponential families

$$\pi_{conj}(\psi|n_0, t_0) \propto c(\psi)^{n_0} exp\left\{n_0 t_0 \psi\right\} = exp\left\{n_0 t_0 \psi - n_0(-log(c(\psi)))\right\}$$

where $n_0 > 0$ and $t_0 \in \mathcal{T} = T(\mathcal{X})$ are the two prior parameters.

Let us verify the conjugacy property .... and derive ....

Conjugate priors for univariate exponential families

$$\begin{aligned}
\pi(\psi|x_1, ..., x_n) &\propto \\
&\propto c(\psi)^{n_0+n} exp\left\{(n_0+n)\left[\frac{n_o}{n_0+n}t_0 + \frac{n}{n_0+n}\bar{T}_n\right] + \psi\right\} \\
&\propto \pi_{conj}\left(\psi|n_0+n, \frac{n_o}{n_0+n}t_0 + \frac{n}{n_0+n}\bar{x}_n\right)
\end{aligned}$$

where $\bar{T}_n = \frac{1}{n}\sum_{i=1}^{n} T(x_i)$

Conjugate priors for univariate exponential families

$$\pi(\psi|n_0, t_0) \propto c(\psi)^{n_0} exp\{nt_0\psi\}$$

Understanding and interpreting the role of the two prior parameters

$t_0 = E[T(X)]$ prior guess for the center of the observable $T(X)$

n_0 prior strenght equivalent to the number of imaginary prior observ

Moreover one can prove that

$$E[T(X)|\psi] = -\frac{c'(\psi)}{c(\psi)}$$

1. The importance of being joint (and dependent!)
2. Understanding/visualizing/elaborating (in)-dependence through graphs
3. DAG
4. Exchangeability (finite and infinite)
5. De Finetti representation theorem

Graph representation of a joint distribution

A graph is an interesting formalization which is often useful for a visual understanding of the dependence structure of a joint distribution

- $G = (V, \mathcal{E})$
- $V = \{v_1, \dots, v_k\}$ set of vertices (or nodes)
- $\mathcal{E} = \{e_1, \dots, e_L\}$ subset of edges (or arcs) connecting couples of nodes $e_\ell = (v_i, v_j)$ or $e_\ell = \{v_i, v_j\}$

We can distinguish two types of graphs:

- undirected graphs
- directed (or oriented) graphs

Graph terminology

Parent: The parents of a node is the set of all nodes that feed into it.

Child: The children of a node is the set of all nodes that feed out of it.

Root: A root is a node with no parents.

Leaf: A leaf is a node with no children.

Ancestors: The ancestors are the parents, grand–parents, etc of a node.

Descendants: The descendants are the children, grand-children, etc of a node.

Neighbors: The neighbors of a node is the set of all immediately connected nodes.

Degree: The degree of a node is the number of neighbors. For directed graphs, we speak of the in–degree and out–degree, which count the number of parents and children.

Cycle or Loop: A cycle is a series of nodes such that we can get back to where we started by following edges. We may speak of a directed or undirected cycles.

Path: A path $s \rightsquigarrow t$ is a series of directed edges leading from s to t.

Tree: An undirected tree is an undirected graph with no cycles and only one parent per child. A directed tree is a DAG in which there are no directed cycles.

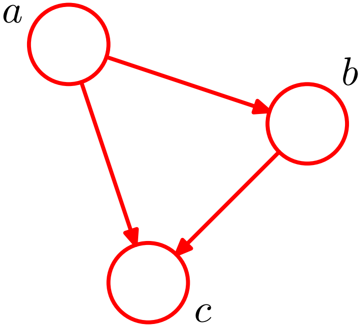
Forest: A forest is a set of trees.

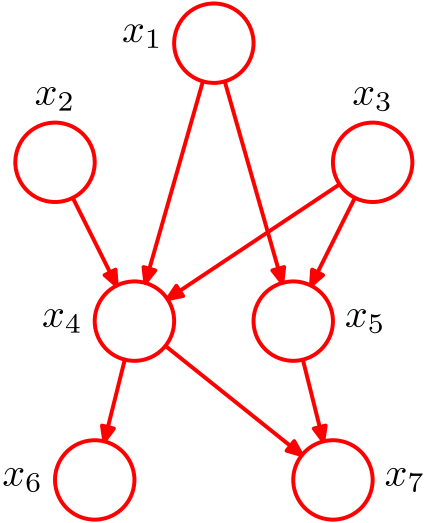
DAG - Directed Acyclic Graph

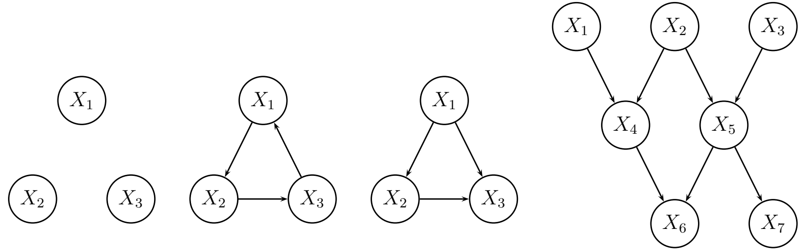
DAG → Directed Acyclic Graph
It is an interesting formalization which is useful not only for a visual understanding of the structure of a statistical model as a joint distribution of random components each represented by a node.

- $G = V, e$
- $V = \{v_1, \ldots, v_k\}$ set of vertices or nodes
- $e = \{e_1, \ldots, e_L\}$ subset of oriented edges connecting [with direction] couple of nodes $v_\ell = (v_i, v_j)$ $v_i \rightarrow v_j$

In a DAG there cannot be cycles.


a
b
c


x_1
x_2
x_3
x_4
x_5
x_6
x_7



(a) A graphical model with no

(b) A directed cyclic graphical model

(c) A directed acyclic graphical model

(d) A more complex DAG

DAG minimal terminology

- child
- parent
- descendants
- ancestors

DAG as a graph representation of a Bayesian model

General formula to represent the **joint** distribution represented by a DAG

$$p(v_1, \ldots, v_k) = \prod_{i=1}^{k} p(v_k | pa(v_k)) \qquad (1)$$

where $pa(v_k)$ denotes the nodes which are *p*arents of the node v_k

NB For this factorization formula to hold nodes must be ordered in a suitable hierarchical way so that there are no links that go from any node to any lower numbered node.

DAG as a graph representation of a joint distribution

Some important preliminary questions (and answers):

- Is it always possible to visualize/represent a joint distribution in terms of a DAG?
- Is the representation of a joint distribution in terms of a DAG unique?
- DAG and more general graphs are useful but not "perfect" tools for representing (in)dependence

DAG as a graph representation of a Bayesian model

Let us try to provide the graphical representation of a Bayesian model of a Bernoulli trial with $n = 5$ observations.

DAG as a graph representation of a Bayesian model

DAGs help us to understand the features of conditional in-dependence structure. The following hods:

Proposition Given its parents, a node is conditionally independent of all of its non-descendants

The above proposition suggests to reinterpret the fundamental factorization in (1) in terms of an operational simulation algorithm.

Graph representation of a probability distribution

Other undirected graphs will be useful to visualize understand the conditional independence structure of a joint (multivariate) distribution.

Exchangeability

A sequence of random variables (a stochastic process) $X_1, \ldots, X_n, \ldots$ is exchangeable if for any k-tuple $(n_1, \ldots, n_k)$ and any permutation $\sigma = (\sigma_1, \ldots, \sigma_k)$ of the first k integers the following holds

$$(X_{n_1}, \ldots, X_{n_k}) \overset{d}{=} (X_{n_{\sigma_1}}, \ldots, X_{n_{\sigma_k}})$$

De Finetti's theorem

For binary random variables $X_i \in \{0,1\}$ the following representation holds

De Finetti's theorem If $X_1,...,X_n,...$ is an exchangeable process of binary random variables there exists a distribution π on $[0,1]$ such that

$$Pr(X_1 = x_1,...,X_n = x_n) = \int_{[0,1]} \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i}\pi(\theta)d\theta$$

The random variables are conditionally independent and i.d. Bernoulli, provided a random $\theta \sim \pi(\theta)$ has been given as their common success probability. Indeed, $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \to \theta \sim \pi(\theta)$ or, equivalently, $\bar{X}_n \approx \pi(\theta)$.

DF theorem holds in a more general case of arbitrary X_i not necessarily binary.

De Finetti's theorem

For binary random variables $X_i \in \{0,1\}$ the following representation holds

De Finetti's theorem If $X_1, \dots, X_n, \dots$ is an exchangeable process of binary random variables there exists a distribution π on $[0,1]$ such that

$$Pr(X_1 = x_1, \dots, X_n = x_n) = \int_{[0,1]} \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i}\pi(d\theta)$$

The random variables are conditionally independent and i.d. Bernoulli, provided a random $\theta \sim \pi(\theta)$ has been fixed as common success probability. Indeed, $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \to \theta \sim \pi(\theta)$ or, equivalently, $\bar{X}_n \approx \pi(\theta)$.

DF theorem holds in a more general case of arbitrary X_i not necessarily binary.

De Finetti's theorem

For random variables $X_i \in \mathcal{X}$ the following representation holds
De Finetti's theorem If $X_1, ..., X_n, ...$ is an exchangeable process of random variables $X_i \in \mathcal{X}$. Then there exists a distribution $\pi(\cdot)$ on the space $\mathcal{P}_{\mathcal{X}}$ of all probability distributions on $(\mathcal{X}, \sigma(\mathcal{X}))$ such that

$$Pr(X_1 \in A_1, ..., X_n \in A_n) = \int_{\mathcal{P}_{\mathcal{X}}} \prod_{i=1}^{n} P(X_i \in A_i) \pi(dP)$$

P is a random probability distribution taking "values" in a functional space $\mathcal{P}_{\mathcal{X}}$

Note that De Finetti's theorem characterizes the exchangeability and proves the existence of a distribution $\pi(\cdot)$ on a suitable space according to which the random observations are conditionally (to the random P) independent and identically distributed. The observations are linked together by sharing the same (random) law P).

Normal statistical model

For observations $Y_1, ..., Y_n$ such that $Y_i|.... \sim N(\theta, \sigma^2)$ we will see how we can:

- easily do the conjugate Bayesian analysis when σ^2 is known [strong unrealistic assumption] using a convenient conjugate class of priors (Normal) for $\theta = \mu$
- less easily do the conjugate Bayesian analysis when both (μ, σ^2) are unknown using a convenient conjugate class of priors for the vector $\omega = (\theta, \sigma^2)$ exploiting a specific form of dependence $\pi(\omega) = \pi(\theta|\sigma^2)\pi(\sigma^2)$ which allows for the prior-to-posterior updating mechanism
- overcome the specific conjugate structure allowing for a simpler independent component prior $\pi(\omega) = \pi(\theta)\pi(\sigma^2)$ so that the simulation-based MCMC approach known as Gibbs Sampling is partially conjugate

2 useful facts

Before providing details on conjugate Bayesian analysis for a Normal sampling model with known variance we state two useful facts:

1. *Reparametrization*. When we are indexing (i.e. parameterizing) a family of probability distributions our way of specifying each member of the family with an index h (i.e. a parameter) is somewhat arbitrary. Instead of using the generic parameter h I could use a one-to-one function of it say $r = g(h)$. As an example instead of using $h = (\tilde{\mu}, \tilde{\sigma}^2)$ to specify a single Normal distribution I could equivalently use the reparametrization $r_{prec} = g_{prec}(h) = (\tilde{\mu}, \frac{1}{\tilde{\sigma}^2})$ where $\frac{1}{\tilde{\sigma}^2}$ is called the *precision*. Similarly, I could equivalently use the reparametrization $r_{sd} = g_{sd}(h) = (\tilde{\mu}, \tilde{\sigma})$ with $\tilde{\sigma}$ being the standard deviation. In other words the parametrization of a family of distributions is a matter of mathematical and/or interpretation convenience.
2. If $\theta \sim p(\theta)$ and
$$p(\theta) \propto exp\left\{-\frac{1}{2}a\theta^2 + b\theta\right\}$$
then $\theta \sim N\left(\tilde{\mu} = \frac{b}{a}, \tilde{\sigma}^2 = \frac{1}{a}\right)$

Useful fact - 1

For the formal rigorous derivation of the previous formulas it is better to state one basic fact which We will consider as known from now on. If we have a density $p(\theta)$ such that

$$p(\theta) \propto exp\left\{-\frac{1}{2}a\theta^2 + b\theta\right\}$$

or, equivalently

$$p(\theta) \propto exp\left\{-\frac{1}{2}\left(a\theta^2 - 2b\theta\right)\right\}$$

then the corresponding distribution is

$$N\left(\tilde{\mu} = \frac{b}{a}, \tilde{\sigma}^2 = \frac{1}{a}\right)$$

Useful fact - 1

$$\exp\left\{-\frac{1}{2\tilde{\sigma}^2}\left(\theta-\tilde{\mu}\right)^2\right\} = \exp\left\{-\frac{\theta^2}{2\tilde{\sigma}^2} + 2\frac{\tilde{\mu}}{2\tilde{\sigma}^2}\theta - \frac{1}{2\tilde{\sigma}^2}\tilde{\mu}^2\right\}$$

$$\begin{aligned} \tilde{\sigma}^2 &= \frac{1}{a} \\ \tilde{\mu} = &= \frac{b}{a} \end{aligned}$$

or, equivalently

$$\begin{aligned} a &= \frac{1}{\tilde{\sigma}^2} \\ b &= \frac{\tilde{\mu}}{\tilde{\sigma}^2} \end{aligned}$$

The complete quadratic expression would be

$$\exp\left\{-\frac{1}{2\tilde{\sigma}^2}\left(\theta-\tilde{\mu}\right)^2\right\} = \exp\left\{-\frac{1}{2}\left(a\theta^2 - 2b\theta + \frac{b^2}{a}\right)\right\} = \exp\left\{-\frac{a}{2}\theta^2 + b\theta - \frac{b^2}{2a}\right\}$$

Useful fact - 1

The complete quadratic expression is then proportional to

$$\begin{aligned}\exp\left\{-\frac{1}{2\tilde{\sigma}^2}\left(\theta-\tilde{\mu}\right)^2\right\} &= \exp\left\{-\frac{a}{2}\theta^2+b\theta\right\}\cdot exp\left\{-\frac{b^2}{2a}\right\} \\ &\propto \exp\left\{-\frac{a}{2}\theta^2+b\theta\right\}\end{aligned}$$

Useful fact - 1

If we consider the entire probability density

$$\frac{1}{\sqrt{2\pi\tilde{\sigma}^2}} \exp\left\{-\frac{1}{2\tilde{\sigma}^2}\left(\theta-\tilde{\mu}\right)^2\right\}$$

and the fact that

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\tilde{\sigma}^2}} \exp\left\{-\frac{1}{2\tilde{\sigma}^2}\left(\theta-\tilde{\mu}\right)^2\right\} d\theta = 1$$

we can also derive the following

$$\int_{-\infty}^{\infty} \frac{\sqrt{a}}{\sqrt{2\pi}} \exp\left\{-\frac{a}{2}\theta^2 + b\theta - \frac{b^2}{2a}\right\} d\theta = 1$$

hence

$$\int_{-\infty}^{\infty} \exp\left\{-\frac{a}{2}\theta^2 + b\theta\right\} d\theta = \frac{\sqrt{2\pi}}{\sqrt{a}} \exp\left\{\frac{b^2}{2a}\right\} = \sqrt{2\pi\sigma^2} \exp\left\{\frac{\mu^2}{2\sigma^2}\right\}$$

Useful fact - 2

Another known fact which will turn out to be useful in the following derivations is the density of a Student-t random variable. We say that $T \sim T_g$ iff

$$T \overset{d}{=} \frac{Z}{\sqrt{\frac{W}{g}}}$$

where $Z \perp\!\!\!\perp W$, $Z \sim N(0,1)$, and $W \sim \chi^2_g$. In that case we say that T has a standard Student-t distribution with g degrees of freedom which has the following density

$$f_T(t|g) = \frac{\Gamma\left(\frac{g+1}{2}\right)}{\Gamma\left(\frac{g}{2}\right)} \frac{1}{\sqrt{\pi}} \frac{1}{\sqrt{g}} \left(1 + \frac{t^2}{g}\right)^{-\frac{g+1}{2}}$$

Useful fact - 3

From the previous distribution one can easily derive the (generalized) Student-t family with g degrees of freedom as well as location-scale parameters (m, s).

$$f_T(t|dof = g, loc = m, scale = s) = \frac{\Gamma\left(\frac{g+1}{2}\right)}{\Gamma\left(\frac{g}{2}\right)} \frac{1}{\sqrt{\pi}} \frac{1}{\sqrt{g}} \frac{1}{s} \left(1 + \frac{1}{g}\left(\frac{t-m}{s}\right)^2\right)^{-\frac{g+1}{2}}$$

It is also well-known that if $g > 1 \implies$ it admits a finite expectation $E[T] = m$ and if $g > 2 \implies V[T] = s^2 \frac{n}{n-2}$

useful facts and their consequences

With the previous useful facts we will show:

1. how Bayes rule works within the Normal sampling model with known variance σ^2 (or known precision $\nu = \frac{1}{\sigma^2}$)
2. the conjugacy of the Normal class of prior distributions for that sampling model

Conjugate analysis for Normal data

(data model) $Y = (Y_1, ..., Y_n) : \;\; Y_i \sim N(\theta, \sigma^2)$ (i.i.d) with σ^2 known
"+"
(prior) $\pi(\theta) = \pi_{\tilde{h}}(\theta) = N(\mu_0, \tau_0^2)[\tilde{h} = (\mu_0, \tau_0^2) = \tilde{h}^{prior}]$
$\overset{\Longrightarrow}{(Bayes)}$ (posterior)
$\pi(\theta|Y = (y_1, .., y_n)) = N(\mu_y, \tau_y^2) = \pi_{\tilde{h}}(\theta)[\tilde{h} = (\mu_y, \tau_y^2) = \tilde{h}^{post}]$
where

$$\begin{aligned}
\mu_y &= \frac{\sigma^2}{\sigma^2 + n\tau_0^2} \cdot \mu_0 + \frac{n\tau_0^2}{\sigma^2 + n\tau_0^2} \bar{y}_n = w\mu_0 + (1 - w)\bar{y}_n = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{n}{\sigma^2}\bar{y}_n}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}} \\
\tau_y^2 &= = \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\tau_0^2}} = \frac{1}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}} = \frac{\tau_0^2 \sigma^2}{n\tau_0^2 + \sigma^2} =
\end{aligned}$$

and

$$\bar{y}_n = \frac{1}{n} \sum_{i=1}^{n} y_i$$

Conjugate analysis for Normal data

We want to show how we update our prior or *initial* or *prior* state of uncertainty about the uknown parameter θ i.e. $\theta \sim \pi(\theta)$ into the *final* or *posterior* state of uncertainty $\theta|y_1, \ldots, y_n$ i.e. $\pi(\theta|y_1, \ldots, y_n)$

$$\theta \sim N(\mu_0, \nu_0) = \pi_{\tilde{\tilde{h}}}(\theta)[(\mu_0, \nu_0) = \tilde{\tilde{h}}^{prior}] \Longrightarrow$$

$$\Longrightarrow \theta|Y = (y_1, .., y_n) \sim \pi(\theta|y_1, \ldots, y_n) = N(\mu_y, \nu_y)[(\mu_y, \nu_y) = \tilde{\tilde{h}}^{posterior}]$$

If we parameterize the normal distribution of the statistical (conditional) model in terms of the single data precision parameter

$$\psi = \frac{1}{\sigma^2}$$

and, similarly, the prior normal distribution with the prior precision

$$\nu_0 = \frac{1}{\tau_0^2}$$

we get a suggestive formula for the prior-posterior updating parameter in the conjugate (Normal) class

$$\begin{aligned} \mu_y &= w \cdot \mu_0 + (1 - w) \cdot \bar{y}_n \\ \nu_y &= \nu_0 + n\psi \end{aligned}$$

Conjugate analysis for Normal data

The conjugate Bayesian analysis in the simplest normal model with known variance boils down to the updating from the prior hyperparameter in the (mean,precision) parameterization

$$h_{prior} = (\mu_0, \nu_0)$$

to the posterior hyperparameter

$$h_{post} = (\mu_y, \nu_y)$$

where

$$\begin{aligned} \mu_y &= w \cdot \mu_0 + (1 - w) \cdot \bar{y}_n \\ \nu_y &= \nu_0 + n\psi \end{aligned}$$

In these formulas $\theta_0 = \mu_0$ can be regarded as our *prior guess*, $\bar{y}_n$ is the sample mean (data evidence) and $w \in (0, 1)$ is a weight which corresponds to the relative weight of the prior information as measured in terms of its *prior strength* (prior precision parameter) relatively to the *data strength* as measured by n times the precision of each single observation in the data vector

$$w = \frac{\nu_0}{\nu_0 + n\psi}$$

Conjugate analysis for Normal data

The conjugate Bayesian analysis in the simplest normal model with known variance boils down to the updating from the prior hyperparameter in the (mean,precision) parameterization

$$h_{prior} = (\mu_0, \nu_0)$$

to the posterior hyperparameter

$$h_{post} = (\mu_y, \nu_y)$$

where

$$\begin{aligned}\mu_y &= w \cdot \mu_0 + (1-w) \cdot \bar{y}_n \\ \nu_y &= \nu_0 + n\psi \overset{PH-notation}{=} \tilde{\tau}_0^2 + n\tilde{\sigma}^2\end{aligned}$$

In these formulas $\theta_0 = \mu_0$ can be regarded as our *prior guess*, $\bar{y}_n$ is the sample mean (data evidence) and $w \in (0,1)$ is a weight which corresponds to the relative weight of the prior information as measured in terms of its *prior strength* (prior precision parameter) relatively to the *data strength* as measured by n times the precision of each single observation in the data vector

$$w = \frac{\nu_0}{\nu_0 + n\psi}$$

Conjugate analysis for Normal data

From PH book, yet another alternative, equivalent, reparameterization of the conjugate prior $\theta \sim \pi(\theta)$ where $\pi(\cdot)$ is $N(\mu_0, \tilde{\tau}_0^2 = \nu_0 = \kappa_0/\sigma^2)$. In this case the prior hyperparameter

$$h_{prior} = (\mu_0, \kappa_0)$$

once the n data points $y_1, \dots, y_n$ are observed, is turned into the posterior hyperparameter

$$h_{post} = (\mu_n, \kappa_n)$$

where

$$\begin{aligned} \mu_n &= w_n \cdot \mu_0 + (1 - w_n) \cdot \bar{y}_n \\ \kappa_n &= \kappa_0 + n \end{aligned}$$

$$w_n = \frac{\kappa_0}{\kappa_0 + n}$$

Conjugate analysis for Normal data - (both parameters unknown)

Suppose now that both parameters of the Normal statistical model are unknown

$$Y_i|(\theta, \lambda) \sim N\left(\theta, \lambda = \frac{1}{\sigma^2}\right)$$

Our conjugate prior on the parameter vector (θ, σ^2) or, equivalently on $(\theta, \lambda = \frac{1}{\sigma^2})$ will be specified hierarchically/sequentially/DAG

$$\pi(\theta, \lambda) = \pi(\theta|\lambda)\pi(\lambda)$$

More precisely,

$$\theta|\lambda \sim N(\mu_0, \kappa_0\lambda) = N\left(\mu_0, \kappa_0/\sigma^2\right)$$

and

$$\frac{1}{\sigma^2} = \lambda \sim Gamma\left(rate = \frac{\nu_0\sigma_0^2}{2}, shape = \frac{\nu_0}{2}\right)$$

Confusing parametrizations and conventions

We are used to say that

$$X \sim Gamma(rate = \alpha, shape = \beta) \quad \Longleftrightarrow \quad Y = \frac{1}{X} \sim InvGamma(\alpha, \beta)$$

Now it is easy to understand that if α represents a rate parameter for X it will be a scale parameter for Y.
So how should we name the first parameter of the Inverse Gamma?
We should better adopt the following convention: whenever we explicitly specify the parameter labels *rate* or *scale* they are referred to the present random variable (Inverse Gamma or, later on, Inverse Wishart) not from the originating ones. If we want to use the *rate* or *scale* terminology for the originating random variable we should explicitly mention this as follows:

$$Y \sim InvGamma(orig - rate = \alpha, \beta)$$

Conjugate analysis for Normal data - (both parameters unknown)

Let us recall our notation for a conjugate class of priors for a specific model characterized in terms of an unknown parameter ω. In our current setting the unknown parameter is the bivariate vector $\omega = (\theta, \lambda)$.

$$\mathcal{P} = \{\pi_h(\omega); h \in \mathcal{H}\}$$

More precisely, in our setting we have that the joint prior on (θ, λ) is characterized uniquely in terms of 4-dimensional hyperparameter $h = (\kappa_0, \mu_0, \nu_0, \sigma_0^2)$

$$\pi_h(\omega) = \pi_{(\kappa_0,\mu_0,\nu_0,\sigma_0^2)}(\theta, \lambda) = \pi_{\kappa_0,\mu_0}(\theta|\lambda)\pi_{\nu_0,\sigma_0^2}(\lambda)$$

Conjugate analysis for Normal data - (both parameters unknown)

It is interesting to derive the marginal distribution of θ in order to better understand the prior info conveyed through the hierarchical prior. Indeed,

$$\pi(\theta) = \int_0^\infty \pi(\theta|\lambda)\pi(\lambda)d\lambda = f_T\left(\theta \middle| \nu_0, \mu_0, \sqrt{\frac{\sigma_0^2}{\kappa_0}, \nu_0} \right)$$

$$\pi(\theta) = \int_0^\infty \pi(\theta|\sigma^2)\pi(\sigma^2)d\sigma^2 = f_T\left(\theta|\nu_0, \mu_0, \sqrt{\frac{\sigma_0^2}{\kappa_0}} \right)$$

Student t density

A random variable T has a Student-t distribution with degrees of freedom d and location-scale parameter (μ, σ) if it has the following density

$$f_T(t; d, \mu, \sigma) = \frac{\Gamma\left(\frac{d+1}{2}\right)}{\Gamma\left(\frac{d}{2}\right)\Gamma\left(\frac{1}{2}\right)} \frac{1}{\sqrt{d}} \frac{1}{\sigma} \left[1 + \frac{1}{d}\left(\frac{t-\mu}{\sigma}\right)^2\right]^{-\frac{d+1}{2}}$$

If $d > 1$ the expectation exists finite

$$E[T] = \mu$$

If $d > 2$ the variance exists finite

$$V[T] = \frac{d}{d-2}\sigma^2$$

When $d = 1$ the Student-t distribution coincides with the Cauchy distribution

Conjugate analysis for Normal data - (both parameters unknown)

Our conjugate prior on the parameter vector (θ, σ^2) will be specified hierarchically/sequentially/DAG

$$\pi(\theta, \sigma^2) = \pi(\theta|\sigma^2)\pi(\sigma^2)$$

More precisely,

$$\theta|\sigma^2 \sim N\left(\mu_0, \frac{\sigma^2}{\kappa_0}\right)$$

and

$$\sigma^2 \sim InvGamma\left(shape = \frac{\nu_0}{2}, origrate = \frac{\nu_0\sigma_0^2}{2}\right)$$

i.e

$$\lambda = \frac{1}{\sigma^2} \sim Gamma\left(rate = \frac{\nu_0\sigma_0^2}{2}, shape = \frac{\nu_0}{2}\right)$$

Conjugate analysis for Normal data - (both parameters unknown)

It is interesting to derive the marginal distribution of θ in order to better understand the prior info conveyed through the hierarchical prior. Indeed,

$$\begin{aligned}\pi(\theta) &= \int_0^\infty \pi(\theta|\sigma^2)\pi(\sigma^2)d\sigma^2 = \\ &= \int_0^\infty \pi(\theta|\lambda)\pi(\lambda)d\lambda = f_T\left(\theta\middle|\nu_0, \mu_0, \sqrt{\frac{\sigma_0^2}{\kappa_0}}\right)\end{aligned}$$

Conjugate analysis for Normal data - (both parameters unknown)

We end up with the following posterior update which again has the posterior distribution for the whole vector (θ, σ^2) with the same hierarchical decomposition

$$\pi(\theta, \sigma^2 | y_1, \ldots, y_n) = \pi(\theta | \sigma^2, y_1, \ldots, y_n) \pi(\sigma^2 | y_1, \ldots, y_n)$$

where

$$\theta | \sigma^2, y_1, \ldots, y_n \sim N\left(\mu_n, \frac{\sigma^2}{\kappa_n}\right)$$

with $\kappa_n = \kappa_0 + n$ and $\mu_n = \frac{\kappa_0}{\kappa_0+n}\mu_0 + \frac{n}{\kappa_0+n}\bar{y}_n$ and

$$\sigma^2 | y_1, \ldots, y_n \sim InvGamma\left(shape = \frac{\nu_n}{2}, origrate = \frac{\nu_n \sigma_n^2}{2}\right)$$

with $\nu_n = \nu_0 + n$ and $\sigma_n^2 = \frac{1}{\nu_n}\left[\nu_0 \sigma_0^2 + (n-1)s^2 + \frac{\kappa_0 n}{\kappa_0+n}(\bar{y}_n - \mu_0)^2\right]$

Conjugate analysis for Normal data - (both parameters unknown)

From our previous derivation on Normal-Inverse Gamma joint distribution we can also state that the marginal posterior distribution on the unknown θ is

$$\theta | y_1, \ldots, y_n \sim T_{\nu_n}\left(loc = \mu_n, scale = \sqrt{\frac{\sigma_n^2}{\kappa_n}} \right)$$

with

- $\nu_n = \nu_0 + n$
- $\mu_n = \frac{\kappa_0}{\kappa_0 + n}\mu_0 + \frac{n}{\kappa_0 + n}\bar{y}_n$
- $\sigma_n^2 = \frac{1}{\nu_n}\left[\nu_0 \sigma_0^2 + (n-1)s^2 + \frac{\kappa_0 n}{\kappa_0 + n}(\bar{y}_n - \mu_0)^2 \right]$
- $\kappa_n = \kappa_0 + n$

Some details for the Normal-InverseGamma (conjugate) prior

The functional kernel of the prior bivariate density $\pi(\theta, \sigma^2) = \pi(\theta|\sigma^2)\pi(\sigma^2)$ is (up to a proportionality constant)

$$(\sigma^2)^{-\frac{1}{2}} \exp\left\{-\frac{\kappa_0}{2\sigma^2}(\theta - \mu_0)^2\right\} (\sigma^2)^{-\frac{\nu_0}{2}-1} \exp\left\{-\frac{\nu_0\sigma_0^2}{2\sigma^2}\right\}$$

$$(\sigma^2)^{-\frac{\nu_0}{2}-1-\frac{1}{2}} \exp\left\{-\frac{\kappa_0}{2\sigma^2}(\theta - \mu_0)^2 - \frac{\nu_0\sigma_0^2}{2\sigma^2}\right\}$$

$$(\sigma^2)^{-\frac{\nu_0}{2}-1-\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma^2}\left[\kappa_0(\theta - \mu_0)^2 + \nu_0\sigma_0^2\right]\right\}$$

Some details for the Normal-InverseGamma (conjugate) prior

We would like to check that the functional kernel of the posterior bivariate density $\pi(\theta, \sigma^2 | y_1, ..., y_n) \propto L_{y_1, ... y_n}(\theta, \sigma^2)\pi(\theta|\sigma^2)\pi(\sigma^2)$ is (up to a proportionality constant w.r.t. both θ and σ^2))

$$(\sigma^2)^{\frac{-\nu_n}{2}-1-\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma^2}\left[\kappa_n\left(\theta - \mu_n\right)^2 + \nu_n\sigma_n^2\right]\right\}$$

for suitable values of the posterior hyperparameters
$H_n = (\kappa_n, \mu_n, \nu_n, \sigma_n^2)$

Some details for the Normal-InverseGamma (conjugate) prior

In fact the likelihood contribution can be expressed as follows:

$$
\begin{aligned}
& L_{y_1,\ldots y_n}(\theta,\sigma^2) \propto \\
& (\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\theta)^2\right\} \\
& (\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\left[(n-1)s_n^2+n(\theta-\bar{y}_n)^2\right]\right\} \\
& (\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\left[(n-1)s_n^2+n(\theta-\bar{y}_n)^2\right]\right\}
\end{aligned}
$$

which in turn makes

$$
\begin{aligned}
& L_{y_1,\ldots y_n}(\theta,\sigma^2)\pi(\theta,\sigma^2) \propto \\
& (\sigma^2)^{-\frac{n+\nu_0}{2}-1-\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma^2}\left[(n-1)s^2+n(\theta-\bar{y}_n)^2+\kappa_0\left(\theta-\mu_0\right)^2+\nu_0\sigma_0^2\right]\right\} \\
& (\sigma^2)^{-\frac{n+\nu_0}{2}-1-\frac{1}{2}} \exp\left\{\frac{1}{\sigma^2}\left[-\frac{n}{2}(\theta-\bar{y}_n)^2-\frac{\kappa_0}{2}\left(\theta-\mu_0\right)^2-\frac{(n-1)s^2+\nu_0\sigma_0^2}{2}\right]\right\}(\star)
\end{aligned}
$$

From this expression we clearly candidate/recognize the hyperparameter component $\nu_n = \nu_0 + n$ and we need to derive an appropriate quadratic form of the kind $(\theta - \mu_n)^2$ within the exponential square brackets.

Some details for the Normal-InverseGamma (conjugate) prior

We can then elaborate the two quadratic expressions as follows

$$-\frac{a}{2}\theta^2 + b\theta - \frac{b^2}{2a} - \frac{a'}{2}\theta^2 + b'\theta - \frac{b'^2}{2a'}$$

with $a = n$, $b = n\bar{y}_n$, $a' = \kappa_0$, $b' = \kappa_0\mu_0$. Hence, we obtain

$$-\frac{a''}{2}\theta^2 + b''\theta - \frac{b^2}{2a} - \frac{b'^2}{2a'}$$

which helps us rewriting the original expression as

$$-\frac{a''}{2}(\theta - \frac{b''}{a''})^2 + \frac{b''^2}{2a''} - \frac{b^2}{2a} - \frac{b'^2}{2a'}$$

with $a'' = n + \kappa_0$, $b'' = n\bar{y}_n + \kappa_0\mu_0$, so that $\frac{b''}{a''} = \frac{n\bar{y}_n + \kappa_0\mu_0}{n+\kappa_0} = \mu_n$

Some details for the Normal-InverseGamma (conjugate) prior

$$
\begin{aligned}
& \frac{b''^2}{2a''} - \frac{b^2}{2a} - \frac{b'^2}{2a'} \\
= \quad & \frac{(n\bar{y}_n + \kappa_0\mu_0)^2}{2(n+\kappa_0)} - \frac{n^2\bar{y}_n^2}{2n} - \frac{\kappa_0^2\mu_0^2}{2\kappa_0} \\
= \quad & \frac{n^2\bar{y}_n^2 + \kappa_0^2\mu_0^2 + 2n\bar{y}_n\kappa_0\mu_0 - n^2\bar{y}_n^2 - n\kappa_0\bar{y}_n^2 - n\kappa_0\mu_0^2 - \kappa_0^2\mu_0^2}{2(\kappa_0+n)} \\
= \quad & \frac{2n\bar{y}_n\kappa_0\mu_0 - n\kappa_0\bar{y}_n^2 - n\kappa_0\mu_0^2}{2(\kappa_0+n)} \\
= \quad & -\frac{(n\kappa_0)(\bar{y}_n-\mu_0)^2}{2(\kappa_0+n)}
\end{aligned}
$$

Some details for the Normal-InverseGamma (conjugate) prior

Finally $(\star)$ becomes

$$(\sigma^2)^{-\frac{n+\nu_0}{2}-1-\frac{1}{2}} \exp\Bigg\{\frac{1}{\sigma^2}\Bigg[-\frac{\kappa_0+n}{2}(\theta-\mu_n)^2 - \frac{(n\kappa_0)(\bar{y}_n-\mu_0)^2}{2(\kappa_0+n)} + \frac{(n-1)s^2+\nu_0\sigma_0^2}{2}\Bigg]\Bigg]\Bigg\}$$

$$(\sigma^2)^{-\frac{n+\nu_0}{2}-1-\frac{1}{2}} \exp\Bigg\{-\frac{\kappa_0+n}{2\sigma^2}(\theta-\mu_n)^2 - \frac{1}{2\sigma^2}\Bigg[\frac{(n\kappa_0)}{(\kappa_0+n)}(\bar{y}_n-\mu_0)^2 + (n-1)s^2 + \nu_0\sigma_0^2\Bigg]\Bigg]\Bigg\}$$

$$(\sigma^2)^{-\frac{\nu_n}{2}-1-\frac{1}{2}} \exp\left\{-\frac{\kappa_0+n}{2\sigma^2}(\theta-\mu_n)^2 - \frac{\nu_n}{2\sigma^2}\frac{\left[\frac{(n\kappa_0)}{(\kappa_0+n)}(\bar{y}_n-\mu_0)^2+(n-1)s^2+\nu_0\sigma_0^2\right]}{\nu_n}\right\}$$

$$(\sigma^2)^{-\frac{\nu_n}{2}-1-\frac{1}{2}} \exp\left\{-\frac{\kappa_0+n}{2\sigma^2}(\theta-\mu_n)^2 - \frac{\nu_n\sigma_n^2}{2\sigma^2}\right\}$$

How many predictive distributions in Bayesian inference?

Suppose we are in the usual context of conditionally i.i.d. observations and suppose the parameter (may be a vector) is $\omega \in \Theta$. In our previous normal model $\omega = (\theta, \sigma^2)$ and $\Theta = \mathbb{R} \times \mathbb{R}^+$

$$\begin{cases} \omega & \sim \pi(\omega) \\ Y_i|\omega & \sim f(y_i|\omega) \end{cases}$$

1. (initial/prior) predictive distribution for a single observable
2. (initial/prior) predictive distribution for a many observables
3. (final/posterior) predictive distribution for a new observation
4. (final/posterior) predictive distribution for multiple new observations

Conjugate analysis for Normal data - (both parameters unknown)

Implementation issues:

- How can we better understand the shape of the conjugate distribution?
- How can we visualize the dependence between θ and σ^2
- Can we simulate a sample from the joint posterior?
- Can we check the shape of the marginal distribution of $\theta | y_1, ..., y_n$? (conditionally on observed data)
- Can we visualize the predictive distribution?

Conjugate analysis for Normal data - (both parameters unknown)

Let us discuss the relation with the classical estimators and let us understand the role of the prior information for the frequentist properties.

- HPD and classical confidence interval for θ (with improper prior $1/\sigma^2$)
- Biasedness of the Bayesian point estimator
- Efficiency in terms of MSE
- Approximate coverage of HPD
- Admissibility

Discrete parameter model: Binomial capture

$$Z_i = \left\{ \begin{array}{c} \text{indicator of} \\ \text{one capture} \end{array} \right\} = \begin{cases} 1 & \text{individual } i \text{ has been captured} \\ 0 & \text{individual } i \text{ has NOT been captured} \end{cases}$$

$$Pr\left(Z_i = 1 | p\right) = p$$

We assume that the capture probability p is known: $p = p_0$

$$Z_1, Z_2, \ldots, Z_N | p_0 \sim Bern(p_0) \Longrightarrow S = \sum_{i=1}^{N} Z_i \bigg| N, p_0 \sim Bin(N, p_0)$$

- S is the experimental outcome and represents the total number of captured (observable) individuals
- N is the unknown polulation size and represents the parameter of interest

Binomial capture example

Parameter space

$N \in \Theta = \mathbb{N}$ can be a countably infinite set but we assume (a priori) that $N \in \Theta = \{1, 2, ..., N_{max}\}$.

Likelihood function

$$\begin{aligned} L_s(N) = Pr\left(S = s | N, p_0\right) &= \binom{N}{s} p_0^s (1 - p_0)^{N-s} I_{\{s,s+1,\dots,N_{max}\}}(N) \\ &= \binom{N}{s} p_0^s \frac{(1-p_0)^N}{(1-p_0)^s} I_{\{s,s+1,\dots,N_{max}\}}(N) \\ &= \binom{N}{s} \left(\frac{p_0}{1-p_0}\right)^s (1-p_0)^N I_{\{s,s+1,\dots,N_{max}\}}(N) \\ &\propto \frac{N!}{(N-s)!} (1-p_0)^N I_{\{s,s+1,\dots,N_{max}\}}(N) \end{aligned}$$

Prior specification

Discrete distribution within a bounded support: we can choose $N \sim Unif\{1, 2, ..., N_{max}\} \implies$

$$\pi(N) = \frac{1}{N_{max}} I_{\{1,2,\dots,N_{max}\}}(N)$$

Binomial capture example

Bayesian inference

Combines $\pi(N)$ and $L_s(N)$ and then yield $\Longrightarrow$ posterior $\pi(N|s)$

$$\begin{aligned}
\pi(N|s) &\propto \frac{N!}{(N-s)!}(1-p_0)^N I_{\{s,s+1,\ldots,N_{max}\}}(N)\frac{1}{N_{max}} I_{\{1,2,\ldots,N_{max}\}}(N) \\
&\propto \frac{N!}{(N-s)!}(1-p_0)^N I_{\{\max\{1,s\},\ldots,N_{max}\}}(N) \\
&\propto (N-s+1)\cdot \ldots . \cdot N \cdot (1-p_0)^N I_{\{\max\{1,s\},\ldots,N_{max}\}}(N)
\end{aligned}$$

Binomial capture example

How can we carry out our Bayesian inference with a single point estimate $\hat{\theta}$?

- posterior mean
- posterior median
- posterior mode[5]

They can all be easily derived numerically. Indeed we can also find a suitable closed form expression for one of them. Let us look for the posterior mode denoted as $Mo\left[N|S=s\right]$.
We look for N^* such that:

- for $N \leq N^*$ one has $\pi(N|s) \geq \pi(N-1|s)$
- for $N \geq N^*$ one has $\pi(N|s) \leq \pi(N-1|s)$

Hence we can discuss for which values of N we get

[5] in this prior setting where we elicited a uniform prior one immediately realize that $Mo\left[N|S=s\right] = \hat{\theta}_{MLE}$

Binomial capture example

$$\frac{\pi(N|s)}{\pi(N-1|s)} \geq 1 \Longleftrightarrow \frac{\frac{N!}{(N-s)!}(1-p_0)^N}{\frac{(N-1)!}{(N-1-s)!}(1-p_0)^{N-1}} \geq 1 \Longleftrightarrow$$

$$\Longleftrightarrow \frac{N}{N-s}(1-p_0) \geq 1 \Longleftrightarrow N - Np_0 \geq N - s \Longleftrightarrow$$

$$N \leq \frac{s}{p_0}$$

Hence it follows that

$$Mo\left[N|S=s\right] = \left\lfloor \frac{s}{p_0} \right\rfloor$$

Binomial capture example

- Can we compute the posterior mean?
- Can we draw a sample from the posterior distribution?
- Can I use a different discrete prior distribution other than the Uniform in a finite range?
- Can I provide an interval estimate?
- Can I carry out hypothesis testing?

Non-conjugate (partially conjugate) analysis for Normal data - (both parameters unknown)

Suppose now that we prefer specifying our prior distribution on the parameter vector $\omega = (\theta, \sigma^2)$ in the following natural way (independent-compontent prior):

$$\pi(\theta, \sigma^2) = \pi(\theta)\pi(\sigma^2)$$

so that $\theta \sim N\left(\mu_0, \tau_0^2\right)$ and $\sigma^2 \sim \mathit{InvGamma}\left(\frac{\nu_0}{2}, \frac{\nu_0 \sigma_0^2}{2}\right)$ and the joint posterior density of $\omega = (\theta, \sigma^2)$ can be written up to a multiplicative constant as follows

$$\text{target}(\omega) = \pi(\omega|\boldsymbol{y}) = \pi(\theta, \sigma^2|\boldsymbol{y}) \propto \pi(\theta)\pi(\sigma^2)L_{\boldsymbol{y}}(\theta, \sigma^2)$$

Can we recognize any standard bivariate distribution?

Non-conjugate (partially conjugate) analysis for Normal data - (both parameters unknown)

Suppose now that we have specified our prior distribution on the parameter vector $\omega = (\theta, \sigma^2)$ in the following way (independent-compontent prior):

$$\pi(\theta, \sigma^2) = \pi(\theta)\pi(\sigma^2)$$

so that $\theta \sim N\left(\mu_0, \tau_0^2\right)$ and $\sigma^2 \sim \mathit{InvGamma}\left(\frac{\nu_0}{2}, \frac{\nu_0 \sigma_0^2}{2}\right)$ and the joint posterior density of $\omega = (\theta, \sigma^2)$ can be written up to a multiplicative constant as follows

$$\text{target}(\omega) = \pi(\omega|\boldsymbol{y}) = \pi(\theta, \sigma^2|\boldsymbol{y}) \propto \pi(\theta)\pi(\sigma^2)L_{\boldsymbol{y}}(\theta, \sigma^2)$$

We **cannot** recognize any standard bivariate distribution! How can we carry out our Bayesian inference based on the posterior distribution of the unknown parameter $\omega = (\theta, \sigma^2)$?

Gibbs sampling - quick intro

- We need a viable solution for approximating all features of interest of the posterior distribution (e.g. the shape of the distribution, its expectation [point-estimate], hpd or equal-tail credible interval, posterior probability of subset of Θ)
- We understand that Monte Carlo techniques and its extensions (Monte Carlo Markov Chain) provide a suitable framework
- If we cannot simulate directly a suitable large number T of realizations of i.i.d. draws from $\pi(\theta|y_1, ..., y_n)$ we can try to simulate the first T components of a Markov Chain which enjoys the needed asymptotic/ergodic properties with respect to the target distribution $\pi(\theta|y_1, ..., y_n)$
- Gibbs sampling (under very mild conditions) is a viable algorithm which allows us to simulate the sequence of (Markov) dependent random variables which enjoys those ergodic properties

Gibbs sampling algorithm ingredients and procedure

Ingredients:

- a target multivariate distribution $\pi(\theta_1, ..., \theta_k)$, typically a density known up to a proportionality constant
- **all** the full conditionals of the target distribution $\pi(\cdot)$, namely $\pi(\theta_i|\theta_{(i)})$ for $i = 1, .., k$, are such that they can be simulated from

Algorithm:

- fix the starting values of the parameter components at time $t = 0$, say $\theta_1^0 = x_1$, $\theta_2^0 = x_2$,...., $\theta_k^0 = x_k$
- for $t = 0, 1, 2, ..., T$ iterate the following Gibbs cycle (loop):
  - $$\theta_j^{t+1} \sim \pi(\theta_j|\theta_1^{t+1}, ..., \theta_{j-1}^{t+1}, \theta_{j+1}^{t}, ..., \theta_k^{t}) \;\; j = 1, ..., k$$

Gibbs sampling with partially conjugate analysis for Normal data

We need (conditionally on the observed values $(y_1^{obs}, \ldots, y_n^{obs})$ to identify the 2 full conditionals:

- $\pi(\theta|\sigma^2)$
- $\pi(\sigma^2|\theta)$

It is straightforward to verify from the functional form of the joint posterior density of $\omega = (\theta, \sigma^2)$ (known up to a multiplicative constant)

$$\text{target}(\omega) = \pi(\omega|\mathbf{y}) = \pi(\theta, \sigma^2|\mathbf{y}) \propto \pi(\theta)\pi(\sigma^2)L_{\mathbf{y}}(\theta, \sigma^2)$$

that:

- $\theta|\sigma^2 \sim N\left(\mu_n(\sigma^2), \sigma_n(\sigma^2)\right)$
  $\Longleftarrow \pi(\theta|\mathbf{y}, \sigma^2) \propto \pi(\theta)\cancel{\pi(\sigma^2)}L_{\mathbf{y}}(\theta, \sigma^2) \propto \pi(\theta)L_{\mathbf{y}}(\theta, \sigma^2)$
- $\sigma^2|\theta \sim InvGamma\left(\frac{\nu_n}{2}, \frac{\nu_n\sigma_n^2(\theta)}{2}\right)$
  $\Longleftarrow \pi(\sigma^2|\mathbf{y}, \theta) \propto \cancel{\pi(\theta)}\pi(\sigma^2)L_{\mathbf{y}}(\theta, \sigma^2) \propto \pi(\sigma^2)L_{\mathbf{y}}(\theta, \sigma^2)$

Gibbs sampling with partially conjugate analysis for Normal data

$$L_{y_1,\ldots y_n}(\theta,\sigma^2)\pi(\sigma^2) \propto$$
$$(\sigma^2)^{-\frac{n+\nu_0}{2}-1}\exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(y_i-\theta)^2+\nu_0\sigma_0^2\right]\right\}$$

Gibbs sampling with partially conjugate analysis for Normal data

where, conditionally on the observed $y_1, \ldots, y_n$, for $\theta|\sigma^2$

- $\mu_n = \left(\frac{\frac{1}{\tau_0^2}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}} \right) \mu_0 + \left(\frac{\frac{n}{\sigma^2}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}} \right) \bar{y}$
- $\tau_n^2 = \left(\frac{1}{\tau_0^2} + \frac{n}{\sigma^2} \right)^{-1}$

while for , for $\sigma^2|\theta$

- $\nu_n = n + \nu_0$
- $\sigma_n^2 = \frac{1}{\nu_n}(\nu_0 \sigma_0^2 + SSR(\theta))$

and

$$SSR(\theta) = \sum_{i=1}^{n} (y_i - \theta)^2$$

is the Sum of Squares of Residuals w.r.t. θ

consider the computational trick for reducing the computational burden in updating each time the sum of a possible very large number of squares .... $SSR(\theta) = (n-1)s_n^2 + n(\bar{y} - \theta)^2$ exploiting the sufficient statistic $(\bar{y}, s^2)$

Remarks on the Gibbs sampling

- Note that the full conditional distribution for a single component of the parameter vector, e.g. $\pi(\theta|\sigma^2, y_1, ..., y_n)$ does not represent (is different from) the posterior distribution $\pi(\theta|y_1, ..., y_n)$
- It is interesting to represent the DAG of the GS simulation

Bayesian models for univariate observables Y_i

At the end of this first part of the course we should be able to understand and implement some fully Bayesian inference on the parameters of univariate distributions for n observables Y_i assumed to be conditionally i.i.d. (hence exchangeable) according to a parametric distribution. We should also be able to make predictions for future observables.

Bayesian models for univariate observables Y_i

In particular, we should be able to do the following:

- specify your Bayesian model: conditional distribution of the observables "+" prior
- elicit (and possibly motivate) your favorite prior
- possibly elaborate on your prior predictive/marginal distribution
- observe data $\boldsymbol{Y} = \boldsymbol{y}^{obs} = (y_1^{obs}, \ldots, y_n^{obs})$
- derive the corresponding posterior distribution
- graphical representation (joint, whenever possible, and/or marginals)
- derive interesting/relevant summaries of the corresponding posterior distribution (possibly using simulations)
- draw conclusions for hypothesis testing problems
- derive posterior predictive/marginal distribution

Alternative approximation strategies for integrals

- **standard numerical methods** - discretize and evaluate the integrand function over a finite grid of points (local approximation$\Longrightarrow$ global) → quadrature formulas → trapezoidal areas

$$\int g(\theta) d\theta$$

- **Monte Carlo methods** - based on large scale simulations of random variables with suitable distribution → asymptotic results

$$E_\pi \left[h(\theta) \right] = \int h(\theta)\pi(\theta) d\theta \text{ or } E_{\pi|\mathbf{x}} \left[h(\theta) \right] = \int h(\theta)\pi(\theta|x_1, \ldots, x_n) d\theta$$

- **other asymptotic/analytic strategies** - asymptotic approximation of the integrand function → Laplace method

$$\int c h(\theta) f(\mathbf{x}|\theta)\pi(\theta) d\theta$$

In vanilla Monte Carlo the integral is directly expressed as an expectation of a quantity of interest $h(\theta)$ where the distribution of θ (a posteriori) [which we assume known for the time being so that we avoid the approximation of the ratio of two integrals] →

$$\begin{aligned} h(\theta) &= \theta \\ h(\theta) &= (\theta - E[\theta|\boldsymbol{x}])^2 \\ h(\theta) &= L(\theta, a) \qquad (L(\theta, a) \text{ is a loss function }) \\ h(\theta) &= I_C(\theta) \text{ e.g for } \theta = (\theta_1, \theta_2) \text{ and } C = \{\theta_1 < \theta_2\} \end{aligned}$$

Sometimes both numerator and denominator (especially this second) could be of specific interest (marginal L. or B. evidence)!
In the next slides we will refer to $\pi(\theta)$ as a probability distribution with respect to (w.r.t.) which integral quantity of interest is expressed (prior or posterior) or indeed integrated.

$$
\begin{aligned}
E_{\pi(\theta|x)}[h(\theta)] &= \int_{\Theta} \theta\pi(\theta|x)d\theta = \int_{\Theta} \theta\pi(d\theta|x) \\
E_{\pi(\theta|x)}[h(\theta)] &= \int_{\Theta} (\theta - E[\theta|\boldsymbol{x}])^2\pi(\theta|x)d\theta \\
E_{\pi(\theta|x)}[h(\theta)] &= \int_{\Theta} L(\theta, a)\pi(\theta|x)d\theta \qquad (\text{ posterior loss }) \\
E_{\pi(\theta|x)}[h(\theta)] &= \int_{\Theta} I_C(\theta)\pi(\theta|x)d\theta = Pr\{\theta \in C|x\}
\end{aligned}
$$

Monte Carlo Method

It relies on the integral quantity expressed as expectation

$$I = \int_\Theta g(\theta) d\theta = \int_\Theta h(\theta)\pi(\theta) d\theta = E_\pi[h(\theta)]$$

and on limit theorems which ensure that appropriate functions of sequences of i.i.d. random variables (i.i.d stochastic process)

If $|I| < \infty$ and $\theta_1, ..., \theta_t$ i.i.d. $\sim \pi$ then from Strong Law of Large Numbers (SLLN) the empirical average is a consistent ”estimator” of I or, more precisely, a a consistent sequence of ”estimators” of I

$$\hat{I}_t = \frac{1}{t}\sum_{i=1}^{t} h(\theta_i) \xrightarrow{a.s.} E_\pi[h(\theta)] = I \qquad t \rightarrow \infty$$

You are probably more familiar with this version of the SLLN.
SLLN (basic) If $Y_1, ..., Y_n, ...$ are i.i.d. with $Y_i \sim f_Y(y)$ and $\mu = E[Y_i]$ then

$$\bar{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i \xrightarrow{a.s.} \mu = E[Y] = \int y\, f_Y(y)\, dy$$

provided the existence and finiteness of $E_{f_Y}[Y_i]$
Indeed we can also restate the result in a slightly more general form dealing with expected values of i.i.d. random variables which are function $h(\cdot)$ of another sequence of underlying random variables X_i.
SLLN If $X_1, ..., X_n$ are i.i.d. with $X_i \sim f_X(x)$ and $I = E[h(X_i)]$ then

$$\frac{1}{n}\sum_{i=1}^{n} h(X_i) \xrightarrow{a.s.} E[h(X)] = I$$

provided the existence and finiteness of $E[h(X_i)]$

In fact the two versions of SLLN are equivalent since

- if $(X_1, \dots, X_n, \dots)$ are i.i.d. then also $(Y_1 = h(X_1), \dots, Y_n = h(X_n), \dots)$ are i.i.d.
- if $E[h(X_i)]$ exists and is finite it can be written as $E[Y_i]$ where

$$\mu = \int y\, f_Y(y)\, dy = E[Y_i] = E[h(X_i)] \overset{(*)}{=} \int h(x)\, f_X(x)\, dx = I$$

(*) a property of the Expectation of a function of a random variable

1. the use of a random value $\hat{I}$ as a way of getting an unknown quantity → typical inferential attitude!!
2. we must ensure a proper **error control** ...
3. when interested to posterior summaries we would rather use two separate approximations for numerator and denominator which are both integral quantities with respect to the prior but I could avoid the presence of two sources of errors if I can directly simulate from the posterior ...
4. Fundamental ingredient will be the ability of simulating random draws from some target distribution $\longrightarrow \pi(\cdot)$ ... possibly known up to a proportionality constant

Error evaluation

$$
\begin{aligned}
E[(\hat{I}-I)^2] = Var[\hat{I}] &= \frac{1}{n} Var[h(X)] = \\
&= \frac{1}{n}\left\{E_\pi[h(X)^2] - E_\pi[h(X)]^2\right\} = \\
&= \frac{K}{n}
\end{aligned}
$$

Indeed I do not know K but I can approximate/estimate K as well, as follows

$$\hat{K} = \widehat{Var}[h(X)] = \frac{1}{n}\sum_{i=1}^{n} h(X_i)^2 - \hat{I}_n^2$$

I can do better when there are sufficient conditions to apply Central Limit Theorem (CLT) and hence I can get a confidence interval for I (probabilistic approximation)

Error evaluation [1]

Unbiasedness of $\hat{I}_n$ is not enough. We are interested in knowing how much $\hat{I}_n$ is close (in distribution) to our target I.
First immediate idea is the use of MSE (Mean Square Error) which in this case coincides with $\hat{I}_n$

$$E[(\hat{I}_n - I)^2] \quad = \quad Var[\hat{I}_n] = \frac{1}{n}\{\mathbb{V}[h(X)]\} = \frac{1}{n}\left\{\mathbb{E}[h(X)^2] - \mathbb{E}[h(X)]^2\right\}$$

I can estimate $K = \mathbb{V}[h(X)]$ with analogous MC approach using the empirical variance of $h(X_i)$

$$\hat{K} = \frac{1}{n}\sum_{i=1}^{t} h(X_i)^2 - \hat{I}_n{}^2$$

(Consistent? unbiased?)

Error evaluation [2]

I can go even further when CLT can be applied since I can get asymptotic confidence intervals for I using the asymptotic distribution of the empirical average $\hat{I}_n$

$$\left[\hat{I}_n - 1.96 \cdot \sqrt{\frac{\hat{K}}{n}}\ ,\ \ \hat{I}_n + 1.96 \cdot \sqrt{\frac{\hat{K}}{n}}\right]$$

or just the thumb rule $\pm 2 \times s.e.$

$$\left[\hat{I}_n - \sqrt{\frac{\hat{K}}{n}}\ ,\ \ \hat{I}_n + \sqrt{\frac{\hat{K}}{n}}\right]$$

Gibbs Sampling

In the most simple case where $\boldsymbol{\omega} = (\omega_1, \omega_2) \in \omega \subset \mathbb{R}^2$
Suppose that

1. our target is $\pi(\boldsymbol{\omega}) = \pi(\omega_1, \omega_2)$;
2. but we do not know how to simulate from π (2-dim);
3. but we know how to simulate from the so called (full conditionals) (1-dim):

$$\pi_1(\omega_1|\omega_2) \ ; \ \pi_2(\omega_2|\omega_1)$$

GS → we define our kernel through the following steps
$q(\boldsymbol{x}, \boldsymbol{\omega}) = f(\boldsymbol{\omega}|\boldsymbol{x})$

$$f(\omega|x) = f((\omega_1, \omega_2)|(x_1, x_2)) = f(\omega_1, \omega_2|x_1, x_2) = \pi_2(\omega_2|\omega_1)\pi_1(\omega_1|x_2)$$

We verify that π is invariant for this GS kernel

$$\int_{\mathbb{R}^2} q(\boldsymbol{x},\boldsymbol{\omega})\pi(\boldsymbol{x})dx \int_{\mathbb{R}^2} f(\boldsymbol{\omega}|\boldsymbol{x})\pi(\boldsymbol{x})dx = \pi(\boldsymbol{\omega}) \quad \forall \boldsymbol{\omega} \in \omega \subset \mathbb{R}^2$$

In fact

$$\begin{aligned}
\int_{\mathbb{R}}\int_{\mathbb{R}} f(\omega_1,\omega_2|x_1,x_2)\pi(x_1,x_2)dx_1dx_2 &= \\
\int_{\mathbb{R}}\int_{\mathbb{R}} \pi_2(\omega_2|\omega_1)\pi_1(\omega_1|x_2)\pi(x_1,x_2)dx_1dx_2 &= \\
\pi_2(\omega_2|\omega_1)\int_{\mathbb{R}} \pi_1(\omega_1|x_2)\pi_2(x_2)dx_2 &= \\
\pi_2(\omega_2|\omega_1)\int_{\mathbb{R}} \pi(\omega_1,x_2)dx_2 &= \\
\pi_2(\omega_2|\omega_1)\pi_1(\omega_1) &= \pi(\omega_1,\omega_2)
\end{aligned}$$

Bivariate Gaussian distribution - toy example

We will use $\sigma_{12} = \rho\sigma_1\sigma_2$. Moreover we consider the vector $\boldsymbol{\theta} = (\theta_1, \theta_2)$ instead of $\boldsymbol{y} = (y_1, y_2)$. Hence

$$f(\theta_1, \theta_2; \mu_1, \mu_1, \sigma_1^2, \sigma_2^2, \sigma_{12}) =$$

$$\frac{1}{2\pi\sqrt{\sigma_1^2\sigma_2^2(1-\rho^2)}} exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(\theta_1-\mu_1)^2}{\sigma_1^2} + \frac{(\theta_2-\mu_2)^2}{\sigma_2^2} - 2\rho\frac{(\theta_1-\mu_1)(\theta_2-\mu_2)}{\sigma_1\sigma_2}\right]\right\}$$

Bivariate Gaussian distribution

Special case with $\mu_1 = 0, \mu_2 = 0, \sigma_1^2 = 1, \sigma_2^2 = 1$

$$f(\theta_1, \theta_2; \mu_1 = 0, \mu_2 = 0, \sigma_1^2 = 1, \sigma_2^2 = 1, \sigma_{12} = \rho) = \frac{1}{2\pi\sqrt{1-\rho^2}} exp\left\{-\frac{1}{2(1-\rho^2)}\left[\theta_1^2 + \theta_2^2 - 2\rho\theta_1 \cdot \theta_2\right]\right\}$$

Bivariate Gaussian distribution

If our target is the bivariate normal density in the particular case where $\mu_1 = \mu_2 = 0$, $\sigma_1^2 = \sigma_2^2 = 1$ and $\sigma_{12} = \sigma_{21} = \rho$ with a fixed $\rho \in (0,1)$ then it is easy (we have done that together in class) to derive the 2 types of full conditional distributions. More precisely,

$$\begin{aligned} \theta_1|\theta_2 = \bar{\theta}_2 &\sim N(\rho\bar{\theta}_2, 1-\rho^2) \\ \theta_2|\theta_1 = \bar{\theta}_1 &\sim N(\rho\bar{\theta}_1, 1-\rho^2) \end{aligned}$$

Bivariate Gaussian distribution

We can see what happens if we implement the GS algorithm trying to understand what is going on if:

- we start the chain at $\boldsymbol{\theta}^{\mathbf{0}} = (\bar{\theta}_1^0, \bar{\theta}_1^0) = (0, 0)$
- then we start once again the chain at a different $\boldsymbol{\theta}^{\mathbf{0}} = (\bar{\theta}_1^0, \bar{\theta}_1^0) = (-100, 100)$ [can you tell the difference?]
- we start randomly from $\theta^0 \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$
- we consider how GS behaves when we have a different target with a different value of ρ (s.t. $|\rho| \approx 1$)

Toy Example (GS1) - Correlated bivariate normal

```
gibbs.biv.norm<-function(nsim,theta1,theta2,rho){
  theta1vec <- rep(NA,nsim+1)
  theta2vec <- rep(NA,nsim+1)
  theta1vec[1] <- theta1
  theta2vec[1] <- theta2
  for(t in 1:nsim){
    theta1vec[t+1]<-rnorm(1,mean=rho*theta2vec[t],
                            sd=sqrt(1-rho^2))
    theta2vec[t+1]<-rnorm(1,mean=rho*theta1vec[t+1],
                            sd=sqrt(1-rho^2))
  }
  gibbssample<-cbind(theta1vec,theta2vec)
  return(gibbssample)
}
```

Toy Example (GS1) - Correlated bivariate normal

Although this is a toy example (you can simulate directly from a multivariate Normal [how-to]) it can be very instructive for beginners:

- learn how to single out f.c. from the joint
- the starting point
- shape of the joint
- simplicity and rigidity of the f.c. scheme
- effect of ρ on the previous items
- trace plots
- acf

Gibbs Sampling: general case

Some preliminary notation when $\omega = (\omega_1, \dots, \omega_i, \dots, \omega_k) \in \omega \subset \mathbb{R}^k$. Given a target $\pi(\omega) = \pi(\omega_1, \dots, \omega_k)$ the corresponding full conditionals will be denoted with

$$\pi(\omega_i | \omega_{(i)}) = \pi(\omega_i | \omega_1, \omega_2, \dots, \omega_{i-1}, \omega_{i+1}, \dots, \omega_k)$$

GS → We start at time $t = 0$ with a fixed initial point $\omega^0 = x = (x_1, \dots, x_k)$ or, equivalently, with initial distribution $\mu = \delta_{\{x\}}$
In the nest time $t = 1$ the state ω^1 will be simulated through the following sequence of (typically univariate, but not necessarily) simulations ...

$$\begin{array}{lllll}
\omega_1^1 & \sim & \pi(\omega_1|x_2,\ldots,x_k) & \rightarrow & \omega_1^1=\bar{\omega}_1^1 \\
\omega_2^1 & \sim & \pi(\omega_2|\bar{\omega}_1^1,x_3,\ldots,x_k) & \rightarrow & \omega_2^1=\bar{\omega}_2^1 \\
\omega_3^1 & \sim & \pi(\omega_3|\bar{\omega}_1^1,\bar{\omega}_2^1,x_4,\ldots,x_k) & \rightarrow & \omega_3^1=\bar{\omega}_3^1 \\
.. & .. & .. & .. & .. \\
\omega_k^1 & \sim & \pi(\omega_k|\bar{\omega}_1^1,\ldots,\bar{\omega}_{k-1}^1) & \rightarrow & \omega_k^1=\bar{\omega}_k^1
\end{array}$$

same sequential procedure (with k intermediate steps) for the next times $t = 2, 3, \ldots$.

In summary the Markov transition kernel corresponding to the Gibbs Sampling is specified in terms of a collection of k full conditionals. More precisely

$$\begin{array}{lcllll}
\omega_1^{t+1} & \sim & \pi(\omega_1|\omega_{(1)}) & = & \pi(\omega_1|\omega_2^t, ...\omega_k^t) = \\
\omega_2^{t+1} & \sim & \pi(\omega_2|\omega_{(2)}) & = & \pi(\omega_2|\omega_1^{t+1}, \omega_3^t...\omega_k^t) \\
.. & .. & .. & .. & .. \\
\omega_k^{t+1} & \sim & \pi(\omega_k|\omega_{(k)}) & = & \pi(\omega_k|\omega_1^{t+1}, \omega_2^{t+1}, ..., \omega_{k-1}^{t+1})
\end{array}$$

If we had to write it more explicitly .... it would be .... like a transition kernel in k steps of a non-homogeneous Markov chain ...

Some advantages:

1. GS requires the simulation of a sequence of distributions (full-conditionals [f.c.]) which are defined over a lower-dimensional spaces rather than on the k-dimensional Ω; usually f.c. are univariate although they need not to be.
2. In Bayesian inference one can easily determine the full conditionals looking at the functional form of the posterior $\pi(\omega|Data)$ simply regarded as a function of one component

$$\pi(\omega_i|\omega_{(i)}, Data) \propto \pi(\omega|Data) \propto \pi(\omega)\ell(\omega; Data)$$

Sometimes one can recognize that the functional form of the full-conditional corresponds to some well-known parametric family of distributions which can be easily simulated from

GS & Markov Chains properties

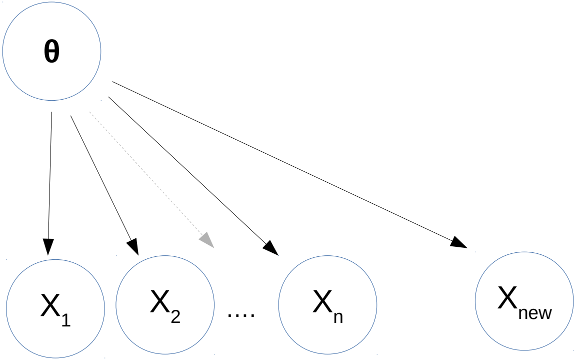
1.

Notice that we build-up our algorithm so that the resulting Markov chain has a kernel and hence a probability law such that at least one important property of the Markov Chain is fulfilled, namely the property that the (transition kernel of the) Markov Chain has the target $\pi(\cdot)$ as its invariant distribution.

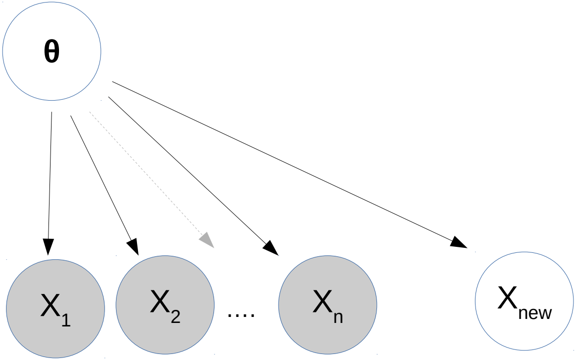
2.

Notice also that we must distinguish between parameter estimation and posterior approximation (litterally quoting from PH)

DAG before data observation



DAG after data observation



Conjugate analysis for Normal data - (both parameters unknown)

Let us discuss the relation with the classical estimators and let us understand the role of the prior information for the frequentist properties.

- HPD and classical confidence interval for θ (with improper prior $1/\sigma^2$)
- Biasedness of the Bayesian point estimator
- Efficiency in terms of MSE
- Approximate coverage of HPD
- Admissibility

Bayesian models for univariate observables Y_i

- specify your Bayesian model: conditional distribution of the observables "+" prior
- elicit (and possibly motivate) your favorite prior
- possibly elaborate on your prior predictive/marginal distribution
- observe data $\boldsymbol{Y} = \boldsymbol{y}^{obs} = (y_1^{obs}, ..., y_n^{obs})$
- derive the corresponding posterior distribution
- graphical representation (joint, whenever possible, and/or marginals)
- derive interesting/relevant summaries of the corresponding posterior distribution (possibly using simulations)
- derive posterior predictive/marginal distribution

Multivariate normal model

$$p_{\boldsymbol{y}}(y_1, \dots, y_p | \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{p}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} exp\left\{-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{\mu})\right\}$$

Meaning of parameters. Restrictions and dimensionality of parameter space. Properties[6]:

P1 linear transformations with $k \times p$ full-rank matrix $\boldsymbol{L}$ with $k \leq p$:

$$\boldsymbol{L}\boldsymbol{y} \sim MVN_k(\boldsymbol{L}\boldsymbol{\mu}, \boldsymbol{L}\boldsymbol{\Sigma}\boldsymbol{L}^T)$$

(indeed, this is a characterizing property when univariate linear transforms (i.e. $k = 1$) are considered)

P2 marginal distributions $\boldsymbol{y}_{[\boldsymbol{b}]} \sim MVN_b(\boldsymbol{\mu}_{[\boldsymbol{b}]}, \boldsymbol{\Sigma}_{[\boldsymbol{b},\boldsymbol{b}]})$

P3 conditional distributions of partitioned sub vectors using two disjoint integer index sets $\boldsymbol{a}$ and $\boldsymbol{b}$ such that $\boldsymbol{a} \cup \boldsymbol{b} = \{1, 2, \dots, p\}$:

$$\boldsymbol{y}_{[\boldsymbol{b}]} | \boldsymbol{y}_{[\boldsymbol{a}]}, \boldsymbol{\mu}, \boldsymbol{\Sigma} \sim MVN_b(\boldsymbol{\mu}_{\boldsymbol{b}|\boldsymbol{a}}, \boldsymbol{\Sigma}_{\boldsymbol{b}|\boldsymbol{a}})$$

[6] K. Mardia, J.T. Kent, J.M. Bibby (1979) Multivariate analysis. Academic Press

Multivariate normal model

… where

$$\boldsymbol{\mu}_{\boldsymbol{b}|\boldsymbol{a}} \equiv \boldsymbol{\mu}_{[\boldsymbol{b}]} + \boldsymbol{\Sigma}_{[\boldsymbol{b},\boldsymbol{a}]}(\boldsymbol{\Sigma}_{[\boldsymbol{a},\boldsymbol{a}]})^{-1}\left(\boldsymbol{y}_{[\boldsymbol{a}]} - \boldsymbol{\mu}_{[\boldsymbol{a}]}\right)$$

and

$$\boldsymbol{\Sigma}_{\boldsymbol{b}|\boldsymbol{a}} \equiv \boldsymbol{\Sigma}_{[\boldsymbol{bb}]} - \boldsymbol{\Sigma}_{[\boldsymbol{ba}]}\boldsymbol{\Sigma}_{[\boldsymbol{aa}]}^{-1}\boldsymbol{\Sigma}_{[\boldsymbol{ab}]}$$

Multivariate normal model

Subset index notation with examples. Let us consider $p = 5$, $\boldsymbol{a} = \{3, 4\}$ and $\boldsymbol{b} = \{1, 2, 5\}$. Let

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \sigma_{14} & \sigma_{15} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} & \sigma_{24} & \sigma_{25} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} & \sigma_{34} & \sigma_{35} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_{44} & \sigma_{45} \\ \sigma_{51} & \sigma_{52} & \sigma_{53} & \sigma_{54} & \sigma_{55} \end{pmatrix}$$

- $b = card\, \boldsymbol{b} = 3$;
- $\boldsymbol{y}_{[\boldsymbol{b}]} = (y_1, y_2, y_5)$;

Multivariate normal model

Subset index notation with examples.

- $$\Sigma_{[\boldsymbol{a},\boldsymbol{a}]} = \begin{pmatrix} \sigma_{33} & \sigma_{34} \\ \sigma_{43} & \sigma_{44} \end{pmatrix}$$

- $$\Sigma_{[\boldsymbol{a},\boldsymbol{b}]} = \begin{pmatrix} \sigma_{31} & \sigma_{32} & \sigma_{35} \\ \sigma_{41} & \sigma_{42} & \sigma_{45} \end{pmatrix}$$

- $$\Sigma_{[\boldsymbol{b},\boldsymbol{a}]} = \begin{pmatrix} \sigma_{13} & \sigma_{14} \\ \sigma_{23} & \sigma_{24} \\ \sigma_{53} & \sigma_{54} \end{pmatrix}$$

- $$\Sigma_{[\boldsymbol{b},\boldsymbol{b}]} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{15} \\ \sigma_{21} & \sigma_{22} & \sigma_{25} \\ \sigma_{51} & \sigma_{52} & \sigma_{55} \end{pmatrix}$$

Useful fact - 4

For the formal rigorous derivation of results related to the conjugate normal model the following known property extends the univariate case. If we have a multivariate p-variate density $p(\boldsymbol{\theta})$ such that for a suitable $p \times p$ matrix $\boldsymbol{A}$ and a $p \times 1$ vector $\boldsymbol{b}$

$$p(\boldsymbol{\theta}) \propto exp\left\{-\frac{1}{2}\boldsymbol{\theta}^T\boldsymbol{A}\boldsymbol{\theta} + \boldsymbol{b}^T\boldsymbol{\theta}\right\} = exp\left\{-\frac{1}{2}\left(\boldsymbol{\theta}^T\boldsymbol{A}\boldsymbol{\theta} - 2\boldsymbol{b}^T\boldsymbol{\theta}\right)\right\}$$

or, equivalently

$$p(\boldsymbol{\theta}) \propto exp\left\{-\frac{1}{2}\boldsymbol{\theta}^T\boldsymbol{A}\boldsymbol{\theta} + \boldsymbol{\theta}^T\boldsymbol{b}\right\} = exp\left\{-\frac{1}{2}\left(\boldsymbol{\theta}^T\boldsymbol{A}\boldsymbol{\theta} - 2\boldsymbol{\theta}^T\boldsymbol{b}\right)\right\}$$

then the corresponding distribution is

$$MVN_p\left(\tilde{\boldsymbol{\mu}} = \boldsymbol{A}^{-1}\boldsymbol{b}\ ,\ \tilde{\boldsymbol{\Sigma}} = \boldsymbol{A}^{-1}\right)$$

The previous facts will help us in rigorously deriving (in a possibly inefficient, but instructive way) the following property.

Proposition – If $\boldsymbol{L}$ is an invertible $p \times p$ matrix and $\boldsymbol{y} \sim MVN_p\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$ then

$$\boldsymbol{z} = \boldsymbol{L}\boldsymbol{y} \sim MVN_p\left(\boldsymbol{L}\boldsymbol{\mu}, \boldsymbol{L}\boldsymbol{\Sigma}\boldsymbol{L}^T\right)$$

Proof:

.

Useful fact - 4 - sketch of the proof

We consider only the particular case where we have a non singular L of order $p \times p$.

- general one-to-one random vector transformation rule $\boldsymbol{z} = g(\boldsymbol{y})$ with $\boldsymbol{z} \in \mathcal{Z}$.
- $$p_{\boldsymbol{z}}(\boldsymbol{z}) = p_{\boldsymbol{y}}(g^{-1}(\boldsymbol{z}))\,|det(J(\boldsymbol{z}))|$$
- here $\boldsymbol{z} = g(\boldsymbol{y}) = L\boldsymbol{y}$
- argue that the inverse function $\boldsymbol{y} = g^{-1}(\boldsymbol{z}) = L^{-1}\boldsymbol{z}$ is a linear transform as well ...
- .. and that the Jacobian matrix $J = J(\boldsymbol{z}) = L^{-1}$ since it has generic entry $J_{ij} = J_{ij}(\boldsymbol{z}) = \frac{\partial y_i}{\partial z_j} = \frac{\partial (g^{-1})_i(z_1,\dots,z_j,\dots,z_p)}{\partial z_j} = \frac{\partial}{\partial z_j} \sum_{r=1}^{p} L^{-1}_{ir} y_r = L^{-1}_{ij}$.

Alternatively one can overlook constants and tediously verify that, after some matrix algebra manipulation, the following holds true

$$\begin{aligned} p_{\boldsymbol{z}}(\boldsymbol{z}) &\propto exp\left\{-\frac{1}{2}\left(\boldsymbol{z}^T\boldsymbol{A}\boldsymbol{z} - 2\boldsymbol{b}\boldsymbol{z}\right)\right\} \propto exp\left\{-\frac{1}{2}\boldsymbol{z}^T\boldsymbol{A}\boldsymbol{z} + \boldsymbol{b}\boldsymbol{z}\right\} \\ &\propto exp\left\{-\frac{1}{2}\boldsymbol{z}^T\boldsymbol{A}\boldsymbol{z} + \boldsymbol{z}^T\boldsymbol{b}^T\right\} \end{aligned}$$

where $\boldsymbol{A} = (L\boldsymbol{\Sigma}L^T)^{-1}$ and $\boldsymbol{b} = \boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}L^{-1}$

Useful fact - 4 - sketch of the proof (2)

Hence it is a kernel of a Multivariate Normal random vector in p dimensions such that the corresponding mean vector $\tilde{\boldsymbol{\mu}}$ and variance covariance matrix $\tilde{\boldsymbol{\Sigma}}$ are

$$\tilde{\boldsymbol{\Sigma}} = \boldsymbol{A}^{-1} = \left((L\boldsymbol{\Sigma}L^T)^{-1}\right)^{-1} = L\boldsymbol{\Sigma}L^T$$

and

$$\begin{aligned}
\tilde{\boldsymbol{\mu}} &= \boldsymbol{A}^{-1}\boldsymbol{b} = L\boldsymbol{\Sigma}L^T(\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}L^{-1})^T = \\
&= L\boldsymbol{\Sigma}L^T(L^{-1})^T(\boldsymbol{\Sigma}^{-1})^T\boldsymbol{\mu} = \\
&= L\boldsymbol{\Sigma}L^T(L^T)^{-1}(\boldsymbol{\Sigma}^{-1})\boldsymbol{\mu} = L\boldsymbol{\mu}
\end{aligned}$$

The previous proposition can be generalized (and can be actually proved more easily using the characterizing property of a MVN random vector)

Proposition – If $\boldsymbol{L}$ is a full-rank $k \times p$ matrix and $\boldsymbol{y} \sim MVN_p\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$ then

$$\boldsymbol{z} = \boldsymbol{L}\boldsymbol{y} \sim MVN_k\left(\boldsymbol{L}\boldsymbol{\mu}, \boldsymbol{L}\boldsymbol{\Sigma}\boldsymbol{L}^T\right)$$

In fact, all the previous matrix manipulation has been a useful excuse .... but we could prove the more general result alternatively using the characterizing properties of multivariate normal vectors ...

With this last more general proposition it is easy to derive the marginal distribution of a multivariate normal vector. Check the case of y_1 with $k = 1$ and $\boldsymbol{L} = (1, 0, 0, ..., 0)$ and generalize the argument to an arbitrary subvector $y_{i_1}, ..., y_{i_k}$.

For instance, write the matrix L you would need to get the first $q < p$ components of the random vector $\boldsymbol{y}$.

Other useful properties are those related to square partitioned matrices denoted as follows:

$$\boldsymbol{\Sigma} = \left[\begin{array}{c|c} \boldsymbol{\Sigma}_{[\boldsymbol{a},\boldsymbol{a}]} & \boldsymbol{\Sigma}_{[\boldsymbol{a},\boldsymbol{b}]} \\ \hline \boldsymbol{\Sigma}_{[\boldsymbol{b},\boldsymbol{a}]} & \boldsymbol{\Sigma}_{[\boldsymbol{b},\boldsymbol{b}]} \end{array}\right]$$

The following properties may be useful in the following:

- (determinant)

$$|\boldsymbol{\Sigma}| = \left|\boldsymbol{\Sigma}_{[\boldsymbol{a},\boldsymbol{a}]}\right| \left|\boldsymbol{\Sigma}_{[\boldsymbol{b},\boldsymbol{b}]} - \boldsymbol{\Sigma}_{[\boldsymbol{b},\boldsymbol{a}]}\boldsymbol{\Sigma}^{-1}_{[\boldsymbol{a},\boldsymbol{a}]}\boldsymbol{\Sigma}_{[\boldsymbol{a},\boldsymbol{b}]}\right|$$

- (inverse) If we denote the partitioned inverse (provided it exists)

$$\boldsymbol{\Sigma}^{-1} = \left[\begin{array}{c|c} \boldsymbol{\Sigma}^{[\boldsymbol{a},\boldsymbol{a}]} & \boldsymbol{\Sigma}^{[\boldsymbol{a},\boldsymbol{b}]} \\ \hline \boldsymbol{\Sigma}^{[\boldsymbol{b},\boldsymbol{a}]} & \boldsymbol{\Sigma}^{[\boldsymbol{b},\boldsymbol{b}]} \end{array}\right]$$

Note that $\boldsymbol{\Sigma}^{[\boldsymbol{a},\boldsymbol{a}]}$ is not the same symbol of $\boldsymbol{\Sigma}_{[\boldsymbol{a},\boldsymbol{a}]}$! Note that $\boldsymbol{\Sigma}^{[\boldsymbol{a},\boldsymbol{a}]}$ is not (in general) the inverse of $\boldsymbol{\Sigma}_{[\boldsymbol{a},\boldsymbol{a}]}$!

with

$$\boldsymbol{\Sigma}^{[\boldsymbol{a},\boldsymbol{a}]} = \left(\boldsymbol{\Sigma}_{[\boldsymbol{a},\boldsymbol{a}]} - \boldsymbol{\Sigma}_{[\boldsymbol{a},\boldsymbol{b}]}\boldsymbol{\Sigma}_{[\boldsymbol{b},\boldsymbol{b}]}^{-1}\boldsymbol{\Sigma}_{[\boldsymbol{b},\boldsymbol{a}]}\right)^{-1}$$

symmetrically for $\boldsymbol{\Sigma}^{[\boldsymbol{b},\boldsymbol{b}]}$
and with

$$\boldsymbol{\Sigma}^{[\boldsymbol{a},\boldsymbol{b}]} = \boldsymbol{\Sigma}_{[\boldsymbol{a},\boldsymbol{a}]}^{-1}\boldsymbol{\Sigma}_{[\boldsymbol{a},\boldsymbol{b}]}\boldsymbol{\Sigma}^{[\boldsymbol{b},\boldsymbol{b}]}$$

Incorrelation and independence

Now it is easy to argue that if

$$\boldsymbol{\Sigma} = \left[\begin{array}{c|c} \boldsymbol{\Sigma}_{[a,a]} & \mathbf{0}_{(a\times b)} \\ \hline \mathbf{0}_{(b\times a)} & \boldsymbol{\Sigma}_{[b,b]} \end{array}\right]$$

then, since

$$|\boldsymbol{\Sigma}| = \left|\boldsymbol{\Sigma}_{[\boldsymbol{a},\boldsymbol{a}]}\right| \left|\boldsymbol{\Sigma}_{[\boldsymbol{b},\boldsymbol{b}]} - \boldsymbol{\Sigma}_{[\boldsymbol{b},\boldsymbol{a}]}\boldsymbol{\Sigma}^{-1}_{[\boldsymbol{a},\boldsymbol{a}]}\boldsymbol{\Sigma}_{[\boldsymbol{a},\boldsymbol{b}]}\right| = \left|\boldsymbol{\Sigma}_{[a,a]}\right| \cdot \left|\boldsymbol{\Sigma}_{[b,b]}\right|$$

and

$$\boldsymbol{\Sigma}^{-1} = \left[\begin{array}{c|c} \boldsymbol{\Sigma}^{[\boldsymbol{a},\boldsymbol{a}]} & \boldsymbol{\Sigma}^{[\boldsymbol{a},\boldsymbol{b}]} \\ \hline \boldsymbol{\Sigma}^{[\boldsymbol{b},\boldsymbol{a}]} & \boldsymbol{\Sigma}^{[\boldsymbol{b},\boldsymbol{b}]} \end{array}\right] = \left[\begin{array}{c|c} \boldsymbol{\Sigma}^{-1}_{[\boldsymbol{a},\boldsymbol{a}]} & \mathbf{0}_{(a\times b)} \\ \hline \mathbf{0}_{(b\times a)} & \boldsymbol{\Sigma}^{-1}_{[\boldsymbol{b},\boldsymbol{b}]} \end{array}\right]$$

we get the following result for the partitioned vector $\boldsymbol{y} = \begin{bmatrix} \boldsymbol{y}_{\boldsymbol{a}} \\ \boldsymbol{y}_{\boldsymbol{b}} \end{bmatrix}$

$$p_{\boldsymbol{y}}(y_1, \ldots, y_p|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = p_{\boldsymbol{y}_{[\boldsymbol{a}]}}(\boldsymbol{y}_{\boldsymbol{a}}|\boldsymbol{\mu}_{\boldsymbol{a}}, \boldsymbol{\Sigma}_{[\boldsymbol{a},\boldsymbol{a}]})p_{\boldsymbol{y}_{[\boldsymbol{b}]}}(\boldsymbol{y}_{\boldsymbol{b}}|\boldsymbol{\mu}_{\boldsymbol{b}}, \boldsymbol{\Sigma}_{[\boldsymbol{b},\boldsymbol{b}]})$$

In fact, in order to prove the last property P3

$$\boldsymbol{y}_{[\boldsymbol{b}]}|\boldsymbol{y}_{[\boldsymbol{a}]}, \boldsymbol{\mu}, \boldsymbol{\Sigma} \sim MVN_b(\boldsymbol{\mu}_{\boldsymbol{b}|\boldsymbol{a}}, \boldsymbol{\Sigma}_{\boldsymbol{b}|\boldsymbol{a}})$$

we will be mainly using algebra of partitioned matrix multiplication. Consider the following matrix $\boldsymbol{L}$ for a suitable instrumental linear transform

$$\boldsymbol{L} = \left[\begin{array}{c|c} L_{[\boldsymbol{a},\boldsymbol{a}]} & L_{[\boldsymbol{a},\boldsymbol{b}]} \\ \hline L_{[\boldsymbol{b},\boldsymbol{a}]} & L_{[\boldsymbol{b},\boldsymbol{b}]} \end{array}\right] = \left[\begin{array}{c|c} \boldsymbol{I}_{(a\times a)} & \boldsymbol{0}_{(a\times b)} \\ \hline -\boldsymbol{\Sigma}_{[\boldsymbol{b},\boldsymbol{a}]}\boldsymbol{\Sigma}^{-1}_{[\boldsymbol{a},\boldsymbol{a}]} & \boldsymbol{I}_{(b\times b)} \end{array}\right]$$

and the corresponding linear transform:

$$\begin{aligned} \boldsymbol{z} &= L\boldsymbol{y} = \left[\begin{array}{c|c} L_{[\boldsymbol{a},\boldsymbol{a}]} & L_{[\boldsymbol{a},\boldsymbol{b}]} \\ \hline L_{[\boldsymbol{b},\boldsymbol{a}]} & L_{[\boldsymbol{b},\boldsymbol{b}]} \end{array}\right] \left[\begin{array}{c} \boldsymbol{y}_{[\boldsymbol{a}]} \\ \hline \boldsymbol{y}_{[\boldsymbol{b}]} \end{array}\right] = \\ &= \left[\begin{array}{c} L_{[\boldsymbol{a},\boldsymbol{a}]}\boldsymbol{y}_{[\boldsymbol{a}]} + L_{[\boldsymbol{a},\boldsymbol{b}]}\boldsymbol{y}_{[\boldsymbol{b}]} \\ \hline L_{[\boldsymbol{b},\boldsymbol{a}]}\boldsymbol{y}_{[\boldsymbol{a}]} + L_{[\boldsymbol{b},\boldsymbol{b}]}\boldsymbol{y}_{[\boldsymbol{b}]} \end{array}\right] = \left[\begin{array}{c} \boldsymbol{y}_{[\boldsymbol{a}]} \\ \hline \boldsymbol{y}_{[\boldsymbol{b}|\boldsymbol{a}]} \end{array}\right] \end{aligned}$$

where $\boldsymbol{y}_{[\boldsymbol{b}|\boldsymbol{a}]} \equiv \boldsymbol{y}_{[\boldsymbol{b}]} - \boldsymbol{\Sigma}_{[\boldsymbol{b},\boldsymbol{a}]}\boldsymbol{\Sigma}^{-1}_{[\boldsymbol{a},\boldsymbol{a}]}\boldsymbol{y}_{[\boldsymbol{a}]}$

Since the vector $(\boldsymbol{y}_{[\boldsymbol{a}]}, \boldsymbol{y}_{[\boldsymbol{b}|\boldsymbol{a}]})$ is jointly multivariate normal, then all I need to look at are the corresponding parameters.

In particular, we should make the following partitioned matrix multiplications for the new mean vector $\tilde{\boldsymbol{\mu}}$

$$\tilde{\boldsymbol{\mu}} = L\boldsymbol{\mu} = \left[\begin{array}{c} L_{[\boldsymbol{a},\boldsymbol{a}]}\boldsymbol{\mu}_{[\boldsymbol{a}]} + L_{[\boldsymbol{a},\boldsymbol{b}]}\boldsymbol{\mu}_{[\boldsymbol{b}]} \\ \hline L_{[\boldsymbol{b},\boldsymbol{a}]}\boldsymbol{\mu}_{[\boldsymbol{a}]} + L_{[\boldsymbol{b},\boldsymbol{b}]}\boldsymbol{\mu}_{[\boldsymbol{b}]} \end{array}\right] = \left[\begin{array}{c} \boldsymbol{\mu}_{[\boldsymbol{a}]} \\ \hline \boldsymbol{\mu}_{[\boldsymbol{b}]} - \boldsymbol{\Sigma}_{[\boldsymbol{b},\boldsymbol{a}]}\boldsymbol{\Sigma}_{[\boldsymbol{a},\boldsymbol{a}]}^{-1}\boldsymbol{\mu}_{[\boldsymbol{a}]} \end{array}\right]$$

..... and .....

the new variance covariance matrix $\tilde{\boldsymbol{\Sigma}}$

$$\begin{aligned}
\tilde{\boldsymbol{\Sigma}} &= \left[\begin{array}{c|c} \tilde{\boldsymbol{\Sigma}}_{[\boldsymbol{a},\boldsymbol{a}]} & \tilde{\boldsymbol{\Sigma}}_{[\boldsymbol{a},\boldsymbol{b}]} \\ \hline \tilde{\boldsymbol{\Sigma}}_{[\boldsymbol{b},\boldsymbol{a}]} & \tilde{\boldsymbol{\Sigma}}_{[\boldsymbol{b},\boldsymbol{b}]} \end{array}\right] = L\boldsymbol{\Sigma}L^T \\
&= \left[\begin{array}{c|c} L_{[\boldsymbol{a},\boldsymbol{a}]} & L_{[\boldsymbol{a},\boldsymbol{b}]} \\ \hline L_{[\boldsymbol{b},\boldsymbol{a}]} & L_{[\boldsymbol{b},\boldsymbol{b}]} \end{array}\right] \left[\begin{array}{c|c} \boldsymbol{\Sigma}_{[\boldsymbol{a},\boldsymbol{a}]} & \boldsymbol{\Sigma}_{[\boldsymbol{a},\boldsymbol{b}]} \\ \hline \boldsymbol{\Sigma}_{[\boldsymbol{b},\boldsymbol{a}]} & \boldsymbol{\Sigma}_{[\boldsymbol{b},\boldsymbol{b}]} \end{array}\right] \left[\begin{array}{c|c} L_{[\boldsymbol{a},\boldsymbol{a}]} & L_{[\boldsymbol{a},\boldsymbol{b}]} \\ \hline L_{[\boldsymbol{b},\boldsymbol{a}]} & L_{[\boldsymbol{b},\boldsymbol{b}]} \end{array}\right]^T \\
&= \left[\begin{array}{c|c} L_{[\boldsymbol{a},\boldsymbol{a}]} & L_{[\boldsymbol{a},\boldsymbol{b}]} \\ \hline L_{[\boldsymbol{b},\boldsymbol{a}]} & L_{[\boldsymbol{b},\boldsymbol{b}]} \end{array}\right] \left[\begin{array}{c|c} \boldsymbol{\Sigma}_{[\boldsymbol{a},\boldsymbol{a}]} & \boldsymbol{\Sigma}_{[\boldsymbol{a},\boldsymbol{b}]} \\ \hline \boldsymbol{\Sigma}_{[\boldsymbol{b},\boldsymbol{a}]} & \boldsymbol{\Sigma}_{[\boldsymbol{b},\boldsymbol{b}]} \end{array}\right] \left[\begin{array}{c|c} L^T_{[\boldsymbol{a},\boldsymbol{a}]} & L^T_{[\boldsymbol{b},\boldsymbol{a}]} \\ \hline L^T_{[\boldsymbol{a},\boldsymbol{b}]} & L^T_{[\boldsymbol{b},\boldsymbol{b}]} \end{array}\right]
\end{aligned}$$

where the most interesting computations are the ones regarding $\tilde{\boldsymbol{\Sigma}}_{[\boldsymbol{a},\boldsymbol{b}]} = \tilde{\boldsymbol{\Sigma}}^T_{[\boldsymbol{b},\boldsymbol{a}]} = \mathbf{0}_{[\boldsymbol{a},\boldsymbol{b}]}$ and $\tilde{\boldsymbol{\Sigma}}_{[\boldsymbol{b},\boldsymbol{b}]} = \boldsymbol{\Sigma}_{[\boldsymbol{b},\boldsymbol{b}]} - \boldsymbol{\Sigma}_{[\boldsymbol{b},\boldsymbol{a}]}\boldsymbol{\Sigma}^{-1}_{[\boldsymbol{a},\boldsymbol{a}]}\boldsymbol{\Sigma}_{[\boldsymbol{a},\boldsymbol{b}]}$

In particular $\tilde{\boldsymbol{\Sigma}}^T_{[\boldsymbol{b},\boldsymbol{a}]} = \mathbf{0}_{[\boldsymbol{a},\boldsymbol{b}]}$ implies that the two subvectors $\boldsymbol{y}_{[\boldsymbol{b}|\boldsymbol{a}]}$ and $\boldsymbol{y}_{[\boldsymbol{a}]}$ are stochastically independent

Now we are ready to argue why

$$\boldsymbol{y}_{[\boldsymbol{b}]}|\boldsymbol{y}_{[\boldsymbol{a}]}, \boldsymbol{\mu}, \boldsymbol{\Sigma} \sim MVN_b(\boldsymbol{\mu}_{\boldsymbol{b}|\boldsymbol{a}}, \boldsymbol{\Sigma}_{\boldsymbol{b}|\boldsymbol{a}})$$

or, more transparently,

$$\boldsymbol{y}_{[\boldsymbol{b}]}|\boldsymbol{y}_{[a]} \sim MVN_b\left(\boldsymbol{\mu}_{[\boldsymbol{b}]} + \boldsymbol{\Sigma}_{[\boldsymbol{b},\boldsymbol{a}]}(\boldsymbol{\Sigma}_{[\boldsymbol{a},\boldsymbol{a}]})^{-1}\left(\boldsymbol{y}_{[\boldsymbol{a}]} - \boldsymbol{\mu}_{[\boldsymbol{a}]}\right), \boldsymbol{\Sigma}_{[\boldsymbol{bb}]} - \boldsymbol{\Sigma}_{[\boldsymbol{ba}]}\boldsymbol{\Sigma}^{-1}_{[\boldsymbol{aa}]}\boldsymbol{\Sigma}_{[\boldsymbol{ab}]}\right)$$

$$\boldsymbol{y}_{[\boldsymbol{b}|\boldsymbol{a}]} \perp\!\!\!\perp \boldsymbol{y}_{[\boldsymbol{a}]} \quad \Longrightarrow \quad \boldsymbol{y}_{[\boldsymbol{b}|\boldsymbol{a}]}|\boldsymbol{y}_{[\boldsymbol{a}]} \overset{d}{=} \boldsymbol{y}_{[\boldsymbol{b}|\boldsymbol{a}]} \sim MVN_b(\tilde{\boldsymbol{\mu}}_{[\boldsymbol{b}]}, \tilde{\boldsymbol{\Sigma}}_{[\boldsymbol{b},\boldsymbol{b}]})$$

where

$$\begin{aligned}
\tilde{\boldsymbol{\mu}}_{[\boldsymbol{b}]} &= \boldsymbol{\mu}_{[\boldsymbol{b}]} - \boldsymbol{\Sigma}_{[\boldsymbol{b},\boldsymbol{a}]}\boldsymbol{\Sigma}^{-1}_{[\boldsymbol{a},\boldsymbol{a}]}\boldsymbol{\mu}_{[\boldsymbol{a}]} \\
\tilde{\boldsymbol{\Sigma}}_{[\boldsymbol{b},\boldsymbol{b}]} &= \boldsymbol{\Sigma}_{[\boldsymbol{b},\boldsymbol{b}]} - \boldsymbol{\Sigma}_{[\boldsymbol{b},\boldsymbol{a}]}\boldsymbol{\Sigma}^{-1}_{[\boldsymbol{a},\boldsymbol{a}]}\boldsymbol{\Sigma}_{[\boldsymbol{a},\boldsymbol{b}]}
\end{aligned}$$

Moreover, conditionally on $\boldsymbol{y}_{[\boldsymbol{a}]}$

$$\begin{aligned}
\boldsymbol{y}_{[\boldsymbol{b}]}|\boldsymbol{y}_{[\boldsymbol{a}]} &\overset{d}{=} \boldsymbol{y}_{[\boldsymbol{b}]} - \boldsymbol{\Sigma}_{[\boldsymbol{b},\boldsymbol{a}]}\boldsymbol{\Sigma}^{-1}_{[\boldsymbol{a},\boldsymbol{a}]}\boldsymbol{y}_{[\boldsymbol{a}]} + \boldsymbol{\Sigma}_{[\boldsymbol{b},\boldsymbol{a}]}\boldsymbol{\Sigma}^{-1}_{[\boldsymbol{a},\boldsymbol{a}]}\boldsymbol{y}_{[\boldsymbol{a}]}|\boldsymbol{y}_{[\boldsymbol{a}]} \\
&\overset{d}{=} \boldsymbol{y}_{[\boldsymbol{b}|\boldsymbol{a}]} + \boldsymbol{\Sigma}_{[\boldsymbol{b},\boldsymbol{a}]}\boldsymbol{\Sigma}^{-1}_{[\boldsymbol{a},\boldsymbol{a}]}\boldsymbol{y}_{[\boldsymbol{a}]}\boldsymbol{y}_{[\boldsymbol{a}]}|\boldsymbol{y}_{[\boldsymbol{a}]} \\
&\sim MVN_b(\tilde{\boldsymbol{\mu}}_{[\boldsymbol{b}]} + \boldsymbol{\Sigma}_{[\boldsymbol{b},\boldsymbol{a}]}\boldsymbol{\Sigma}^{-1}_{[\boldsymbol{a},\boldsymbol{a}]}\boldsymbol{y}_{[\boldsymbol{a}]}, \tilde{\boldsymbol{\Sigma}}_{[\boldsymbol{b},\boldsymbol{b}]}) \\
&= MVN_b\left(\boldsymbol{\mu}_{[\boldsymbol{b}]} + \boldsymbol{\Sigma}_{[\boldsymbol{b},\boldsymbol{a}]}\boldsymbol{\Sigma}^{-1}_{[\boldsymbol{a},\boldsymbol{a}]}\left(\boldsymbol{y}_{[\boldsymbol{a}]} - \boldsymbol{\mu}_{[\boldsymbol{a}]}\right), \boldsymbol{\Sigma}_{[\boldsymbol{bb}]} - \boldsymbol{\Sigma}_{[\boldsymbol{ba}]}\boldsymbol{\Sigma}^{-1}_{[\boldsymbol{aa}]}\boldsymbol{\Sigma}_{[\boldsymbol{ab}]}\right) \\
&= MVN_b(\boldsymbol{\mu}_{\boldsymbol{b}|\boldsymbol{a}}, \boldsymbol{\Sigma}_{\boldsymbol{b}|\boldsymbol{a}})
\end{aligned}$$

We can then derive the regression function of $\boldsymbol{y}_{[\boldsymbol{b}]}$ with respect to $\boldsymbol{y}_{[\boldsymbol{a}]}$. In fact we can verify that

$$E\left[\boldsymbol{y}_{[\boldsymbol{b}]}|\boldsymbol{y}_{[\boldsymbol{a}]}\right] = \boldsymbol{\mu}_{[\boldsymbol{b}]} - \boldsymbol{\Sigma}_{[\boldsymbol{b},\boldsymbol{a}]}\boldsymbol{\Sigma}^{-1}_{[\boldsymbol{a},\boldsymbol{a}]}\boldsymbol{\mu}_{[\boldsymbol{a}]} + \boldsymbol{\Sigma}_{[\boldsymbol{b},\boldsymbol{a}]}\boldsymbol{\Sigma}^{-1}_{[\boldsymbol{a},\boldsymbol{a}]}\boldsymbol{y}_{[\boldsymbol{a}]}$$

For instance we may consider the particular case with $\boldsymbol{a} = \{2, 3, ..., p\}$ and $\boldsymbol{b} = \{1\}$. It will be $\beta_1 = \boldsymbol{\mu}_{[\boldsymbol{b}]} - \boldsymbol{\Sigma}_{[\boldsymbol{b},\boldsymbol{a}]}\boldsymbol{\Sigma}^{-1}_{[\boldsymbol{a},\boldsymbol{a}]}\boldsymbol{\mu}_{[\boldsymbol{a}]}$ a scalar and $(\beta_1, ..., \beta_p) = \boldsymbol{\Sigma}_{[\boldsymbol{b},\boldsymbol{a}]}\boldsymbol{\Sigma}^{-1}_{[\boldsymbol{a},\boldsymbol{a}]}\boldsymbol{y}_{[\boldsymbol{a}]}$ a $(p-1)$ row vector with

$$\boldsymbol{\beta}^T = (\beta_1, \beta_2, ..., \beta_p)$$

We can then derive the general multidimensional regression function of $\boldsymbol{y}_{[\boldsymbol{b}]}$ with respect to $\boldsymbol{y}_{[\boldsymbol{a}]}$ and verify it is linear

$$E\left[\boldsymbol{y}_{[\boldsymbol{b}]}|\boldsymbol{y}_{[\boldsymbol{a}]}\right] = \boldsymbol{c} + \boldsymbol{D}\boldsymbol{y}_{[\boldsymbol{a}]}$$

Prior distribution for a MVN model

We need to specify a joint prior distribution on the whole $(p + p \times (p+1)/2)$ dimensional parameter $(\boldsymbol{\theta}, \boldsymbol{\Sigma})$ with restricted subspace due to the symmetry and positive definiteness condition on the variance covariance matrix.

A suitable *semiconjugate* or *partially conjugate* prior distribution is the following

$$\pi(\boldsymbol{\theta}, \boldsymbol{\Sigma}) = \pi(\boldsymbol{\theta})\pi(\boldsymbol{\Sigma})$$

where

$$\boldsymbol{\theta} \sim MVN_p(\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0)$$

$$\boldsymbol{\Sigma} \sim InvWishart(\nu_0, \boldsymbol{S}_0^{-1})$$

Wishart and its relationship with MVN

Similarly to the $\chi^2_{\nu_0}$ distribution (and hence to the Gamma(shape=$\frac{\nu_0}{2}$,rate=$\frac{1}{2}$) distribution) we obtain a random matrix with Wishart distribution summing up ν_0 independent quadratic forms of independent and identically distributed multivariate random vectors $\boldsymbol{x}_i \sim MVN_p(\boldsymbol{0}, \boldsymbol{\Lambda})$. More precisely,

$$\boldsymbol{X}^T\boldsymbol{X} = \sum_{i=1}^{\nu_0} \boldsymbol{x}_i \boldsymbol{x}_i^T \sim Wishart(\nu_0, \boldsymbol{\Lambda})$$

where $\boldsymbol{x}_i^T$ (of dimension $1 \times p$) is the i-th row of the $\nu_0 \times p$ matrix $\boldsymbol{X}$

$$E[\boldsymbol{X}^T\boldsymbol{X}] = \nu_0 \boldsymbol{\Lambda}$$

Wishart distribution and singularity

Note that if we take $\nu_0 < p$ rows to build up the matrix $\boldsymbol{X}$ then the random matrix resulting from

$$\boldsymbol{W} = \boldsymbol{X}^T \boldsymbol{X}$$

turns out to be almost surely singular (no inverse can be uniquely defined). If $\nu_0 \geq p$ then the random matrix $\boldsymbol{W} = \boldsymbol{X}^T \boldsymbol{X}$ has an absolutely continuous distribution on the space of symmetric and definite positive square matrices with **probability density** given by

$$f(\boldsymbol{W}) = c \cdot |\boldsymbol{W}|^{\frac{\nu_0}{2} - \frac{p+1}{2}} exp \left\{ -\frac{1}{2} tr \left(\boldsymbol{\Lambda}^{-1} \boldsymbol{W} \right) \right\}$$

with multiplicative constant

$$c = 2^{\frac{\nu_0}{2} p} \Gamma_p \left(\frac{\nu_0}{2} \right) |\boldsymbol{\Lambda}|^{\frac{\nu_0}{2}}$$

If the random matrix $\boldsymbol{W}$ is almost surely non-singular ($\nu_0 \geq p$) then it is well defined the random matrix corresponding to its inverse

$$\boldsymbol{W}^{-1} \sim InvWishart(\nu_0, \boldsymbol{\Lambda})$$

Inverse Wishart distributions

In PH's book notation $\boldsymbol{\Sigma} \sim InvWishart(\nu_0, \boldsymbol{S}_0^{-1})$ if the random $p \times p$ matrix has the following density

$$p(\boldsymbol{\Sigma}|\nu_0, \boldsymbol{S}_0^{-1}) \quad = \quad \left[2^{\nu_0 p/2} \pi^{\binom{p}{2}/2} |\boldsymbol{S}_0|^{-\nu_0/2} \prod_{j=1}^{p} \Gamma([\nu_0 + (1-j)/2])\right]^{-1} \times$$

$$\times |\boldsymbol{\Sigma}|^{-(\nu_0+p+1)/2} exp\left\{-\frac{1}{2} tr\left(\boldsymbol{S}_0 \boldsymbol{\Sigma}^{-1}\right)\right\}$$

This somehow generalizes in $p \geq 2$ dimensions the $InvGamma\left(shape = \frac{\nu_0}{2}, origrate = \frac{s_0}{2}\right)$ or, equivalently, $InvGamma\left(shape = \frac{\nu_0}{2}, origscale = 2s_0^{-1}\right)$

Confusing parametrizations and conventions

We are used to say that

$$X \sim Gamma(shape = \alpha, rate = \beta) \quad \Longleftrightarrow \quad Y = \frac{1}{X} \sim InvGamma(\alpha, \beta)$$

Now it is easy to understand that if β represents a rate parameter for X it will be a scale parameter for Y.
So how should we name the second parameter of the Inverse Gamma?
Perhaps, we should better adopt the following convention: whenever we explicitly specify the parameter labels *rate* or *scale* they are referred to the present random variable (Inverse Gamma or Inverse Wishart) not from the originating ones.

Confusing parametrizations and conventions

Symmetrically, we can say that

$$X \sim Gamma(shape = \alpha, scale = \beta) \quad \Longleftrightarrow \quad Y = \frac{1}{X} \sim InvGamma(\alpha, \beta)$$

Now it is easy to understand that if β represents a scale parameter for X it will be a rate parameter for Y.
So how should we name the second parameter of the Inverse Gamma?
Perhaps, we should better adopt the following convention: whenever we explicitly specify the parameter labels *rate* or *scale* they are referred to the present random variable (Inverse Gamma or Inverse Wishart) not from the originating ones.

Confusing parametrizations and conventions

Similar considerations can be carried out for the Wishart and Inverse Wishart distributions.
In PH's book the second parameter $\boldsymbol{S}_0^{-1}$ for the InverseWishart distribution is related to a generalized version of the rate parameter of the inverse Gamma. In fact, up to a suitable multiplicative constant $2\boldsymbol{S}_0^{-1}$ acts as a generalized *rate* parameter and hence its inverse $\boldsymbol{S}_0/2$ acts as a generalized *scale* parameter for the corresponding InverseWishart which, in turn, corresponds to the *origrate* parameter of the originating Wishart distribution. Finally the matrix $2\boldsymbol{S}_0^{-1}$ acts as a generalized *scale* parameter for the corresponding Wishart which corresponds to the prior on the precision matrix.

If $\nu_0 > (p+1)$ and $\boldsymbol{\Sigma} \sim InvWishart(d.o.f. = \nu_0, rate = \boldsymbol{S}_0^{-1})$ then

$$E\left[\boldsymbol{\Sigma}\right] = \frac{\boldsymbol{S}_0}{\nu_0 - p - 1}$$

Inverse Wishart distributions

An alternative way of writing the density which more closely resembles the univariate case of an inverse $\chi^2_{\nu_0}$ density is the following where $\boldsymbol{S}_0^{-1}$ can be thought of as a generalized ”rate” matrix for the inverse Wishart and hence its inverse generalizes the $Gamma(shape = \nu_0/2, rate = 1)$ random variable appropriately ”scaled” with $\boldsymbol{S}_0^{-1}/2$

$$p(\boldsymbol{\Sigma}|\nu_0, \boldsymbol{S}_0^{-1}) \quad = \quad \frac{|\boldsymbol{S}_0|^{\frac{\nu_0}{2}}}{2^{p\frac{\nu_0}{2}}} \frac{|\boldsymbol{\Sigma}|^{-\frac{\nu_0+p+1}{2}}}{\boldsymbol{\Gamma}_p(\frac{\nu_0}{2})} exp\left\{-\frac{1}{2} tr\left(\boldsymbol{S}_0 \boldsymbol{\Sigma}^{-1}\right)\right\}$$

where

$$\boldsymbol{\Gamma}_p(x) = \pi^{\frac{\binom{p}{2}}{2}} \prod_{j=1}^{p} \Gamma(x + [1-j]/2) = \pi^{\frac{p(p-1)}{4}} \prod_{j=1}^{p} \Gamma(x + [1-j]/2)$$

is the multivariate (p-variate) Gamma function. It is defined for a real $x > \frac{p-1}{2}$ and we use it for $x = \frac{\nu_0}{2}$.

Inverse Wishart distributions

In fact,

$$\boldsymbol{\Sigma} \sim InvWishart(\nu_0, rate = \boldsymbol{S}_0^{-1}) \iff \boldsymbol{\Sigma}^{-1} \sim Wishart(\nu_0, scale = \boldsymbol{S}_0^{-1})$$

and, for $\nu_0 > (p+1)$,

$$E\left[\boldsymbol{\Sigma}\right] = \frac{1}{\nu_0 - p - 1}\boldsymbol{S}_0$$

while

$$E\left[\boldsymbol{\Sigma}^{-1}\right] = \nu_0 \boldsymbol{S}_0^{-1}$$

Inverse Wishart elicitation

In order to ”center” the Inverse Wishart distribution around a prior guess $\mathbf{\Sigma}^*$ we could

- center it tightly settting

$$\begin{aligned} \nu_0 & \quad \text{large} \\ \boldsymbol{S}_0 & = (\nu_0 - p - 1)\mathbf{\Sigma}^* \end{aligned}$$

- center it loosely (vaguely) setting

$$\begin{aligned} \nu_0 & = p + 2 \\ \boldsymbol{S}_0 & = \mathbf{\Sigma}^* \end{aligned}$$

In both cases we would get

$$\mathbf{\Sigma} \sim InvWishart(\nu_0, rate = \mathbf{\Sigma}^{*-1}) \iff E\left[\mathbf{\Sigma}\right] = \mathbf{\Sigma}^*$$

Partially conjugate structure

Now let us come back to our Bayesian model for conditionally multivariate normally distributed random observations.
In the previously introduced partially conjugate prior structure we will show that, after observing a sample of n random vectors $\boldsymbol{Y}_1 = \boldsymbol{y}_1, \ldots, \boldsymbol{Y}_n = \boldsymbol{y}_n$ from our multivariate Normal model, the following conditional distributions hold:

$$\boldsymbol{\theta} | \boldsymbol{\Sigma}, \boldsymbol{Y}_1 \ldots, \boldsymbol{Y}_n \sim MVN_p(\boldsymbol{\mu}_n, \boldsymbol{\Lambda}_n)$$

$$\begin{aligned} \boldsymbol{\Sigma} | \boldsymbol{\theta}, \boldsymbol{Y}_1 \ldots, \boldsymbol{Y}_n &\sim InvWishart(\nu_n, scale = \boldsymbol{S}_n) = \\ &\sim InvWishart(\nu_n, rate = \boldsymbol{S}_n^{-1}) \end{aligned}$$

where $\boldsymbol{\mu}_n$ and $\boldsymbol{\Lambda}_n$ may depend on $\boldsymbol{\Sigma}$ (and the observations) as well as ν_n and $\boldsymbol{S}_n$ may depend on $\boldsymbol{\theta}$ (and the observations).

Gibbs sampling for partially conjugate MVN model

Simulate, sequentially, for $t = 1, ..., T_{sim}$ times

$$\boldsymbol{\theta}^{(t)}|\boldsymbol{\Sigma}^{(t-1)}, \boldsymbol{Y}_1 ..., \boldsymbol{Y}_n \sim MVN_p(\boldsymbol{\mu}_n^{(t-1)}, \boldsymbol{\Lambda}_n^{(t-1)})$$

where

$$\begin{aligned}
\boldsymbol{\Lambda}_n &= \left(\Lambda_0^{-1} + n(\Sigma^{(t-1)})^{-1}\right)^{-1} \\
\boldsymbol{\mu}_n &= \boldsymbol{\Lambda}_n \left(\Lambda_0^{-1}\mu_0 + n(\Sigma^{(t-1)})^{-1}\bar{\boldsymbol{y}}_n\right)
\end{aligned}$$

$$\boldsymbol{\Sigma}|\theta^{(t)}, \boldsymbol{Y}_1 ..., \boldsymbol{Y}_n \sim InvWishart(\nu_n^{(t)}, (\boldsymbol{S}_n^{(t)})^{-1})$$

where

$$\begin{aligned}
\nu_n^{(t)} &= \nu_0 + n \\
\boldsymbol{S}_n^{(t)} &= \boldsymbol{S}_0 + \boldsymbol{S}_{\boldsymbol{\theta}^{(t)}}
\end{aligned}$$

and use the sampled vectors and matrices as if they were a sample from the posterior (i.e. use the Monte Carlo method main idea)

Approximate posterior analysis via simulation

The previous two conditional distributions are not a direct way of providing the joint distribution of the posterior for $(\boldsymbol{\theta}, \boldsymbol{\Sigma})$.

However, it can be proved that a sequential alternate simulation from these two conditional distributions will provide a sequence of simulated values which can be used as a sample from the posterior distribution.

The empirical (simulated) sample can be legitimately used as an approximation of the posterior similarly to the Monte Carlo approach. Indeed, those simulations are not i.i.d.. In fact they are a realization from a *Markov chain*. The sequential sampling algorithm is in fact known as Gibbs sampling (GS). Full details of the underlying theory of MCMC will be provided .... soon

Slightly more general formulation of the Gibbs Sampler

Under suitable regularity conditions ..... if we have a joint distribution of

$$(W_1, ..., W_k) \sim j(w_1, ..., w_k)$$

we can get an approximate multivariate sample from $j = j_{\boldsymbol{W}}$ iterating cyclically the simulation from the full conditionals

$$j(w_i | w_{(i)}) = j(w_i | w_1, ..., w_{i-1}, w_{i+1}, ..., w_k) \qquad i = 1, 2, ..., k$$

Remark: each intermediate simulation will update the conditioning argument for the following full conditionals

Thinking in abstract as Bayesians ...

... and get ready to handle missing data!
Suppose we have a joint distributions on 6 random quantities:

- A,B,C,D,E,F
- What do we know?
- How can we represent our knowledge?
- What do we need to specify?
- Suppose we have observed $C = c^{obs}$ and $F = f^{obs}$
- How can we learn from this new knowledge? the observations only?

Naive approaches

- Remove all the rows with at least one missing data $\Longrightarrow$ $n_{avail} < n \Longrightarrow$ loss of precious data info
- Replace (how?) missing data with imputed values $\Longrightarrow$ artificial way of reducing variability in the data and possibly biasing the data info $\Longrightarrow$ artificial way of reducing variability inferential outcome

Hence, the above remarks urge to find out a suitable approach for avoiding this shortcomings

Thinking in abstract as Bayesians ...

... and exploiting computational simulation-based approximation techniques when needed

$$(A, B, C, D, E, F) \sim j(a, b, c, d, e, f)$$

GS is a useful computational device to "approximately" draw a "sample" from $j(a, b, c, d, e, f)$ iterating sequentially (not nececessarely in this alphabetic order)

- $A|B = b, C = c, D = d, E = e, F = f \quad \rightarrow A = a^{new}$
- $B|A = a^{new}, C = c, D = d, E = e, F = f \quad \rightarrow B = b^{new}$
- $C|A = a^{new}, B = b^{new}, D = d, E = e, F = f \quad \rightarrow C = c^{new}$
- $D|A = a^{new}, B = b^{new}, C = c^{new}, E = e, F = f \quad \rightarrow D = d^{new}$
- $E|A = a^{new}, B = b^{new}, C = c^{new}, D = d^{new}, F = f \quad \rightarrow E = e^{new}$
- $F|A = a^{new}, B = b^{new}, C = c^{new}, D = d^{new}, E = e^{new} \quad \rightarrow F = f^{new}$

Then set $a = a^{new}$, $b = b^{new}$, $c = c^{new}$, $d = d^{new}$, $e = e^{new}$, $f = f^{new}$ for next Gibbs cycle

Thinking in abstract as Bayesians ...

Conceptually ... no problem if the only joint randomness left is on fewer letters, say (A, B, D, E).

$$\tilde{j}_{updated}(a, b, d, e) \propto j(a, b, c^{obs}, d, e, f^{obs})$$

GS remains a useful computational device to "approximately" draw a "sample" from $\tilde{j}(a, b, d, e)$ iterating sequentially (not necessarily in this alphabetic order)

- $A|B = b, C = c^{obs}, D = d, E = e, F = f^{obs} \quad \rightarrow A = a^{new}$
- $B|A = a^{new}, C = c^{obs}, D = d, E = e, F = f^{obs} \quad \rightarrow B = b^{new}$
- ~~$C|A = a^{new}, B = b^{new}, D = d, E = e, F = f^{obs}$~~ → no update since $C = c^{obs}$
- $D|A = a^{new}, B = b^{new}, C = c^{obs}, E = e, F = f^{obs} \quad \rightarrow D = d^{new}$
- $E|A = a^{new}, B = b^{new}, C = c^{obs}, D = d^{new}, F = f^{obs} \quad \rightarrow E = e^{new}$
- ~~$F|A = a^{new}, B = b^{new}, C = c^{obs}, D = d^{new}, E = e^{obs}$~~ → no update $F = f^{obs}$

Then set $a = a^{new}$, $b = b^{new}$, $d = d^{new}$, $e = e^{new}$ for next Gibbs cycle

Dealing with missing data

These are the random quantities involved

$$(\boldsymbol{O}, \boldsymbol{Y}, \omega)$$

where $\boldsymbol{Y} = \boldsymbol{y} = (y_1, \ldots, y_p)$ are the observables, $\boldsymbol{O} = \boldsymbol{o}$ is the realized "observation pattern" i.e. the subset of indexes which correspond to the p components which are actually observed, $\boldsymbol{\omega}$ are the parameters. We need to specify

$$j(\boldsymbol{o}, \boldsymbol{y}, \omega)$$

We understand $\boldsymbol{o}$ as a subset of the first p integers instead of a multivariate binary indicator ... using a slightly different (but indeed equivalent) notation w.r.t. PH book. We also denote the missing components with $\boldsymbol{m} = \{1, \ldots, p\} \setminus \boldsymbol{o}$

For a completely-at-random missingness mechanism we model

$$p(\boldsymbol{o}) f(\boldsymbol{y}|\omega) \pi(\omega) = p(\boldsymbol{o}) f(\boldsymbol{y}_{[\boldsymbol{o}]}, \boldsymbol{y}_{[\boldsymbol{m}]}|\omega) \pi(\omega)$$

Dealing with missing data

In a MCAR model there is no learning from the missingness pattern so, the only relevant effect of the actually observed components, is determining a suitable partition/decompostion of the observable components

$$(\boldsymbol{y}_{[\boldsymbol{o}]}, \boldsymbol{y}_{[\boldsymbol{m}]})$$

Likelihoods (1) - observed

It is useful to define and distinguish the *observed data* likelihood and the so-called the *completed data* likelihood.
The *observed data* likelihood is defined by the joint distribution of what has been actually observed

$$p(\boldsymbol{o}) \int_{\mathcal{Y}_{[\boldsymbol{m}]}} f(\boldsymbol{y}_{[\boldsymbol{o}]}, \boldsymbol{y}_{[\boldsymbol{m}]} | \boldsymbol{\omega}) d\boldsymbol{y}_{[\boldsymbol{m}]} = L_{\boldsymbol{y}_{[\boldsymbol{o}]}}(\boldsymbol{\omega}) = f(\boldsymbol{y}_{[\boldsymbol{o}]} | \boldsymbol{\omega})$$

although the great advantage of the Bayesian framework is to deal with the whole set of potentially observable quantities and apply the Gibbs sampling strategy in order to sample from the conditional joint distribution of all unobserved/unknown quantities.

Likelihoods (2) - complete data

In fact, within the Gibbs sampling, we use the *complete data* likelihood

$$p(\boldsymbol{o})f(\boldsymbol{y}_{[\boldsymbol{o}]}, \boldsymbol{y}_{[\boldsymbol{m}]}|\boldsymbol{\omega}) = p(\boldsymbol{o})f(\boldsymbol{y}_{[\boldsymbol{o}]}|\boldsymbol{y}_{[\boldsymbol{m}]}, \boldsymbol{\omega})f(\boldsymbol{y}_{[\boldsymbol{m}]}|\boldsymbol{\omega})$$

In the rhs we have highlighted that we could conceive the missing data $\boldsymbol{y}_{[\boldsymbol{m}]}$ <u>as if</u> it were an additional parameter of the model we are not interested in.

What is the focus of our Bayesian inference?

Having started from the joint $j(\boldsymbol{y}_{[\boldsymbol{o}]}, \boldsymbol{y}_{[\boldsymbol{m}]}, \omega)$, after observing $\boldsymbol{Y}_{[\boldsymbol{o}]} = \boldsymbol{y}_{[\boldsymbol{o}]}$ we have a final uncertainty based on the conditional joint distribution

$$j(\omega, \boldsymbol{y}_{[\boldsymbol{m}]} | \boldsymbol{y}_{[\boldsymbol{o}]}) \propto j(\boldsymbol{y}_{[\boldsymbol{o}]}, \boldsymbol{y}_{[\boldsymbol{m}]}, \omega)$$

Suppose we are only interested in inferring the parameter $\boldsymbol{\omega}$. In that case, in our posterior analysis on the parameter of interest we should focus on the "marginal" posterior of the parameter $\boldsymbol{\omega}$ only.

$$j(\omega | \boldsymbol{y}_{[\boldsymbol{o}]}) = \pi(\omega | \boldsymbol{y}_{[\boldsymbol{o}]})$$

On the other hand, if we are interested in inferring the missing data we could look at the marginal conditional distribution

$$j(\boldsymbol{y}_{[\boldsymbol{m}]} | \boldsymbol{y}_{[\boldsymbol{o}]})$$

What is the focus of our Bayesian inference?

Both these conditional distributions

$$\begin{aligned} j(\omega|\boldsymbol{y}_{[\boldsymbol{o}]}) \\ j(\boldsymbol{y}_{[\boldsymbol{m}]}|\boldsymbol{y}_{[\boldsymbol{o}]}) \end{aligned}$$

can be easily approximately ”explored” (and their functionals can be approximated) by means of sequentially (approximately) sampling from the joint conditional

$$j(\omega, \boldsymbol{y}_{[\boldsymbol{m}]}|\boldsymbol{y}_{[\boldsymbol{o}]})$$

and possibly applying some $h(\cdot)$ transformation of the simulations $(\omega, \boldsymbol{y}_{[\boldsymbol{m}]})^{(1)}, ..., (\omega, \boldsymbol{y}_{[\boldsymbol{m}]})^{(i)}, ..., (\omega, \boldsymbol{y}_{[\boldsymbol{m}]})^{(t)}$ or ignoring the simulated components which are not of interest

MVN model with missing measurements

Dealing with the complete data randomness of $\boldsymbol{\omega} = (\boldsymbol{\theta}, \boldsymbol{\Sigma})$ and $\boldsymbol{y}_{[\boldsymbol{m}]}$ makes life easier in order to "*draw an approximate sample*" from the posterior distribution using GS since the following full-conditionals

(a) $\boldsymbol{\Sigma}^{(t)} | \boldsymbol{\theta}^{(t-1)}, \boldsymbol{y}_{[\boldsymbol{m}]}^{(t-1)}, \boldsymbol{y}_{[\boldsymbol{o}]}$

(b) $\boldsymbol{\theta}^{(t)} | \boldsymbol{\Sigma}^{(t)}, \boldsymbol{y}_{[\boldsymbol{m}]}^{(t-1)}, \boldsymbol{y}_{[\boldsymbol{o}]}$

(c) $\boldsymbol{y}_{i,[\boldsymbol{m}_i]}^{(t)} | \boldsymbol{\theta}^{(t)}, \boldsymbol{\Sigma}^{(t)}, \boldsymbol{y}_{i,[\boldsymbol{o}_i]}$ for $i = 1, \ldots, n$.

can be explicitly recognized and random draws from them can then be implemented: (a) and (b) are standard as in the all-available data case, while for (c) one should only consider that, conditionally on the parameters of the MVN, for each observations i we should only look at the conditional distributions of MVN subvectors for each missing pattern of the observation i independently of the other ones (look also at the DAG!)

MVN model with missing measurements

We now focus on explaining the (c) part of the GS implementation for our MVN model with missing data. Indeed the joint distribution of the original Bayesian model can be rewritten as follows with a slight abuse of notation

$$
\begin{aligned}
& \pi(\boldsymbol{\theta})\pi(\boldsymbol{\Sigma})\prod_{j=1}^{n} f_{MVN}(\boldsymbol{y}_{j,[\boldsymbol{m}_j]},\boldsymbol{y}_{j,[\boldsymbol{o}_j]}|\boldsymbol{\theta},\boldsymbol{\Sigma})p(\boldsymbol{o}_j) = \\
= \quad & f_{MVN}(\boldsymbol{y}_{i,[\boldsymbol{m}_i]},\boldsymbol{y}_{i,[\boldsymbol{o}_i]}|\boldsymbol{\theta},\boldsymbol{\Sigma})p(\boldsymbol{o}_i) \times \pi(\boldsymbol{\theta})\pi(\boldsymbol{\Sigma})\prod_{j\neq i} f_{MVN}(\boldsymbol{y}_{j,[\boldsymbol{m}_j]},\boldsymbol{y}_{j,[\boldsymbol{o}_j]}|\boldsymbol{\theta},\boldsymbol{\Sigma})p(\boldsymbol{o}_j)
\end{aligned}
$$

Hence could focus in turn on the conditional randomness of the missing measurements of the *i*-th observation

$$
f_{MVN}(\boldsymbol{y}_{i,[\boldsymbol{m}_i]}|\boldsymbol{y}_{i,[\boldsymbol{o}_i]},\boldsymbol{\theta},\boldsymbol{\Sigma}) \propto f_{MVN}(\boldsymbol{y}_{i,[\boldsymbol{m}_i]},\boldsymbol{y}_{i,[\boldsymbol{o}_i]}|\boldsymbol{\theta},\boldsymbol{\Sigma})
$$

where we do know how to simulate from the conditional distribution from the well known properties of the MVN distribution

Remarks on implementation

The implementation is available in the code
`2025-W-10-multivariate-normal-with-missing-data-example.R`

The order of the Gibbs-Sampling steps is arbitrary. Instead of updating the three blocks of still random quantities $[\boldsymbol{\Sigma}], [\boldsymbol{\theta}], [\boldsymbol{y}_{i,[\boldsymbol{m}_i]} i = 1, \ldots, n]$ we could use another order like $[\boldsymbol{y}_{i,[\boldsymbol{m}_i]} i = 1, \ldots, n], [\boldsymbol{\Sigma}], [\boldsymbol{\theta}]$. In that case we should be careful with the notation of the (t) or $(t-1)$ and the proper implementation of the full-conditionals. In fact we should modify as follows:

(a) $\boldsymbol{y}_{i,[\boldsymbol{m}_i]}^{(t)} | \boldsymbol{\theta}^{(t-1)}, \boldsymbol{\Sigma}^{(t-1)}, \boldsymbol{y}_{i,[\boldsymbol{o}_i]}$ for $i = 1, \ldots, n$.

(b) $\boldsymbol{\Sigma}^{(t)} | \boldsymbol{\theta}^{(t-1)}, \boldsymbol{y}_{[\boldsymbol{m}]}^{(t)}, \boldsymbol{y}_{[\boldsymbol{o}]}$

(c) $\boldsymbol{\theta}^{(t)} | \boldsymbol{\Sigma}^{(t)}, \boldsymbol{y}_{[\boldsymbol{m}]}^{(t)}, \boldsymbol{y}_{[\boldsymbol{o}]}$

After one updates the missing data entries $\boldsymbol{y}_{i,[\boldsymbol{m}_i]}^{(t)}$ for $i = 1, \ldots, n$, one also needs to update summary statistics which are relevant for the updating the other components $\boldsymbol{\Sigma}$ and $\boldsymbol{\theta}$. In particular the complete data (observed and imputed missing) sample average $\bar{\boldsymbol{y}}_n$ (`y_bar_n`)

Remark on the difference between Bayesian Inference and Computational Tools

- Bayesian inference
- (Bayesian) computational tools based on a large number of simulations: MC & MCMC

In the second item the main point is how to get ......

Remark on the difference between Bayesian Inference and Computational Tools

- Bayesian inference $\longleftrightarrow$ posterior distribution + posterior predictive distribution
- (Bayesian) computational tools based on a large number of simulations: MC & MCMC

In the second item the main point is how to get an **approximation** of the posterior distribution and all its relevant summaries and derived distributions

Monte Carlo Method

It relies on the integral quantity expressed as expectation

$$I = \int_{\Theta} g(\theta) d\theta = \int_{\Theta} h(\theta)\pi(\theta) d\theta = E_{\pi}[h(\theta)]$$

and on limit theorems which ensure that appropriate functions of sequences of i.i.d. random variables (i.i.d stochastic process)

If $|I| < \infty$ and $\theta_1, ..., \theta_t$ i.i.d. $\sim \pi$ then from **Strong Law of Large Numbers** (SLLN) the empirical average can be regarded as a *consistent* "estimator" of I or, more precisely, a *consistent* sequence of "estimators" of I

$$\hat{I}_t = \frac{1}{t} \sum_{i=1}^{t} h(\theta_i) \xrightarrow{a.s.} E_{\pi}[h(\theta)] = I \qquad t \to \infty$$

You are probably more familiar with the following version of the SLLN.

SLLN (basic) If $Y_1, ..., Y_n, ...$ are i.i.d. with $Y_i \sim f_Y(y)$ and $\mu = E[Y_i]$ then

$$\bar{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i \xrightarrow{a.s.} \mu = E[Y] = \int y \, f_Y(y) \, dy$$

provided the existence and finiteness of $E_{f_Y}[Y_i]$

Indeed we can also restate the result in a slightly more general form dealing with expected values of i.i.d. random variables which are function $h(\cdot)$ of another sequence of underlying random variables X_i.

SLLN If $X_1, ..., X_n$ are i.i.d. with $X_i \sim f_X(x)$ and $I = E[h(X_i)]$ then

$$\frac{1}{n}\sum_{i=1}^{n} h(X_i) \xrightarrow{a.s.} E[h(X)] = I$$

provided the existence and finiteness of $E[h(X_i)]$

In fact the two versions of SLLN are equivalent since

- if $(X_1, ..., X_n, ...)$ are i.i.d. then also $(Y_1 = h(X_1), ..., Y_n = h(X_n), ...)$ are i.i.d.
- if $E[h(X_i)]$ exists and is finite it can be written as $E[Y_i]$ where

$$\mu = \int y\, f_Y(y)\, dy = E[Y_i] = E[h(X_i)] \overset{(*)}{=} \int h(x)\, f_X(x)\, dx = I$$

(*) a property of the Expectation of a function of a random variable

1. the use of a random value $\hat{I}$ as a way of getting an unknown quantity → typical (frequentist) inferential attitude!!
2. we must ensure a proper **error control** ...
3. when interested to posterior summaries we would rather use two separate approximations for numerator and denominator which are both integral quantities with respect to the prior but I could avoid the presence of two sources of errors if I can directly simulate from the posterior ...
4. Fundamental ingredient will be the ability of simulating random draws from some *target* distribution $\longrightarrow \pi(\cdot)$ ... possibly known up to a proportionality constant

Error evaluation

$$\begin{aligned} E[(\hat{I} - I)^2] = Var[\hat{I}] &= \frac{1}{n} Var[h(X)] = \\ &= \frac{1}{n} \left\{ E_\pi[h(X)^2] - E_\pi[h(X)]^2 \right\} = \\ &= \frac{K}{n} \end{aligned}$$

Indeed I do not know K but I can approximate/estimate K as well, as follows

$$\hat{K} = \widehat{Var}[h(X)] = \frac{1}{n} \sum_{i=1}^{n} h(X_i)^2 - \hat{I}_n^2$$

I can do better when there are sufficient conditions to apply Central Limit Theorem (CLT) and hence I can get a confidence interval for I (probabilistic approximation)

Error evaluation [1]

Unbiasedness of $\hat{I}_n$ is not enough. We are interested in knowing how much $\hat{I}_n$ is close (in distribution) to our target I.
First immediate idea is the use of MSE (Mean Square Error) which in this case coincides with $\hat{I}_n$

$$E[(\hat{I}_n - I)^2] \quad = \quad Var[\hat{I}_n] = \frac{1}{n}\{\mathbb{V}[h(X)]\} = \frac{1}{n}\left\{\mathbb{E}[h(X)^2] - \mathbb{E}[h(X)]^2\right\}$$

I can estimate $K = \mathbb{V}[h(X)]$ with analogous MC approach using the empirical variance of $h(X_i)$

$$\hat{K} = \frac{1}{n}\sum_{i=1}^{t} h(X_i)^2 - \hat{I}_n{}^2$$

(Consistent? unbiased?)

Error evaluation [2]

I can go even further when CLT can be applied since I can get asymptotic confidence intervals for I using the asymptotic distribution of the empirical average $\hat{I}_n$

$$\left[\hat{I}_n - 1.96 \cdot \sqrt{\frac{\hat{K}}{n}} \;,\; \hat{I}_n + 1.96 \cdot \sqrt{\frac{\hat{K}}{n}}\right]$$

or just the thumb rule $\pm 2 \times s.e.$

$$\left[\hat{I}_n - \sqrt{\frac{\hat{K}}{n}} \;,\; \hat{I}_n + \sqrt{\frac{\hat{K}}{n}}\right]$$

1. the use of a random value $\hat{I}$ as a way of getting an unknown quantity → typical inferential attitude!!
2. we must ensure a proper **error control** ...
3. when interested to posterior summaries we would rather use two separate approximations for numerator and denominator which are both integral quantities with respect to the prior but I could avoid the presence of two sources of errors if I can directly simulate from the posterior ...
4. Fundamental ingredient will be the ability of simulating random draws from some *target* distribution $\longrightarrow \pi(\cdot)$ ... possibly known up to a proportionality constant

Main messages

1. If we cannot compute *exactly* a finite quantity like

$$I = E_\pi[h(\theta)] = \int_\Theta h(\theta)\pi(d\theta)$$

   we can **approximate** it with $\hat{I}_n$ using a (possibly long) stream of n **simulations**, $\theta_1, ..., \theta_i, ..., \theta_t$

2. instead of integrating w.r.t $\pi(d\theta)$ we integrate with respect to the empirical distribution of the simulations

$$\hat{I}_t = H[F_t] = \int_\Theta h(\theta)F_t(d\theta) = \frac{1}{t}\sum_{i=1}^{t} h(\theta_i)$$

3. $\hat{I}_t$ is random!! (Monte Carlo)
4. this method works as long as $\hat{I}_t$ represents consistent estimator of I. In the case of i.i.d. simulations this is guaranteed by the SLLN
5. we should take under control the approximation error in a probabilistic sense: MSE or confidence intervals

Some additional messages

1. we can extend this idea and method for approximation as long as there is a sequence of "estimators" (not necessarily sample averages) which can be proved to be a consistent sequence of estimators for a functional of a probability distribution (variance, quantiles, etc). In the univariate case, if we denote with $F_t(\cdot)$ the "empirical distribution" of the i.i.d simulations θ_i from π

   $$F_t(u) = \frac{1}{t}\sum_{i=1}^{t} I_{(-\infty,u)}(\theta_i)$$

   we can write the Monte Carlo principle as follows

   $$H[F_t] \approx H[F_\pi] \text{ or } H[\pi]$$

2. when θ is multivariate we can also consider transformations h involving different components (separately or jointly)

Monte Carlo as approximation strategy .... (extended)

More generally, for unknown quantities H of interest which, similarly to expectations) are functionals of some fixed known distribution $H = H[\pi]$, one can use consistent (i.e. a.s. converging) estimators based on i.i.d. simulations in order to get an approximation for the corresponding a.s. limit H. For instance approximation of the variance can be obtained from empirical variances. When π is multidimensional the same can be done. For instance for a bivariate random vector one can approximate the correlation using the empirical correlation computed after n i.i.d. simulations form the bivariate random vector.

What can we approximate through MC?

- integrals as conceived ad expectations with respect to a suitable distribution F
- probability of events
- areas, volumes, hyper-volumes of bounded regions in $\mathbb{R}^d$
- a whole distribution → histograms
- other related quantities → quantiles
- … and much more … (maximization)

Introduction with historical note

$$I = \int_A g(x)dx \qquad\qquad A \subset \mathbb{R}^d$$

- Exact evaluation $\longleftrightarrow$ Approximation
- Approximation as a random guess !?
- Metropolis & Ulam (1949) - JASA

Alternative for exact evaluation of integrals

- methods based on direct evaluation of the integrand functions over a finite grid of points or, more generally, on an approximation of the whole integrand function and with known integral evaluation
    - Riemann sums
    - trapezoidal rule (linear interpolation)
    - Simpson's rule (quadratic interpolation)
    - other more sophisticated approximations (splines etc.)
    - methods based on orthogonal polynomials
    - quadrature formulas (optimal choice of grid points)
    - Laplace method

Riemann sums

$$I = \int_a^b g(x)dx = \lim_{n\to\infty} \sum_{i=0}^{n} g(x_{i,n}) \left(x_{i+1,n} - x_{i,n}\right)$$

$a = x_{0,n} < x_{1,n} < x_{2,n} < \ldots < x_{n,n} < x_{n+1,n} = b$.
It can be generalized to multidimensional (d-dimensional) integrals over subsets

$$
\begin{aligned}
\tilde{I}_{RS} &= \sum_{i=0}^{n-1} g(x_{i,n})\left(x_{i+1,n} - x_{i,n}\right) \\
\tilde{I}_{TR} &= \frac{1}{2}\sum_{i=0}^{n}\left[g(x_{i+1,n}) + g(x_{i,n})\right]\left(x_{i+1,n} - x_{i,n}\right) \\
\tilde{I}_{SIMPS} &= \frac{\delta}{3}\left\{g(a) + 4\sum_{i=1}^{n/2} g(x_{2i-1}) + 2\sum_{i=1}^{n/2-1} g(x_{2i}) + g(b)\right\}
\end{aligned}
$$

with n even, $a = x_0$, $b = x_n$ and $\delta = \frac{b-a}{n} = (x_{j+1} - x_j)$

Polynomial approximation of a function

Lagrange Polynomial
Explicit formula for a $p-1$-order polynomial which passes through p points on the graph of an integral function $f(x)$ on the interval $[x_1, x_p]$
Let the points be $(x_j, f(x_j))$ $(j = 1, ..., p)$

$$P_{p-1}(x) = \sum_{j=1}^{p} f(x_j) \prod_{i \neq j}^{p} \frac{x - x_i}{x_j - x_i}$$

Take $p = 3$ and obtain a second order polynomial from which you can derive Simpson's rule deriving explicitly the evaluation of the integrals within equally spaced subintervals $(y_h, y_h + \Delta)$ where $y_h = x_1$ e $y_h + \Delta = x_p = x_3$.

Polynomial approximation of a function and its exact integral

Lagrange Polynomial

Let us try to compute it for $n = 2$: a quadratic function which passes through $(a, f(a))$, $(m, f(m))$, $(b, f(b))$ with $m = \frac{a+b}{2}$

Quadrature formulas

Quadrature formulas uses similar formulas

$$I \approx \sum_{i=1}^{n} w_i f(x_i)$$

but relying on different tools for approximating integrand functions. The underlying theory is based on the construction of a system of orthogonal polynomials whose roots will allow to make an exact evaluation of integral of polynomials of degree at most $2n-1$.

Laplace's Method

Let us give some detail on first order approximation and suppose the integrand $g(x)$ is positive

$$\int_a^b g(x)dx = \int_a^b e^{n\cdot\frac{1}{n}\log g(x)} = \int_a^b e^{n\cdot L(x)}dx$$

Taking the argmax point x_0 for $g(x)$ (hence also for $L(x) = n^{-1}\log g(x)$) we have $L'(x_0) = 0$ e $L''(x_0) < 0$ and $L(x) = L(x_0) + L'(x_0)(x - x_0) + \frac{L''(x_0)}{2!}(x - x_0)^2 + R_n(x)$ so that we can rely on a Gaussian cdf representation as follows

$$\int_a^b g(x)dx \quad \approx \quad e^{n\cdot L(x_0)}\sqrt{\frac{2\pi}{-nL''(x_0)}}\left[F_{N(x_0,\sigma_0^2)}(b) - F_{N(x_0,\sigma_0^2)}(a)\right]$$

with $\sigma_0^2 = -\frac{1}{nL''(x_0)}$.

Laplace's Method (cont.)

In fact,

$$\begin{aligned}\int_a^b g(x)dx &= \int_a^b e^{n\cdot L(x)} \approx \int_a^b e^{n\cdot\left[L(x_0)+L'(x_0)(x-x_0)+\frac{L''(x_0)}{2!}(x-x_0)^2\right]}dx \\ &\approx e^{n\cdot L(x_0)}\sqrt{\frac{2\pi}{-nL''(x_0))}}\int_a^b\sqrt{\frac{-nL''(x_0))}{2\pi}}e^{-\frac{1}{2}(x-x_0)^2(-nL''(x_0))}dx \\ &\approx e^{n\cdot L(x_0)}\sqrt{2\pi\sigma_0^2}\left[F_{N(x_0,\sigma_0^2)}(b) - F_{N(x_0,\sigma_0^2)}(a)\right]\end{aligned}$$

with $\sigma_0^2 = -\frac{1}{nL''(x_0)}$.

Pros e cons

Of course there is no overall superiority

- MC seems unaffected by the curse of dimensionality
- MC exploits the probabilistic nature of the model and gives less importance to low probability regions
- we can generate many variants of the starting problem without imposing to remaster the approximation strategy
- we can achieve simultaneous and joint analysis of different synthetic features of a target distribution of interest e.g in Bayesian inference
- we (statisticians/data scientists) are well trained to deal with random variables and their ability to yield estimation of unknown quantities (approximation)

... but random ... need not be optimal!

Basic steps for vanilla Monte Carlo

A) rewrite the integral as an expectation with respect to a probability distribution F. For instance if X is an a.c. random variable with cdf F and density $f(x)$ strictly positive on $\mathcal{X} \supset A$

$$I = \int_A g(x)dx = \int_{\mathcal{X}} I_A(x)\frac{g(x)}{f(x)}f(x)dx = \int_{\mathcal{X}} h(x)f(x)dx = \mathbb{E}\left[h(X)\right]$$

B) device to generate the realization of a sequence of i.i.d. random variable with distribution F

C) use the powerful tools of asymptotic theory of i.i.d. processes + the related asymptotic inference like Strong Law of Large Numbers (SSLN - Kolmogorov) and $\hat{I}$ as a strongly consistent estimator of I

Vanilla Monte Carlo and more

$$\hat{I}_n = I(X_1, \ldots, X_n) = \frac{1}{n} \sum_{i=1}^{n} h(X_i) \qquad \xrightarrow{n \to \infty} \qquad I = \mathbb{E}[h(X_i)] = \mathbb{E}[Y_i]$$

... we could use the same asymptotic results as Kolmogorov SSLN for i.i.d. process $Y_i = h(X_i)$ $i = 1, 2, \ldots$.

using a more general structure of stochastic processes with non i.d or dependent components

- SSLN - CLT (Lindberg-Levy, Liapunov, Lindberg-Feller ...)
- Ergodic theorems for Markov Chains

Monte Carlo as approximation strategy ....

Our goal is to evaluate I

$$I = \int_A g(x)dx = \int_{\mathcal{X}} h(x) f_X(x) dx \qquad A \subset \mathcal{X} \subset \mathbb{R}^d$$

for simplicity let us suppose that $\mathcal{X}$ coincides with the support of X.
The Monte Carlo metod of approximation is based on

$$\hat{I}_n = \frac{1}{n}\sum_{i=1}^{n} h(X_i)$$

where $X_1, ..., X_n$ is simulated by and i.i.d. process with $X_i \sim f_X$
We can easily derive that $\hat{I}_n$ is unbiased, i.e.

$$\mathbb{E}[\hat{I}_n] = I$$

Is $\hat{I}_n$ close to our target I ?

Error evaluation

$$\begin{aligned} E[(\hat{I}-I)^2] = Var[\hat{I}] &= \frac{1}{n} Var[h(X)] = \\ &= \frac{1}{n}\left\{E_\pi[h(X)^2] - E_\pi[h(X)]^2\right\} = \\ &= \frac{K}{n} \end{aligned}$$

Indeed I do not know K but I can approximate/estimate K as well, as follows

$$\hat{K} = \widehat{Var}[h(X)] = \frac{1}{n}\sum_{i=1}^{n} h(X_i)^2 - \hat{I}_n^2$$

I can do better when there are sufficient conditions to apply Central Limit Theorem (CLT) and hence I can get a confidence interval for I (probabilistic approximation)

Error evaluation [1]

Unbiasedness of $\hat{I}_n$ is not enough. We are interested in knowing how much $\hat{I}_n$ is close (in distribution) to our target I.
First immediate idea is the use of MSE (Mean Square Error) which in this case coincides with $\hat{I}_n$

$$E[(\hat{I}_n - I)^2] \quad = \quad Var[\hat{I}_n] = \frac{1}{n}\{\mathbb{V}[h(X)]\} = \frac{1}{n}\left\{\mathbb{E}[h(X)^2] - \mathbb{E}[h(X)]^2\right\}$$

I can estimate $K = \mathbb{V}[h(X)]$ with analogous MC approach using the empirical variance of $h(X_i)$

$$\hat{K} = \frac{1}{n}\sum_{i=1}^{t} h(X_i)^2 - \hat{I}_n{}^2$$

(Consistent? unbiased?)

Error evaluation [2]

I can go even further when CLT can be applied since I can get asymptotic confidence intervals for I using the asymptotic distribution of the empirical average $\hat{I}_n$

$$\left[\hat{I}_n - 1.96 \cdot \sqrt{\frac{\hat{K}}{n}} \;,\; \hat{I}_n + 1.96 \cdot \sqrt{\frac{\hat{K}}{n}}\right]$$

or just the thumb rule $\pm 2 \times s.e.$

$$\left[\hat{I}_n - 2\sqrt{\frac{\hat{K}}{n}} \;,\; \hat{I}_n + 2\sqrt{\frac{\hat{K}}{n}}\right]$$

How can we simulate $X_1 \dots X_n$ i.i.d. according to F ?

- Monte Carlo roulette o game of lotto
- Physical devices like frequency pulses of radioactive emissions or of an electronic signal (RAND project, anni '50, random.org) [`library(random)`]
- Pseudo random sequences
- Uniform [discrete/continuous/univariate/multivariate] → OK. But what about the i.i.d. simulation from a generic F?

1. Discrete Uniform
2. "Continuous" Uniform
3. Integral Transform (and its generalization)
4. *ad hoc* specific transforms ("*all random variables are relative to each other*"!)
   - 4.1 simple: exponential, chi-square, Gamma, location-scale family
   - 4.2 ~~Exponential-Poisson transform~~
   - 4.3 ~~Multiple transform: Box-Muller (ratio of uniforms)~~
5. Completions, Mixtures and Marginalization.
6. ~~Fundamental Theorem of Simulation~~
7. Rejection Sampling
8. Truncation
9. ~~Ratio of uniforms~~
10. ....

Uniform pseudo random number generation

- linear congruential generators for integer numbers with uniform distribution $m_0, m_1, \dots, m_n, m_{n+1}, \dots.$ with $m_n \in \{0, \dots, M\}$

$$m_{n+1} = (a \cdot m_n + c) \bmod (M + 1)$$

- easily transformed as random deviates in the unit interval $[0, 1]$
- how can we say they are random'? → randomness tests (`library(RDieHarder)`)
- seed, periodicity and maximal useful length of the sequence (2^{38} with $a = 5^{17}$; $M = 2^{40}$; $m_0 = 1$)
- variations (e.g. permutation, coupling , ...) and alternative methods
  (up to periods like $2^{19937} - 1$)

Non randomness, original sin and reproducibility

Shift registers

Binary representation of an integer $x \in \{0, 1, ..., 2^k - 1\}$

$$x = \sum_{i=0}^{k-1} e_i 2^i$$

where $e_i \in \{0, 1\}$ is a binary digit

To generate $x_{n+1} = D(x_n)$ one can use a matrix multiplication (mod 2) on the binary digits denoting this multiplication as follows

$$x_{n+1} = Tx_n$$

where T a square matrix with binary entries

Example:

$$T_L = \begin{pmatrix} 1 & 1 & \dots & & \\ 0 & 1 & \dots & & 0 \\ & & & & \\ & & \dots & 1 & 1 \\ & & \dots & 0 & 1 \end{pmatrix} = (I + L)$$

Shift registers

If we only have 1 digits in positions i and j in the i-th row of T the in the i-th position of the transformed binary vector there will be $(e_{in} + e_{jn}) mod 2$

Kiss generator

It uses three sequences I_n, J_n and K_n and re-combines them to get a single pseudo-random sequence X_{n+1} as follows

$$\begin{aligned} I_{n+1} &= 69069 \times I_n + 23606797 mod(2^{32}) \\ J_{n+1} &= (I + L^{15})(I + R^{17}) J_n mod(2^{32}) \\ K_{n+1} &= (I + L^{13})(I + R^{18}) K_n mod(2^{31}) \end{aligned}$$

and finally

$$X_{n+1} = (I_{n+1} + J_{n+1} + K_{n+1}) mod(2^{32})$$

Random variable simulation with given non-uniform target π

Pseudo-random number generation:

Starting point: simulation of a pseudo-random sequence of draws which behave $\underline{\text{similarly}}$ to a i.i.d. process with uniform components in the unit interval $(0, 1)$

Every statistical/mathematical software (R, SAS, Mathematica, Matlab, Octave (even Excel/OpenOffice!) contains basic functions which return pseudo-random deviates from the main classes (discrete and continuous) of probability distributions. Usually the extensions from uniform in $(0, 1)$ to arbitrary distributions are based on probability calculus of random variable transformation.

1. integral trasform
2. ad hoc specific transforms → examples → efficiency
3. Rejection Sampling

Integral Transform Method

We should be looking for general rules which could be of help when we will deal with distribution whose functional form cannot be recognized to belong to known/standard parametric classes of distributions

$$X \sim F_X(x) = Pr\{X \leq x\}$$

If $U_i \sim U(0,1) \Longrightarrow X_i = F^{-1}(U_i) \sim F_X$
To simplify the proof let us assume that F is strictly increasing and continuous over all its support (although there exist a general proof without this simplifying assumption)

$$\begin{aligned} Pr\{X_i \leq x\} &= Pr\{F_X^{-1}(U_i) \leq x\} \\ &= Pr\{F(F_X^{-1}(U_i)) \leq F(x)\} = Pr\{U_i \leq F_X(x)\} = F_X(x) \end{aligned}$$

Integral Transform Method

From pseudo random deviates which are Uniform on $(0,1)$ to (standard i.e. with scale parameter $\lambda = 1$) exponentially distributed random deviates
$\Longrightarrow F_X(x) = 1 - e^{-x} \Longrightarrow F_X^{-1}(y) = -\log(1-y)$
Hence from U_i i.i.d. Uniform(0,1)

$$X_i = -\log(1-U_i) \sim exp(\lambda = 1)$$

..... although we do not actually need to do that in practice with standard statistical software ....

Discrete distributions

- in this case what do we mean with $F_X^{-1}(y)$? ($\Longrightarrow$ generalized cdf inverse)
- Bernoulli
- Generic discrete distribution with finite support

$$\{(x_i, p_i); i = 1, 2, 3, \ldots, k\}$$

- Generic discrete distribution with infinite support

$$\{(x_i, p_i); i = 1, 2, 3, \ldots\}$$

- *Can we simulate from a distribution with infinite support in a finite time if we know exactly the probability distribution?*
- *Computing time?*
- *What about if we had the distribution only up to a multiplicative constant?*

Generalized CDF Inverse (1)

We can generalize the Integral Transform Method for target discrete distributions for which F_X is neither continuous nor invertible
We can define a generalized inverse cdf as follows:

$$F_X^{*-1}(u) = inf\left\{x : F_X(x) > u\right\}$$

Example If $X \sim Ber(p)$ then its cdf is

$$F_X(x) = \begin{cases} 0 & -\infty < x < 0 \\ 1 - p & 0 \le x < 1 \\ 1 & x \ge 1 \end{cases}$$

Generalized CDF Inverse (2)

Its generalized inverse is ${F_X^*}^{-1} : (0,1) \to \mathbb{R}$

$$\begin{aligned} {F_X^*}^{-1}(u) &= \\ \begin{cases} 0 & 0 < u < 1-p \\ 1 & 1-p \leq u < 1 \end{cases} &= I_{[1-p,1)}(u) \end{aligned}$$

Remark

It is easy to verify that if $U \sim \mathit{Unif}(0,1)$ then $X = I_{[1-p,1)}(U) \sim \mathit{Ber}(p)$. Moreover,

$$X \equiv I_{[1-p,1)}(U) \overset{d}{=} X' \equiv I_{(0,p]}(U')$$

Hence in order to simulate a Bernoulli r.v. with success probability one can also use a uniform r.v. U' on $(0,1)$ and have a success when $U' \leq p$

Box-Muller Transform

Here is one of the remarkable methods to simulate from a Gaussian distribution. suppose U_1 and U_2 are two independent random variable with Uniform distribution in $[0, 1]$. Then the following random variables

$$\begin{cases} Z_1 = \sqrt{-2 \log U_1} \cos(2\pi U_2) \\ Z_2 = \sqrt{-2 \log U_1} \sin(2\pi U_2) \end{cases}$$

are independent and they both have a standard normal distribution The proof follows from well-known rules of random vector transformations

Completion

Definition Let us consider a target distribution $f(y)$ for a random variable $Y \sim f(y)$. Let us consider a bivariate random vector (Z, Y) taking values in $\mathcal{Z} \times \mathcal{Y} \subseteq \mathbb{R}^2$ such that $(Z, Y) \sim j(z, y)$. We say that the joint distribution $j(z, y)$ is a *completion* of the distribution $f(y)$ (probability density or probability mass function on $\mathcal{Y}$) if the following holds:

$$(Z, Y) \sim j(z, y) \Longrightarrow Y \sim f(y)$$

If we have probability densities this amounts to verify that

$$f(y) = \int_{\mathcal{Z}} j(z, y) dz$$

If we have probability mass functions

$$f(y) = \sum_{z \in \mathcal{Z}} j(z, y)$$

Completion: example 1

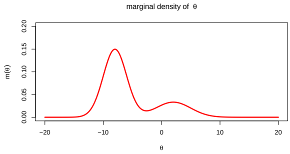
The following joint distribution $j(z, y)$

$$j(z,y) = 0.75(1-z)\frac{1}{\sqrt{2\pi 4}} exp\left\{-0.5\frac{(y+8)^2}{4}\right\} + 0.25z\frac{1}{\sqrt{2\pi 9}} exp\left\{-0.5\frac{(y-2)^2}{9}\right\}$$

with $(z, y) \in \mathcal{Z} \times \mathcal{Y} = \{0,1\} \times (-\infty, \infty)$ i.e with $z \in \{0,1\}$ and $y \in (-\infty, \infty)$, is the **completion** of a two-component finite mixture of normals $f(y) = w_0 f_{N(\mu_0,\sigma_0^2)}(y) + w_1 f_{N(\mu_1,\sigma_1^2)}(y)$ with $w_0 + w_1 = 1$:

$$f(y) = 0.75\frac{1}{\sqrt{2\pi 4}} exp\left\{-0.5\frac{(y+8)^2}{4}\right\} + 0.25\frac{1}{\sqrt{2\pi 9}} exp\left\{-0.5\frac{(y-2)^2}{9}\right\}$$

whose density is as follows



Completion: example 1 (cont.)

The easiest way to simulate from a two-component finite mixture of normals $f(y) = w_1 f_{N(\mu_1,\sigma_1^2)}(y) + w_0 f_{N(\mu_0,\sigma_0^2)}(y)$ with $w_0 + w_1 = 1$: it to draw a sample from

$$Z \sim p(z) = Prob\{Z = z\} = \begin{cases} w_1 & z = 1 \\ w_0 = 1 - = w_1 & z = 0 \end{cases}$$

and then

$$Y|Z = z \sim N(\mu_z, \sigma_z^2)$$

so that

$$(Z, Y) \sim j(z, y)$$

and hence

$$Y \sim f(y)$$

Completion: example 2

Suppose that

$$\begin{cases} Z & \sim Gamma\left(rate = \frac{g}{2}, shape = \frac{g}{2}\right) \\ Y|Z = z & \sim N\left(0, \frac{1}{z}\right) \end{cases}$$

so that

$$g_{X,Y}(z,y) = g_{Y|Z}(y|z)g_Z(z)$$

One can prove that this is a completion of a Student T distribution with g degrees of freedom

$$f_{T_g}(y) = \frac{1}{\sqrt{\pi g}} \frac{\Gamma\left(\frac{g+1}{2}\right)}{\Gamma\left(\frac{g}{2}\right)} \left(1 + \frac{y^2}{g}\right)^{-\frac{g+1}{2}} = \int g_{Z,Y}(z,y)dz$$

Completion: example 2

Suppose that $Y \sim Gamma\left(rate = \frac{g}{2}, shape = \frac{g}{2}\right)$
$X|Y = y \sim N\left(0, \frac{1}{y}\right)$ so that

$$g_{X,Y}(x,y) = g_{X|Y}(x|y)g_Y(y)$$

One can prove that this is a completion of a Student T distribution with g degrees of freedom

$$f_{T_g}(x) = \frac{1}{\sqrt{\pi g}}\frac{\Gamma\left(\frac{g+1}{2}\right)}{\Gamma\left(\frac{g}{2}\right)}\left(1+\frac{x^2}{g}\right)^{-\frac{g+1}{2}} = \int g_{X,Y}(x,y)dy$$

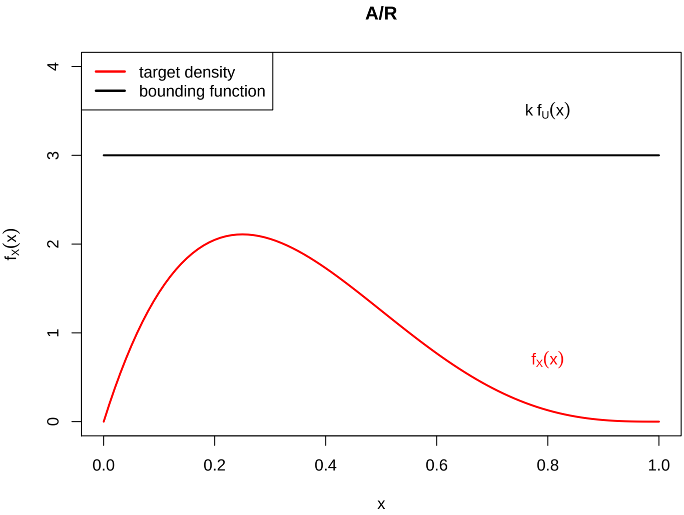
Acceptance-rejection method

Let us introduce it with a very basic intuition: we would like to simulate $X_i \sim f_X$, where f_X satisfies this boundedness condition (see figure):

$$f_X(x) \leq k = kf_U(x)$$

Basic intuitive idea: simulate $Y_1, \ldots, Y_i, \ldots$ (*candidates*) from f_U (say a uniform), but *keep* just a subset of these simulated values according to a random (acceptance-rejection) rule which tells you when you should accept it or not according the intuition that you should penalize and reject more often those values which have low density f_X (when compared to the candidate f_U).

$$\begin{array}{llcccccc}
\text{from } U(0,1) & \rightarrow & Y_1 & Y_2 & Y_3 & \ldots & Y_i & \ldots \\
\text{A/R rule} & \rightarrow & A(Y_1) & R(Y_2) & A(Y_3) & \ldots & R(Y_i) & \ldots \\
 & & & & & & & \\
\text{A/R rule} & \rightarrow & Y_1^A = Y_1 & Y_2^R = Y_2 & Y_3^A = Y_3 & \ldots & Y_i^R = Y_i & \ldots \\
 & & & & & & & \\
\text{from } f_X & \leftarrow & X_1 = Y_1^A & X_2 = Y_3^A & X_3 = Y_7^A & \ldots & X_i = Y_j^A & \ldots
\end{array}$$


A/R
target density
bounding function
$k\,f_U(x)$
$f_X(x)$
4
3
2
1
0
0.0
0.2
0.4
0.6
0.8
1.0
x
$f_X(x)$

In other words the A-R idea is: take pseudo-random $Y_1, ..., Y_N$ from a *candidate* distribution which is different from your target ad keep just of final X_i's only a subset of the simulated Y_j using a random rule based on an auxiliary experiment (say E_i)

Y_1	Y_2	Y_3	Y_4	Y_5	..	..	Y_{100}
$E_1 \to Acc$	$E_2 \to Rif$	$E_3 \to Rif$	$E_4 \to Acc$	$E_5 \to Rif$	..	..	$E_{100} \to Acc$
Y_1^A	Y_2^R	Y_3^R	Y_4^A	Y_5^R	..	..	Y_{100}^A
X_1	$--$	$--$	X_2	$--$	..	..	X_{45}

$$X \equiv Y^A$$

Acceptance-Rejection rule (particular case)

For each $Y \sim f_U$ we use an auxiliary experiment

$$Y \to \begin{cases} Y^A & \text{if } E = A \\ Y^R & \text{if } E = R \end{cases}$$

- the subset of the accepted Y_i^A is a subset of independent random deviates
- each Y_i^A has the desired *target* distribution $f_X(x)$
- The number of accepted Y_i^A, say T, is random and it less than or equal to the number of draws, say t, from the original distribution f_U. The final random size T of random draws form the target depends on acceptance probability

Acceptance-Rejection rule (particular case)

What is the distribution of the auxiliary experiment? It is basically a Bernoulli random variable such that $E = A = 1$ and $E = R = 0$. Its distribution depends on the simulated $Y = y$ in the following sense:

$$Pr(E = 1|Y = y) = \frac{f_X(y)}{kf_U(y)} = \frac{f_X(y)}{k} \in [0, 1]$$

Hence we should think the whole algorithm in terms of the joint vector (Y, E) with the following joint distribution

$$\begin{aligned} Y &\sim f_U(y) \\ E|Y = y &\sim Ber\left(p(y) = \frac{f_X(y)}{kf_U(y)}\right) \end{aligned}$$

Acceptance-Rejection rule (particular case)

The distribution of the accepted randow draws is in fact ... a conditional distribution

$$F_X(x) = Pr\left\{X \leq x\right\} = Pr\left\{Y \leq x | Y = Y^A\right\} = Pr\left\{Y \leq x | E = 1\right\}$$

Let us evaluate the acceptance probability i.e. the probability of the event $Y = Y^A$ i.e. $E = 1$. Remind that it depends on the joint distribution of (Y, E). Let us start from the conditional acceptance probability $Pr(E = 1|Y = y)$ hence the (unconditional) acceptance probability of each

$$\begin{aligned}
Pr\left\{E = 1\right\} &= \int_{(0,1)} Pr\left\{E = 1|Y = y\right\} f_U(y)dy \\
&= \int_{(0,1)} \frac{f_X(y)}{kf_U(y)} f_U(y)dy \\
&= \int_{(0,1)} \frac{f_X(y)}{k} dy \\
&= \frac{1}{k} \int_{(0,1)} f_X(y)dy \\
&= \frac{1}{k}
\end{aligned}$$

Now we can check whether we will eventually get a simulation from our *target* distribution $f_X(\cdot)$ $X \equiv Y^A$

$$
\begin{aligned}
& Pr\left\{X \le x\right\} = Pr\left\{Y^A \le x\right\} = \\
= \quad & Pr\left\{Y \le x | E = 1\right\} \\
= \quad & \frac{Pr\left\{Y \le x, E = 1\right\}}{Pr\left\{E = 1\right\}} \\
= \quad & \frac{Pr\left\{E = 1, Y \in [0, x]\right\}}{Pr\left\{E = 1, Y \in [0, 1]\right\}} \\
= \quad & \frac{\int_0^x Pr\left\{E = 1 | Y = y\right\} f_U(y) dy}{\int_0^1 Pr\left\{E = 1 | Y = y\right\} f_U(y) dy} = \frac{\int_0^x \frac{f_X(y)}{k f_U(y)} f_U(y) dy}{\int_0^1 \frac{f_X(y)}{k f_U(y)} f_U(y) dy} \\
= \quad & \frac{\frac{1}{k}\int_0^x f_X(y) dy}{\frac{1}{k}\int_0^1 f_X(y) dy} = \int_0^x f_X(y) dy = F_X(x) = Pr\left\{X \le x\right\} \qquad \Longrightarrow X \sim f_X(\cdot)
\end{aligned}
$$

Remarks

- "difficult" part for implementing this scheme is find out a suitable k which allows the corresponding auxiliary density f_U generating candidates Y_i to become f_U a density $k \cdot f_U$ dominating the target f_X
- one can generalize this algorithm to an arbitrary density for candidates (need not be a probability, need not be univariate) as long as it is **dominating** and **easy-to-simulate** from
- Efficiency $\propto$ Acceptance Probability
- This scheme can then be applied even if we do not have the target distribution completely known but only known up to a multiplicative constant. The only important point is that its scaled version can be computed at each point (as, for instance, in the case of a posterior density [numerator])

  In fact the algorithm works thanks to the ratio

  $$\frac{f_X(y)}{kf_U(y)} = \frac{f_X(y)}{k}$$

  [Look at the last line in the previous proof]

Acceptance-Rejection Sampling (general case)

Suppose we need to simulate from a *target* distribution whose density $f_X(x)$ is known up to an unknown finite positive constant c so that $\tilde{f}(x) = c \cdot f_X(x)$ is the corresponding probability density. Suppose we know a distribution with probability density $q(\cdot)$ on $\mathcal{Y}$ a suitable constant k such that

$$f_X(x) \leq kq(x) \quad \forall x \in \mathcal{Y}$$

1. $Y \sim q$
2. $E|Y = y \sim Ber\left(\frac{f_X(Y)}{k \cdot q(Y)}\right)$
3. If $E = 1$ then $X = Y$ else restart from 1.

Key assumptions of the general version of AR sampling

(a) $q(\cdot)$ must be a probability density with support denoted as $\mathcal{Y}$

(b) We should be able to simulate from $q(\cdot)$

(c) The following bounding condition holds

$$f_X(x) \leq kq(x) \quad \forall x \in \mathcal{Y}$$

(d) we must be able to compute all the required elements in the ratio $\frac{f_X(x)}{kq(x)}$ for any x in the support of q denoted as $\mathcal{Y}$

Important consequences of the underlying randomness

(a) There is a joint distribution of the two random quantities

$$(Y, E)$$

Can you recover the constructive way of defining this joint distribution?

(b) An interesting quantity is related to the marginal distribution of E.

(c) The most interesting distribution is the distribution of $Y|E = 1$.

(d) We can give a nice geometric interpretation of $P(E = 1)$ starting from the area under the dominating density $kq(y)$ and the area under the dominated density $f_X(y)$ (need not be [and a.a. it is not] a probability densities) .

Important consequences of the underlying randomness

(a) There is a joint distribution of the two random quantities

$$(Y, E)$$

Can you recover the constructive way of defining this joint distribution?

(b) An interesting quantity is related to the marginal distribution of E. In fact,

$$P(Acc) = P(E = 1) = \frac{\int_{\mathcal{X}} f_X(x)dx}{\int_{\mathcal{X}} kq(y)dy}$$

(c) The most interesting distribution is the distribution of $Y|E = 1$. Indeed we define (and exploit) $X = Y_{acc} = Y|E = 1$

(d) We can give a nice geometric interpretation of $P(E = 1)$ starting from the area under the densities $kq(y)$ (need not be [and a.a. it is not] a probability density) and $f_X(y)$.

Looking forward to Homeworks or future tests. For A/R in the general case:

1. Compute the acceptance probability
2. Verify that the accepted θ are random draws from f_X
3. Show how in Bayesian inference you could use simulations from the prior (auxiliary density) to get a random draw from the posterior (target distribution) without knowing the proportionality constant
4. Illustrate analytically possible difficulties of this approach with a simple conjugate model
5. Verify your conclusions implementing the A-R approach with your conjugate model

Some other insights on A/R and the joint $J(y, e)$

- E is a Bernoulli distributed r.v.
- Y can be regarded as a completion (or a finite mixture if you like)
- $E_1, ..., E_n$ i.i.d Bernoulli
- $\tau = \min\{i : Y_i = Y_i^A = X_1\}$ has essentially a geometric distribution i.e., more precisely, a shifted (by 1) Negative Binomial with parameters $p = P(E = 1)$ and $m = 1$
- simulating from a conditional distribution is equivalent to taking a random subsample!

A/R and restricted/truncated distributions

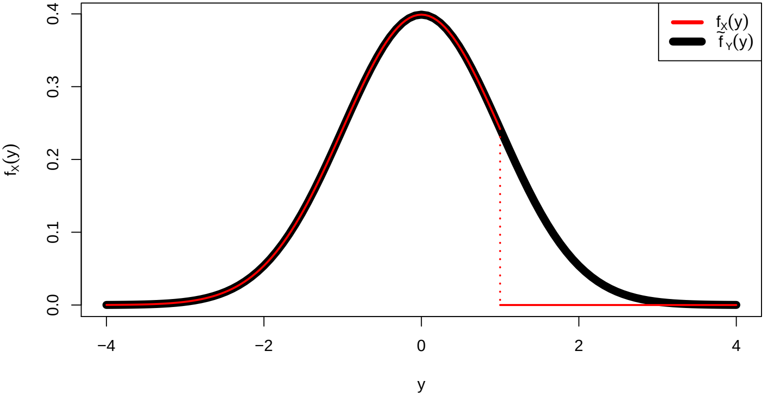
Now we can easily address the simulation of a random variable Y whose probability density (or mass) $\tilde{f}_Y(y)$ is restricted to a proper subset $\mathcal{X}$ of its original support $\mathcal{Y}$ without affecting its functional form on the subset where it is not null. A part from the region where the density is made null the main aspect of the restricted probability distribution which is affected is the normalization constant.
Suppose that $Y \sim \tilde{f}_Y(y)$ with $y \in \mathcal{Y}$ and that we want to consider the distribution of X such that its density, $f_X(\cdot)$, possibly up to a positive proportionality constant is given by

$$f_X(y) = \tilde{f}_Y(y) I_{\mathcal{X}}(y) \qquad\qquad \forall y \in \mathcal{X} \subset \mathcal{Y}.$$

Of course, since $\mathcal{X} \subset \mathcal{Y}$ then

$$I_{\mathcal{X}}(y) \leq 1 \qquad \forall y \in \mathcal{Y}$$

A/R and restricted/truncated distributions



A/R and restricted/truncated distributions

Indeed this means that $\tilde{f}_X(\cdot) \propto \tilde{f}_Y(y) I_{\mathcal{X}}(y)$ with the the probability density of X being

$$\tilde{f}_X(y) = \frac{\tilde{f}_Y(y) I_{\mathcal{X}}(y)}{\int_{\mathcal{Y}} I_{\mathcal{X}}(z) \tilde{f}_Y(z) dz} = \frac{1}{\int_{\mathcal{X}} \tilde{f}_Y(z) dz} \tilde{f}_Y(y) I_{\mathcal{X}}(y)$$

We can then verify that we are in the condition of the A/R algorithm since

$$f_X(y) = \tilde{f}_Y(y) I_{\mathcal{X}}(y) \leq \tilde{f}_Y(y) \qquad \forall y \in \mathcal{Y}$$

Hence if we implement the A/R sampling we have

1. Simulate a candidate $Y \sim \tilde{f}_Y(\cdot)$
2. If $Y = y \in \mathcal{X}$ then acept it and set $X = Y = y$ else go to 1.

In fact the ratio which gives the acceptance probability of the auxiliary Bernoulli experiment is

$$\frac{f_X(y)}{kq(y)} = \frac{\tilde{f}_Y(y) I_{\mathcal{X}}(y)}{\tilde{f}_Y(y)} = I_{\mathcal{X}}(y)$$

hence the auxiliary conditional Bernoulli experiment leads with probability 1 or 0 to the acceptance/rejection.

Consider for example the simulation from a truncated Normal distribution whose density is proportional to

$$f_X(y) \propto I_{(a,b)}(y) \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{-\frac{1}{2\sigma^2}(y-\mu)^2\right\}$$

We can show that the conditions for A/R occur and one can simulate from the restricted distribution in a very intuitive manner:

1. simulate a candidate from the auxiliary distribution $Y \sim N(\mu, \sigma^2)$
2. if $Y = y \in (a, b) = \mathcal{X}$ then set $X = Y = y$ else go to 1.

A/R and restricted/truncated distributions

On the other hand, one can always write

$$\begin{aligned}
\tilde{f}_Y(y) &= \tilde{f}_Y(y) I_{\mathcal{X}}(y) + \tilde{f}_Y(y) I_{\mathcal{Y}\setminus\mathcal{X}}(y) \\
&= \left(\int_{\mathcal{X}} \tilde{f}_Y(z)dz \right) \frac{\tilde{f}_Y(y) I_{\mathcal{X}}(y)}{\int_{\mathcal{X}} \tilde{f}_Y(z)dz} + \left(\int_{\mathcal{Y}\setminus\mathcal{X}} \tilde{f}_Y(z)dz \right) \frac{\tilde{f}_Y(y) I_{\mathcal{Y}\setminus\mathcal{X}}(y)}{\int_{\mathcal{Y}\setminus\mathcal{X}} \tilde{f}_Y(z)dz} \\
&= w f_1(y) + (1-w) f_0(y)
\end{aligned}$$

where

$$f_1(y) = \tilde{f}_X(y)$$

Fundamental Theorem of Simulation

This is a theoretical result which, despite its simplicity, it deserves to be named *fundamental Theorem of Simulation* (sic!).
Let us suppose $f(x)$ is a density on $\mathbb{R}$.
From the trivial identity

$$f(x) = \int_0^{f(x)} dw$$

we can realize that the following subset of the plane $H = \{(u, w) : 0 < w < f(u)\} \subset \mathbb{R}^2$ has Lebesgue measure in $\mathbb{R}^2$ exactly equal to ... . Hence if we take its indicator function $I_H(x, w)$ we get that

$$f(x) = \int I_H(x, w) dw$$

Fundamental Theorem of Simulation

Let us fill the gaps!

In fact the function $I_H(x, w)$ can be interpreted as a probability distribution for .... (random object) in ... (support/domain) whereas ... $f(x)$ corresponds to its ...

Fundamental Theorem of Simulation (cont'd)

Simulating $X \sim f(x)$ and $U|X \sim Unif[0, f(X)]$ corresponds to simulating $(X, U) \sim Unif(H)$ that is a vector (X, U) with joint density

$$f(x, u) = I_H(x, u)$$

Graphically ... there is a connection with the A-R algorithm

In fact we may generalize this for X with support in $\mathbb{R}^n$.

A-R and FTS

In order to simulate uniformly on a subset $A \subset \mathbb{R}^d$ ....
... I can simulate uniformly on $B \subset \mathbb{R}^d$ such that $A \subset B$ and reject all $X = x$ such that

$$x \in B \setminus A$$

In $A - R$ the subsets A and B correspond to $H_f \subset H_{kq}$ and $d = 1 + 1$ or $d = n + 1$ in the general case

ABC basics

In some circumstances have to deal with a statistical model for which

1. we are not able to **compute** $f(x|\theta)$ ...
2. ... but we are able to simulate $Z \sim f(\cdot|\theta)$ for any fixed $\theta \in \Theta$

Suppose we observe a set of data x_{obs} from our statistical model .... how can we make Bayesian inference if we cannot compute the likelihood?

ABC basics

Suppose $X \sim f(x|\theta)$ and $\theta \sim \pi(\theta)$

vanilla ABC algorithm (valid only for discrete observations x_{obs}):

1. Simulate $\theta \sim \pi(\cdot)$
2. Simulate $Z \sim f(\cdot|\theta)$
3. If $Z = x_{obs} \implies$ Accept $\theta_{acc} = \theta$ (although it is not necessary, we could consider the Bernoulli outcome E of the acceptance event)

Indeed we are simulating the "triple" (θ, Z, E) from a joint "density".

- are we able to write the joint distribution?
- are we able to characterize the accepted θ_{acc}?

ABC basics

Indeed the accepted sample of θ_{acc} are simulated from the conditional distribution of $(\theta | Z = x_{obs}, E = 1)$ which can be easily shown to be proportional to to

$$\pi(\theta) f(x_{obs}|\theta)$$

In fact, the joint distribution can be written as follows:

$$\pi(\theta) f(z|\theta) \left[e I_{\{x_{obs}\}}(z) + (1 - e)(1 - I_{\{x_{obs}\}}(z)) \right]$$

Hence the distribution of $\theta | E = 1$ is the same of $\theta | Z = x_{obs}$ (or, equivalently, of $\theta | Z = x_{obs}, E = 1$)

$$\pi(\theta_{acc}) \propto \pi(\theta) f(x_{obs}|\theta)$$

In principle, this would allow us to make posterior inference **without ever computing the likelihood function** but just using random draws and comparing simulated values of Z to the observations x_{obs}. However, there are theoretical and practical drawbacks fot this vanilla ABC idea.

ABC implementation

Difficulty in generalizing the basic discrete settting:

- For typical multivariate observations $(x_1, ..., x_n)$ the event $Z = x_{obs}$ becomes very rare hence one may consider a (sufficent) statistics $T(x_1, ..., x_n)$ or a set of statistics.
- In the continuous case we have that the event $Z = x_{obs}$ has probability 0 so we cannot expect the acceptance to occur in a finite amount of time.

One can "approximate" the acceptance event through the use of a distribution $q_\varepsilon(\cdot)$ which can be thought of as close to be degenerate at x_{obs} [7] as follows:

$$\pi_{\epsilon,ABC}(\theta|x_{obs}) \propto \pi(\theta) \int_{\mathcal{Z}} f(z|\theta) h_\varepsilon\left(S(z) - S(x_{obs})\right) dz$$

[7] for example $h_\varepsilon(u) = \frac{1}{2\varepsilon} I_{[-\varepsilon,\varepsilon]}(u)$ and $q_\varepsilon(x_{obs}, z) = h_\varepsilon\left(S(z) - S(x_{obs})\right)$; technically $q_\varepsilon(x, z)$ csan be regarded as a kernel/transition

ABC error

- replacing the indicator function we modify the target. As $\varepsilon \to 0$ we can get the original as a limit
- non sufficiency of the statistics $S(z)$ alters the target
- usual Monte Carlo error

Bayesian model uncertainty and model selection

We give a very essential list of main approaches in the presence of model uncertainty

- posterior distribution within model index space (straight fully Bayesian approach)
- Marginal likelihood and Bayesian model evidence
- Bayes factor for selecting between two competing models
- DIC as a model index criterion

Marginal likelihood and Bayesian inference ...

In the presence of alternative models $m = 1, 2, \ldots, M$. Each model m has a prior mass $\mathrm{pri}(m)$ and a data distribution

$$f(data|\theta, m)$$

with parameter $\theta \in \Theta_m$ which has in turn its prior distribution on Θ_m denoted with $\pi(\theta|m)$ The joint distribution of *data* and model m only

$$J(data, m) \quad =$$

Marginal likelihood and Bayesian inference ...

In the presence of alternative models $m = 1, 2, \dots, M$. Each model m has a prior mass $\mathrm{pri}(m)$ and a data distribution

$$f(data|\theta, m)$$

with parameter $\theta \in \Theta_m$ which has in turn its prior distribution on Θ_m denoted with $\pi(\theta|m)$ The joint distribution of *data* and model m only pops out if we integrate out the parameter θ

$$\begin{aligned} J(data, m) &= \int_{\Theta_m} f(data|\theta, m)\pi(\theta|m)\, d\theta\, \mathrm{pri}(m) \\ &= \mathrm{pri}(m) \int_{\Theta_m} f(data|\theta, m)\pi(\theta|m)\, d\theta \\ &= \mathrm{pri}(m) b(data|m) \end{aligned}$$

where

$$b(data|m) = \int_{\Theta_m} f(data|\theta, m)\pi(\theta|m) d\theta$$

is the marginal likelihood also named Bayesian evidence of model m provided by *data*

Marginal likelihood and Bayesian inference ...

In fact, the posterior model probabilitiy for model m given the data is derived from the joint distribution of the data and the model

$$J(data, m) = \text{pri}(m) b(data|m)$$

so that

$$\text{post}(m|data) = \frac{J(data, m)}{\sum_{m'} J(data, m')} = \frac{\text{pri}(m) b(data|m)}{\sum_{m'} \text{pri}(m') b(data|m')}$$

Hence, the posterior odds between two alternative models (m_i and m_j) can be factorized as

$$\begin{aligned}
\frac{\text{post}(m_i|data)}{\text{post}(m_j|data)} &= \frac{\frac{q(data|m_i)}{\sum_{m'} q(data|m')}}{\frac{q(data|m_j)}{\sum_{m'} q(data|m')}} \\
&= \frac{q(data|m_i)}{q(data|m_j)} \\
&= \frac{\text{pri}(m_i)}{\text{pri}(m_j)} \frac{b(m_i|data)}{b(m_j|data)}
\end{aligned}$$

Marginal likelihood and Bayes factor

Where the ratio

$$BF_{ij} = \frac{b(m_i|data)}{b(m_j|data)}$$

is called Bayes Factor of model m_i with respect to model m_j regarded as the ratio of posterior odds to prior odds.

Bayes factor calibration

You may find alternative guidelines on the scale interpretation. The following is one of the most cited[8]

BF_{ij}	Interpretation
Under 1	Supports model j
1-3	Weak support for model *i* - not worth more than a bare mer
3-20	Support for model i
20-150	Strong support for model i
Over 150	Very strong support for model i

[8] Kass and Raftery (1995), *Bayes Factors*, JASA

Penalized Likelihood Criteria

There is an alternative approach which mimics the classical recipe and relies on the deviance $-2\log f(y \mid \theta)$

- We want to compare several models.
- Let p denote the number of parameters in the model and n the number of data points.
- Define the *deviance* as

$$D(y, \theta) = -2\log f(y \mid \theta).$$

- Also define

$$D_{\hat{\theta}}(y) = D(y, \hat{\theta}(y))$$

  as the deviance evaluated at a "representative" point $\hat{\theta}$ (usually the posterior mean, or the posterior mode).

Penalized Likelihood Criteria

In the classical/frequentist world one numeric index used to address model selection is the Akaike Information Criterion

$$AIC = D_{\hat{\theta}}(y) + 2p$$

The AIC is used to choose models that have good “out-of-sample” predictive capabilities.

Penalized Likelihood Criteria

Define the posterior mean deviance as:

$$\begin{aligned} D_{avg}(y) &= E_{\theta\,|\,y}(D(y,\theta)\,|\,y) \\ &= \int D(y,\theta)\pi(\theta\,|\,y)d\theta \\ &= \int -2\log f(y\,|\,\theta)\pi(\theta\,|\,y)d\theta \\ \hat{D}_{avg}(y) &\approx \frac{1}{M}\sum_j -2\log(f(y\,|\,\theta^{(j)})) \end{aligned}$$

The difference between the posterior mean deviance and $D_{\hat{\theta}}(y)$ represents the effect of model fitting and has been used as a measure of the effective number of parameters in a Bayesian model.

The effective number of parameters is

$$p_D = \hat{D}_{\text{avg}}(y) - D_{\hat{\theta}}(y)$$

Penalized Likelihood Criteria

DIC has been introduced in Spiegelhalter, D. J.; Best, N. G.; Carlin, B. P.; van der Linde, A. (2002). "Bayesian measures of model complexity and fit (with discussion)". in Journal of the Royal Statistical Society, Series B. Although it has been shown to suffer some unexpected/undesirable behavior it works in many practical examples and it is currently widely used.

p_D can be thought of as the number of "unconstrained" parameters in the model, where a parameter counts as: 1 if it is estimated with no constraints or prior information; 0 if it is fully constrained or if all the information about the parameter comes from the prior distribution; or an intermediate value if both the data and prior distributions are informative.

Penalized Likelihood Criteria

We then use the deviance information criterion (DIC), which is a generalization of AIC. Define

$$\begin{aligned} \text{DIC} &= D_{\hat{\theta}}(y) + 2p_D \\ &= \hat{D}_{avg}(y) + p_D \\ &= 2\hat{D}_{avg}(y) - D_{\hat{\theta}}(y) \end{aligned}$$

Lower values of the penalized likelihood criteria are better. These criteria do not have an absolute scale and should be used only to rank models.

DIC

- For non-hierarchical models, p_D should be approximately the true number of parameters.
- In models with negligible prior information, DIC will be approximately equivalent to AIC.
- Slightly different values of $\hat{D}_{avg}(y)$ (and hence p_D and DIC) can be obtained depending on the parameterization used for the prior distribution.

DIC

- The minimum DIC estimates the “best” model in the same spirit as Akaike’s criterion. However, if the difference in DIC is, say, less than 5, and the models make very different inferences, then it could be misleading just to report the model with the lowest DIC.
- DICs are comparable only over models with exactly the same observed data, but there is no need for them to be nested.
- DIC can be used only as a **comparative index**: there is no way of considering DIC on a standard absolute scale so it is useful only when different models are compared on the same data set

Marginal Likelihood, Model comparison and Bayes Factor

$$\mathcal{E} = m(x) = \int_{\Theta} L_x(\theta)\pi(\theta)d\theta$$

- $\hat{\mathcal{E}}^{AM}$: Arithmetic Mean estimator (naive Monte Carlo)
- $\hat{\mathcal{E}}^{AM}$: Importance Sampling estimator (using suitable approximating auxiliary $q(\cdot)$)
- $\hat{\mathcal{E}}^{HM}$ Harmonic mean (Newton and Raftery, 1994)
- $\hat{\mathcal{E}}^{GHM}$ Generalized Harmonic mean (Gelfand and Day, 1994)
- $\hat{\mathcal{E}}^{Chib}$ Chib's Candidate formula (Chib, 1995)
- $\hat{\mathcal{E}}^{Lap}$ Laplace approximation
- Bridge Sampling
- Path Sampling or Thermodynamic Integration

Marginal Likelihood approximation: $\hat{\mathcal{E}}^{AM}$

Simulate $\theta_1, ..., \theta_t$ from the prior $\pi(\theta)$ (likely you could directly simulate i.i.d)

$$\hat{\mathcal{E}}^{AM} = \frac{1}{t}\sum_{i=1}^{t} L(\theta_i)$$

Problem: bad approximation in the usual peaked-likelihood situation. Very few points (if any at all) are simulated in the more relevant high-likelihood region.

Marginal Likelihood approximation: $\hat{\mathcal{E}}^{HM}$

Simulate (MCMC sample) $\theta_1, ..., \theta_t$ from the posterior $\pi(\theta|x)$

$$\hat{\mathcal{E}}^{HM} = \frac{1}{\frac{1}{t}\sum_{i=1}^{t} \frac{1}{L(\theta_i)}}$$

Problem: No finite variance guarantee (some ∞-variance examples with conjugate Gaussian models). Many examples of bad behaviour.

Simulation for Approximating quantities of interest

(in particular in our Bayesian inference framework)

The **first important message** is that we are not confined to "simple" (conjugate) class of Bayesian models to derive the posterior quantities of interest since **we can use i.i.d. simulation** combined with Monte Carlo methods and its variants to approximate them

However, in order to implement in practice this idea we need

- to have a way of implementing i.i.d simulation from our target distribution which, most of the times is the posterior distribution which is only known/computable up to a proportionality constant

$$\mathit{target}(\theta) \propto \pi(\theta) L_{y_1,\ldots,y_n}(\theta)$$

- to have the ability of understanding and controlling the error of approximation: the error is random, the MSE (i.e. the variance) should be assessed (usually estimated) and possibly reduced (variance reduction techniques, IS as an example)

Simulation for Approximating quantities of interest

(in particular in our Bayesian inference framework)

The **second important message** is that when **i.i.d. simulation** is not available there is another type of simulation, namely **Markov Chain simulation**, which allows us to achieve the same goal: the approximation of an expected value relying on asymptotic arguments like the convergence of the empirical average to the theoretical mean. However, there are some big important differences that must be understood from the beginning:

- Markov Chains have a more complex structure which makes the relation between the way of simulating the sequence of random variables and the target distribution less direct and requires to get a clear understanding of the stationarity (and asymptotic) properties of the Markov Chain
- the error control requires a different formula in which the presence of dependence (correlation) among the simulated random variables usually increases/inflated the approximation error. Moreover, the way we can estimate this error (in terms of its variance) is less straightforward

Simulation for Approximating quantities of interest

(in particular in our Bayesian inference framework)

The **second important message** is that when **i.i.d. simulation** is not available there is another type of simulation, namely **Markov Chain simulation**, which allows us to achieve the same goal: the approximation of of an expected value relying on asymptotic arguments like the convergence of the empirical average to the theoretical mean. However, there are some big important differences that must be understood from the beginning:

- Markov Chains have a more complex structure which makes the relation between **the way of simulating** the sequence of random variables and the **target distribution** less direct and requires to get a clear understanding of the **stationarity** (and asymptotic) properties of the Markov Chain
- the **error control** requires a different formula in which the presence of dependence (**correlation**) among the simulated random variables usually increases/**inflated the approximation error**. Moreover, the way we can **estimate this error** (in terms of its variance) is less straightforward

MC $\longrightarrow$ MCMC

MC difficulty: direct i.i.d. simulation of $\theta_1, ..., \theta_t, ...$ from π can be hard for distributions outside well known parametric families and becomes prohibitive when π is defined on $\Theta \subset \mathbb{R}^k$ with large k.
Solution: it may be easier to find out a suitable stationary stochastic process $\theta_0, \theta_1, ..., \theta_t, ...$ with some **dependence structure** such that the random behaviour of its coordinates are somehow related to π so that one can establish asymptotic relations with the target π
We are interested in defining stochastic processes for which we can keep on applying the same basic approximation strategy used for vanilla MC i.e.

$$\hat{I} = \frac{1}{t}\sum_{i=1}^{t} h(\theta_i) \xrightarrow{a.s.} E_\pi[h(\theta)] = I \qquad t \to \infty$$

We will achieve this goal relying on the appropriate ergodic properties of Markov processes where π will represent the so-called stationary distribution (invariant) as well as the limit distribution $(t \to \infty)$ of the component θ_t.

The emphasized notions are probably familiar to some of you in the discrete state space setting

In order to understand in more depth MCMC ...

... one should be aware and confident with the following:

- definition of a stochastic process and the main ingredients to set up its probability law
- definition of a **stationary** stochastic process
- definition of a Markov property for a stochastic process
- 2 main ingredients to set up a Markov chain
- definition of **invariant distribution** for the transition (kernel) of a Markov chain
- the connection between the invariance distribution and the stationary distribution for a (stationary) Markov chain
- Monte Carlo error control in MCMC: variance formula and the estimation of the asymptotic variance
- DBC and reversibility
- MCMC algorithms and kernels: kernel composition and hybrid algorithms

Stochastic Process

$\{X_t; t \in \mathcal{T}\}$	indexed collection of random variables
$X_t \in \mathcal{S}$	state space
$\mathcal{T}$	index set

The simplest way to regard a stochastic process is to consider it as a collection of random variables indexed by some set $\mathcal{T}$, taking values in some set $\mathcal{S}$. $\mathcal{T}$ is the **index set**, usually time (ordered), e.g. $\mathbb{Z},^+$, $\mathbb{R}$, $\mathbb{R}^+$. $\mathcal{S}$ is the **state space**, e.g.$\{1, 2, ..., k\}$, $\{a, b, c\}$ Z^+, $\mathbb{R}$, $\mathbb{R}^d$. Stochastic processes are classified according to both the index set (discrete or continuous) and the state space (finite, countable or uncountable/continuous).

Stochastic Process

The simplest way to regard a stochastic process is to consider it as a collection of random variables indexed by some set $\mathcal{T}$, taking values in some set $\mathcal{S}$. $\mathcal{T}$ is the **index set**, usually time (ordered), e.g. Z^+, $\mathbb{R}$, $\mathbb{R}^+$. $\mathcal{S}$ is the **state space**, e.g.$\{1, 2, ..., k\}$, $\{a, b, c\}$ Z^+, $\mathbb{R}$, $\mathbb{R}^d$. Stochastic processes are classified according to both the index set (discrete or continuous) and the state space (finite, countable or uncountable/continuous).

Probability law of a stochastic process

The most general way of specifying uniquely the probability law of a stochastic process is to assign its finite-dimensional distributions (fi-dis)

$$\mu_{t_1,t_2,\ldots,t_k}(A_1,\ldots,A_j,\ldots,A_k) = Pr\{X_{t_1} \in A_1,\ldots,X_{t_k} \in A_k\}$$

for any collection of k-tuples (possibly ordered $t_1 < t_2 < \ldots < t_k$) of indexes t_j in the index set $\mathcal{T}$ and any $A_j \in \sigma(\mathcal{S})$.

Probability law of a stochastic process

Finite-dimensional distributions uniquely specify the law of a stochastic process if a suitable "compatibility" condition hold (Kolmogorov conditions) for any collection of k-tuples and p-tuples $\{t_1, t_2, ..., t_k\} \subset \{s_1, s_2, ..., s_p\}$ of the index set $\mathcal{T}$. Slightly informally the finite dimensional distribution of $(X_{t_1}, ..., X_{t_k})$ must coincide with the marginal distribution of $(X_{s_1}, ..., X_{s_p})$.

Example for t-uples $\{1, 3, 8\} \subset \{1, 2, 3, 5, 7, 8\}$

$$\mu_{1,3,8}(A_1, A_3, A_8) = \mu_{1,2,3,5,7,8}(A_1, \mathcal{S}, A_3, \mathcal{S}, \mathcal{S}, A_8)$$

Probability law of a stochastic process

Finite-dimensional distributions uniquely specify the law of a stochastic process if suitable "compatibility" conditions hold (Kolmogorov conditions) for any collection of k-tuples and p-tuples ($\{t_1, t_2, ..., t_k\} \subset \{t_1, t_2, ..., t_k, t_{k+1}, ..., t_{k+m}\}$ $(k + m = p)$ of the index set $\mathcal{T}$. Slightly informally the finite dimensional distribution of $(X_{t_1}, ..., X_{t_k})$ must coincide with the marginal distribution of $(X_{t_1}, ..., X_{t_p})$.

Example

$$\begin{aligned}
\mu_{1,3,8}(A_1, A_3, A_8) &= Pr\{X_1 \in A_1 \cap X_3 \in A_3 \cap X_8 \in A_8\} \\
&= Pr\{X_1 \in A_1 \cap X_2 \in \mathcal{S} \cap X_3 \in A_3 \cap X_5 \in \mathcal{S} \cap X_7 \in \mathcal{S} \cap X_8 \in A_8\} \\
&= \mu_{1,2,3,5,7,8}(A_1, \mathcal{S}, A_3, \mathcal{S}, \mathcal{S}, A_8, \mathcal{S})
\end{aligned}$$

Probability law of a stochastic process

Indeed, the previous formulas highlight also that from the finite dimensional distributions corresponding to consecutive time indexes one can derive the finite dimensional distributions corresponding to non-consecutive time indexes

Example

$$\begin{aligned}
\mu_{1,3,6}(A_1, A_3, A_6) &= Pr\{X_1 \in A_1 \cap X_3 \in A_3 \cap X_6 \in A_6\} \\
&= Pr\{X_1 \in A_1 \cap X_2 \in \mathcal{S} \cap X_3 \in A_3 \cap X_4 \in \mathcal{S} \cap X_5 \in \mathcal{S} \cap X_6 \in A_6\} \\
&= \mu_{1,2,3,4,5,6}(A_1, \mathcal{S}, A_3, \mathcal{S}, \mathcal{S}, A_6)
\end{aligned}$$

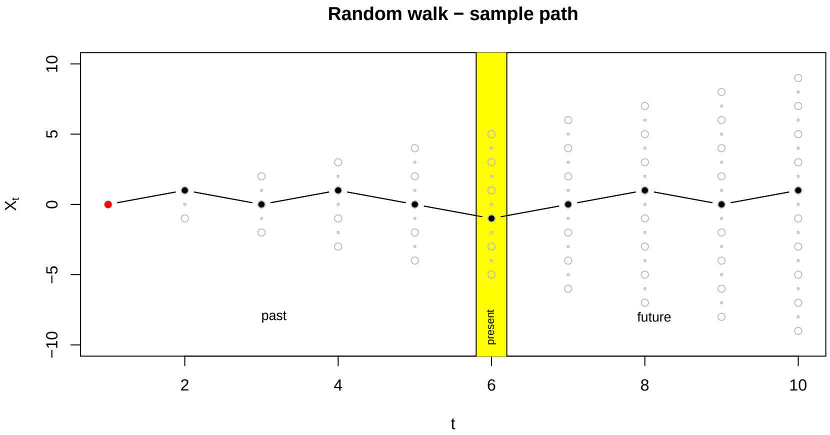
Definition A Markov chain on a discrete state space $\mathcal{S}$ is a stochastic process indexed by a discrete time index t $\{X_t; t = 0, 1, ...\}$ such that $\forall i, j, .., r, s \in \mathcal{S}$

$$Pr\{X_{t+1} = r | X_0 = i, X_1 = j, ..., X_t = s\} = Pr\{X_{t+1} = r | X_t = s\}$$

Markov ≈ The future is independent from the past given the present

We actually need a more general state space $\mathcal{S}$ and hence we will refer to a more general theory (and notation).

The future is independent from the past given the present



Definition A Markov chain on a discrete state space $\mathcal{S}$ is homogeneous if

$$Pr\{X_{t+1} = r | X_t = s\} = p_{sr} \qquad\qquad \forall\, t \in \mathcal{T}$$

The (stochastic) matrix P with generic entry p_{sr} is called transition probability matrix (t.p.m.)

Remarkable example of a Markov chain is the description of the accumulation of money for a better. It is a special example of Markov chain called random walk. In principle the state space is $\mathcal{S} = \mathbb{Z}$. In this case the transition probability matrix is as follows

$$P = \begin{pmatrix} \dots & & & \dots & & & \dots \\ \dots & 1-p & 0 & p & 0 & 0 & \dots \\ \dots & 0 & 1-p & 0 & p & 0 & \dots \\ \dots & 0 & 0 & 1-p & 0 & p & \dots \\ \dots & & & \dots & & & \dots \end{pmatrix}$$

Let us consider another very simple example with $\mathcal{S} = \{1, 2\}$. In this case we set the transition probability matrix as follows

$$P = \begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix} = \begin{pmatrix} 0.8 & 0.2 \\ 0.5 & 0.5 \end{pmatrix}$$

Stability properties of a stochastic process

Remind that we will be interested in exploiting a Markov chain in order to simulate $X_0, ..., X_t$ and then take

$$\hat{I}_t = \frac{1}{t} \sum_{i=1}^{t} h(\theta_i)$$

in the hope that the process will have suitable ergodic properties which will ensure that

$$\hat{I}_t \xrightarrow{a.s.} E_\pi[h(\theta)] = I \qquad t \to \infty$$

similarly to the standard Monte Carlo recipe.
Some fundamental distinctions:

- the simulated random variables will no longer be independent
- it is not (yet) apparent how π is related to our stochastic process since $\pi(\cdot)$ is not (explicitly) in the basic ingredients

Stationarity

Strict Stationarity
The process $(X_t, t \in \mathcal{T})$ is said to be strictly stationary if for any k-tuples $\{t_1, t_2, ..., t_k\}$ the finite dimensional distribution of $(X_{t_1}, ..., X_{t_k})$ is the same distribution of $(X_{t_1+h}, ..., X_{t_k+h})$.

Remark Strict stationarity implies that all existing moments (whenever they exist) are time invariant

Two examples implementing simulation of a finite portion of a Markov chain

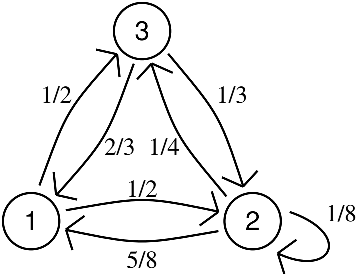
- in the (finite) discrete case
- in the absolutely continuous case

We will see possibly "unstable" behaviour

```
#### Discrete Markov Chain simulation
#### with only two states
#### S={1,2}

alpha<-0.2
beta<-0.5
mpt<-matrix(c(1-alpha,alpha,beta,1-beta),nrow=2,byrow=T)
S=c(1,2)  # discrete state space

x1<-2 ## starting value of the chain

nsample<-1000
chain<-rep(NA,nsample+1) # vector that will hold
                         # all the simulates values
chain[1]<-x1             # starting value x1 assigned to chain[1]
for(t in 1:nsample){
  chain[t+1]<-sample(S,size=1,prob=mpt[chain[t],])
}
plot(chain,ylim=c(0,4))
table(chain)
```

Another example

[9]



[9] This is NOT a DAG!!

Using the two ingredients and the Markov property I can easily write the joint finite dimensional of the random states in first t times

$$
\begin{aligned}
&Pr\{X_0 = i, X_1 = j, X_2 = k, ..., X_{t-1} = r, X_t = s\} = \\
&Pr\{X_0 = i\}Pr\{X_1 = j|X_0 = i\}Pr\{X_2 = k|X_1 = j\}...Pr\{X_t = s|X_{t-1} = r\}
\end{aligned}
$$

$$Pr\{X_0 = i, X_1 = j, X_2 = k, ..., X_{t-1} = r, X_t = s\} = \mu_i p_{ij} p_{jk} ... p_{rs}$$

From this one can easily compute any

$$
\begin{aligned}
&Pr\{X_0 \in A_0, X_1 \in A_1, X_2 \in A_2 ..., X_{t-1} \in A_{t-1}, X_t \in A_t\} = ... \\
&\sum_{i \in A_0} \sum_{j \in A_1} .... \sum_{r \in A_{t-1}} \sum_{s \in A_t} \mu_i p_{ij} p_{jk} ... p_{rs}
\end{aligned}
$$

In the discrete case let us derive $Pr\left\{X_1 = j\right\}$ the probability of $X_1 = j$ if the Markov Chain starts randomly at time 0 according to the probability distribution represented by the (column) vector μ

$$X_0 \sim \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \dots \\ \mu_r \\ \dots \\ \mu_d \end{pmatrix}$$

$$P = \begin{pmatrix} p_{11} & p_{12} & \dots & p_{1j} & \dots & p_{1d} \\ p_{21} & p_{22} & \dots & p_{2j} & \dots & p_{2d} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ p_{r1} & p_{r2} & \dots & p_{rj} & \dots & p_{rd} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ p_{d1} & p_{d2} & \dots & p_{dj} & \dots & p_{dd} \end{pmatrix}$$

$$\mu^T P$$

$$\begin{pmatrix} \mu_1 & \mu_2 & \dots & \mu_r & \dots & \mu_d \end{pmatrix} \begin{pmatrix} p_{11} & p_{12} & \dots & p_{1j} & \dots & p_{1d} \\ p_{21} & p_{22} & \dots & p_{2j} & \dots & p_{2d} \\ & \dots & \dots & \dots & \dots & \dots \\ p_{r1} & p_{r2} & \dots & p_{rj} & \dots & p_{rd} \\ & \dots & \dots & \dots & \dots & \dots \\ p_{d1} & p_{d2} & \dots & p_{dj} & \dots & p_{dd} \end{pmatrix}$$

$$\mu^T P = \begin{pmatrix} Pr\left\{X_1 = 1\right\} & Pr\left\{X_1 = 2\right\} & \dots & Pr\left\{X_1 = j\right\} & \dots & Pr\left\{X_1 = d\right\} \end{pmatrix}$$

Remark – The matrix product $\mu^T P$ basically allows us to derive all the probabilities $Pr\left\{X_1 = j\right\}$ for any $j = 1, \dots, d$.

```
#### Markov chain simulation on a continuous state-space:
####this is an autoregressive process AR(1)
mcsize=20

present=3.9
chain=c(present)
print(present)
# initialization at time t=0

# loop for sequential updating from present to next future
# using a transition kernel corresponding to an
# absolutely continuous (gaussian) distribution
# whose mean depends on the present state

for(i in 1:mcsize){

rho=0.7 # try also rho=1.0001 or rho=0.999
next.future=rnorm(1,mean=rho*present,sd=sqrt(1))
present=next.future
print(present)
chain=c(chain,present)
}
```

Basic ingredients for setting-up a Markov Chain on a general state space

Let $t \in \{0, 1, 2, \dots\}$ be the index of the process and $\mathcal{S} \subset \mathbb{R}^k$ the general space of states.
In simple words:

1. *Start* :: starting state at time 0: fixed $(X_0 = x_0)$ or random $(X_0 \sim \mu)$
2. *Transition* :: randomness of the state in the next future time-point once the state of the present time point is known

In graphical notation:

1. *Start* :: the first node
2. *Transition* :: the arrow and the node starting from the previous time-point node

Basic ingredients for setting-up a Markov Chain on a general state space

More formally:
Let $t \in \{0, 1, 2, \dots\}$ be the index of the process and $\mathcal{S} \subset \mathbb{R}^k$ the general space of states.

1. (*Start*) $\mu \rightarrow$ Initial distribution at time $t = 0$
2. (*Transition(s)*) $\underline{\text{Transition kernel(s)}}$
   $K_t(x, A) = Pr\{X_{t+1} \in A | X_t = x\}$ for each $t = 1, 2, \dots$

The transition kernel (at time t) is a function
$K_t(\cdot, \cdot) : \mathcal{S} \times \mathcal{B}(\mathcal{S}) \rightarrow [0, 1]$

- $\forall x \in \mathcal{S}$ $K_t(x, \cdot)$ is a probability measure
- $\forall A \in \mathcal{B}(\mathcal{S})$ $K_t(\cdot, A)$ is measurable

If $K_t(\cdot, A) = K(\cdot, A)$ the Markov Chain is said to be $\underline{\textit{homogeneous}}$

- When $K_t(x, A) = K_1(x, A)$ for all $t \Longrightarrow$ we say that the chain is homogeneous
- particular cases:
  - when the kernel indexes absolutely continuous (conditional) distributions it can be represented in terms of a transition density usually denoted with $q(x, y) = f_{X_{t+1}|X_t}(y|x)$
  - when the kernel indexes discrete distributions $Pr_{X_{t+1}|X_t}(y|x) = p_{xy}$ one can equivalently use a [→ transition probability matrix $P = (p_{xy})$]

Now let us consider a specific type of MC where we have absolutely continuous distribution conditional distribution $f(z|x)$ as basic ingredient:

- easy to simulate any initial part of the chain $(\theta_1, ..., \theta_t)$
- easy to derive the joint probability law $j(\theta_1, ..., \theta_t)$

It is very easy to simulate the initial path $(\theta_1, ..., \theta_t)$ of a Markov chain defined through the transition density $q(x, z) = f(z|x)$
Let us fix the initial state $\theta_0 = x_0$

$$\begin{array}{ll}
\text{get the realization } \theta_1 = \theta'_1 & \text{simulating } \theta_1 \sim f(\cdot|x_0) = q(x_0, \cdot) \\
\text{get the realization } \theta_2 = \theta'_2 & \text{simulating } \theta_2 \sim f(\cdot|\theta'_1) = q(\theta'_1, \cdot) \\
\text{get the realization } \theta_3 = \theta'_3 & \text{simulating } \theta_3 \sim f(\cdot|\theta'_2) = q(\theta'_2, \cdot) \\
\dots & \dots \\
\text{get the realization } \theta_t = \theta'_t & \text{simulating } \theta_t \sim f(\cdot|\theta'_{t-1}) = q(\theta'_{t-1}, \cdot)
\end{array}$$

$$j_{x_0}(\theta'_1, \ldots, \theta'_t) = f(\theta'_1 | x_0) \cdot f(\theta'_2 | \theta'_1) \cdot \ldots \cdot f(\theta'_t | \theta'_{t-1})$$

How to determine all the distributions of interest for the process

if the process at the initial time $(t = 0)$
starts from a fixed state x_0 $\Longrightarrow$ $P_{x_0}\{X_0 \in A\} = \delta_{x_0}(A)$

$$\begin{aligned}
P_{x_0}\{X_1 \in A\} &= K(x_0, A) = \int_A K(x_0, dy) \\
P_{x_0}\{(X_1, X_2) \in A_1 \times A_2\} &= \int_{A_1} K(y, A_2) K(x_0, dy) \\
P_{x_0}\{(X_1, X_2, X_3) \in A_1 \times A_2 \times A_3\} &= \int_{A_2} K(z, A_3) \int_{A_1} K(y, dz) K(x_0, dy) \\
P_{x_0}\{(X_1, X_2, \dots, X_t) \in A_1 \times A_2 \times \dots \times A_t\} &= \dots
\end{aligned}$$

$$P^t_{x_0}(A) = P_{x_0}\{X_t \in A\} = K^t(x_0, A) = \int_{\mathcal{S}} K(y, A) K^{t-1}(x_0, dy)$$

Example of a Markov chain on a discrete space

- How let us look at how we represent the kernel
  Given a (conditional) distribution $\{p_x(x_i); x_i \in \mathcal{S}\}$ on a discrete state space $\mathcal{S} \Longrightarrow K(x, A) = \sum_{i: x_i \in A} p_x(x_i)$
- how we can simulate a realization of the chain in t consecutive times
  Let us take a very simple case:
  $\mathcal{S} = \{1, 2\}$

$$K(x, A) \to \begin{bmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{bmatrix}$$

(suppose for simplicity that $\mu = \delta_{\{1\}}$)

Some parallel notations in the absolutely continuous case

To understand the new general notation let us consider the following particular case using the notation you are more familiar with. Suppose that a random vector with two components (X_1, X_2) is a.c. with joint density $f_{(X_1,X_2)}(y,z)$ which can be also decomposed as a product $f_{(X_1,X_2)}(y,z) = f_{X_1}(y) f_{X_2|X_2}(z|y)$. If we need to compute the probability $Pr\{(X_1, X_2) \in A_1 \times A_2\}$ we can write

$$\begin{aligned} Pr\{(X_1, X_2) \in A_1 \times A_2\} &= \int_{A_1 \times A_2} f_{(X_1,X_2)}(y,z) dy\, dz = \\ &= \int_{A_1} \int_{A_2} f_{X_1}(y) f_{X_2|X_1}(z|y) dz\, dy \\ &= \int_{A_1} \left[\int_{A_2} f_{X_2|X_1}(z|y)\, dz \right] f_{X_1}(y) dy \end{aligned}$$

Some parallel notations in the absolutely continuous case

Integrating a function of y with respect to the measure represented by $f_{X_1}^{x_0}(y)dy$ is rephrased now in a more general way as the integration with respect to $K(x_0, dy)$ (consider that the x_0 plays no role for the time being ....)

$$\begin{aligned}
Pr\{(X_1, X_2) \in A_1 \times A_2\} &= \int_{A_1 \times A_2} f_{(X_1,X_2)}(y, z)dydz = \\
&= \int_{A_1} \int_{A_2} f_{X_1}^{x_0}(y) f_{X_2|X_2}(z|y)dz\,dy \\
&= \int_{A_1} \left[\int_{A_2} f_{X_2|X_2}(z|y)\,dz \right] f_{X_1}^{x_0}(y)dy \\
&= \int_{A_1} [K(y, A_2)]\; K(x_0, dy)
\end{aligned}$$

Let us now consider the case where the Markov kernel can be represented in terms of (conditional) densities:

$$K(x, A) = \int_A q(x, y)dy = \int_A f(y|x)dy$$

$q(x, y)$ is the (one-step) transition density from x to y
Analogous manipulations allow us to get the transition densities in 2,...,t steps
$q^1(x, y) = q(x, y)$
$q^2(x, y) = \int_{\mathcal{S}} q(x, z)q(z, y)dz$
...
$q^t(x, y) = \int_{\mathcal{S}} q^{t-1}(x, z)q(z, y)dz$

Some parallel notations in the absolutely continuous case

.... earlier was ....

$$P_{x_0}\{X_1 \in A\} = K(x_0, A) = \int_A K(x_0, dy) =$$

$$P_{x_0}\{(X_1, X_2) \in A_1 \times A_2\} = \int_{A_1} K(y, A_2)K(x_0, dy)$$

$$P_{x_0}\{(X_1, X_2, X_3) \in A_1 \times A_2 \times A_3\} = \int_{A_2} K(z, A_3) \int_{A_1} K(y, dz)K(x_0, dy)$$

$$\dots$$

$$P^t_{x_0}(A) = P_{x_0}\{X_t \in A\} = K^t(x_0, A) = \int_{\mathcal{S}} K(y, A)K^{t-1}(x_0, dy)$$

Some parallel notations in the absolutely continuous case

.... in the a.c. case is ....

$$
\begin{aligned}
&P_{x_0}\{X_1 \in A\} = K(x_0, A) = \int_A q(x_0, y)dy = \int_A f(y|x_0)dy \\
&P_{x_0}\{(X_1, X_2) \in A_1 \times A_2\} = \int_{A_1}\left[\int_{A_2} q(y,z)dz\right] q(x_0, y)dy \\
&P_{x_0}\{(X_1, X_2, X_3) \in A_1 \times A_2 \times A_3\} = \int_{A_1}\int_{A_2}\int_{A_3} q(z,w)dw\ q(y,z)dz\ q(x_0,y)dy \\
&\ldots \\
&P^t_{x_0}(A) = P_{x_0}\{X_t \in A\} = \int_{\mathcal{A}} q^t(x_0, w)dw = \int_{\mathcal{S}}\int_{\mathcal{A}} q(y,z)dz\ q^{t-1}(x_0, y)dy
\end{aligned}
$$

if the process at the initial time $(t = 0)$

starts from a random state

according to the distribution $\mu(\cdot)$ $\Longrightarrow$

$$
\begin{aligned}
P_\mu\{X_0 \in A\} &= \mu(A) \\
P_\mu\{X_1 \in A\} &= \int_{\mathcal{S}} K(x, A)\mu(dx) \\
\dots\dots \\
\dots\dots
\end{aligned}
$$

It is very easy to simulate the initial path $(\theta_1, ..., \theta_t)$ of a Markov chain defined through the transition density $q(x, z) = f(z|x)$
Let us fix the initial state $\theta_0 = x_0$

get the realization $\theta_1 = \theta'_1$ simulating $\theta_1 \sim f(\cdot|x_0) = q(x_0, \cdot)$
get the realization $\theta_2 = \theta'_2$ simulating $\theta_2 \sim f(\cdot|\theta'_1) = q(\theta'_1, \cdot)$
get the realization $\theta_3 = \theta'_3$ simulating $\theta_3 \sim f(\cdot|\theta'_2) = q(\theta'_2, \cdot)$
... ...
get the realization $\theta_t = \theta'_t$ simulating $\theta_t \sim f(\cdot|\theta'_{t-1}) = q(\theta'_{t-1}, \cdot)$

After t steps I get a random value which comes from the following:
$\theta_t \sim P^t_{x_0}(\cdot) = K^t(y, \cdot)$
We want to exploit the ergodic properties (which hold under appropriate conditions, not always) such that
$P^t_{x_0}(\cdot) \to P^\infty_{x_0}(\cdot) = \pi(\cdot)$
or, more specifically,

$$\hat{I} = \frac{1}{t}\sum_{i=1}^{t} h(\theta_i) \xrightarrow{a.s.} E_\pi[h(\theta)] = I \qquad t \to \infty$$

or (even better)

$$\hat{I} = \frac{1}{t}\sum_{i=T_0}^{T_0+t} h(\theta_i) \xrightarrow{a.s.} E_\pi[h(\theta)] = I \qquad t \to \infty$$

Invariant measure and stationarity of a Markov chain

A measure π (need not be a probability measure) is said to be invariant with respect to $K(\cdot,\cdot)$ if

$$\pi(A) = \int_{\mathcal{S}} K(y, A)\pi(dy)$$

If π is an invariant probability measure w.r.t. $K(\cdot,\cdot)$ and the initial distribution of the Markov chain $X_0 \sim \mu$ is set $\mu = \pi$ then the chain stays invariant or in equilibrium for all t and corresponds to a (strongly) stationary process such that for all $t > 0$ $Pr_\pi\{X_t \in A\} = \pi(A)$.

Invariant measure and stationarity of a Markov chain

Of course if we start from a degenerate distribution at $t = 0$, i.e. $X_0 \sim \mu(\cdot) = \delta_{x_0}(\cdot)$ we expect the distributions $P^t_{x_0}(\cdot)$ to be different from the starting one. However, the invariant probability measure is the natural candidate that allows us to obtain stronger results like $P^t_{x_0}(\cdot) \to P^\infty(\cdot) = \pi(\cdot)$ i.e. ergodic results no matter what the fixed starting point x_0 is.
In fact, we will see that if the chain enjoys the irreducibility property the invariant probability measure is unique.

Invariant measure and stationarity of a Markov chain

In the **absolutely continuous case** when $K(x, A) = \int_A q(x, z)\, dz$ and $\pi(A) = \int_A \pi(z)\, dz$ we say that the measure $\pi(\cdot)$ is invariant with respect to the kernel $K(\cdot, \cdot)$ or, equivalently, that the density $\pi(z)$ is invariant with respect to the transition density $q(x, y)$ when the following holds

$$\pi(y) = \int_S q(x, y)\pi(x) dx \qquad\qquad \forall y \in \mathcal{S}$$

Invariant measure and stationarity of a Markov chain

In the **discrete case** when $K(x, A)$ can be uniquely represented in terms of a transition probability matrix P (basically considering only subsets made of singletons $A = \{s\}$ so that $K(x, \{s\}) = p_{xs}$ we say that the discrete probability measure $\pi(\cdot)$, i.e. the vector π of masses π_s, is invariant[10] with respect to the kernel $K(\cdot, \cdot)$, i.e. to the transition probability matrix $P = (p_{rs})$, when the following holds with π^T being a row vector

$$\pi^T P = \pi^T$$

or, equivalently[11],

$$P^T \pi = \pi$$

in the usual column vector notation for π.

[10] 2025-Stationary-distribution-finite-state.pdf

[11] see our 2-state example in the pdf document here and the corresponding .Rmd code here

In the discrete case let us derive $Pr\left\{X_1 = j\right\}$ the probability of $X_1 = j$ if the Markov Chain starts randomly at time 0 according to the probability distribution represented by the (column) vector π

$$X_0 \sim \pi = \begin{pmatrix} \pi_1 \\ \pi_2 \\ \dots \\ \pi_r \\ \dots \\ \pi_d \end{pmatrix}$$

$$P = \begin{pmatrix} p_{11} & p_{12} & \dots & p_{1j} & \dots & p_{1d} \\ p_{21} & p_{22} & \dots & p_{2j} & \dots & p_{2d} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ p_{r1} & p_{r2} & \dots & p_{rj} & \dots & p_{rd} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ p_{d1} & p_{d2} & \dots & p_{dj} & \dots & p_{dd} \end{pmatrix}$$

$\pi^T P$

$$\pi^T P = \begin{pmatrix} Pr\left\{X_1 = 1\right\} & Pr\left\{X_1 = 2\right\} & \dots & Pr\left\{X_1 = j\right\} & \dots & Pr\left\{X_1 = d\right\} \end{pmatrix}$$

$$\begin{pmatrix} \pi_1 & \pi_2 & \dots & \pi_r & \dots & \pi_d \end{pmatrix} \begin{pmatrix} p_{11} & p_{12} & \dots & p_{1j} & \dots & p_{1d} \\ p_{21} & p_{22} & \dots & p_{2j} & \dots & p_{2d} \\ & \dots & \dots & \dots & \dots & \dots \\ p_{r1} & p_{r2} & \dots & p_{rj} & \dots & p_{rd} \\ & \dots & \dots & \dots & \dots & \dots \\ p_{d1} & p_{d2} & \dots & p_{dj} & \dots & p_{dd} \end{pmatrix}$$

Remark – The matrix product $\pi^T P$ basically allows us to derive all the probabilities $Pr\left\{X_1 = j\right\}$ for any $j = 1, \dots, d$.

$$\mu^T P$$

$$\begin{pmatrix} \mu_1 & \mu_2 & \dots & \mu_r & \dots & \mu_d \end{pmatrix} \begin{pmatrix} p_{11} & p_{12} & \dots & p_{1j} & \dots & p_{1d} \\ p_{21} & p_{22} & \dots & p_{2j} & \dots & p_{2d} \\ & \dots & \dots & \dots & \dots & \dots \\ p_{r1} & p_{r2} & \dots & p_{rj} & \dots & p_{rd} \\ & \dots & \dots & \dots & \dots & \dots \\ p_{d1} & p_{d2} & \dots & p_{dj} & \dots & p_{dd} \end{pmatrix}$$

$$\mu^T P = \begin{pmatrix} Pr\{X_1 = 1\} & Pr\{X_1 = 2\} & \dots & Pr\{X_1 = j\} & \dots & Pr\{X_1 = d\} \end{pmatrix}$$

Remark – The matrix product $\mu^T P$ basically allows us to derive all the probabilities $Pr\{X_1 = j\}$ for any $j = 1, \dots, d$.

Stationarity of the Markov chain defined by a Markov kernel $K(x, \cdot)$ depends on the possibility of finding a suitable π such that

$$X_n \sim \pi \Longrightarrow X_{n+1} \sim \pi$$

However, one can have approximate stationarity of the Markov chain whenever regularity conditions of the Markov kernel can ensure that

$$K^{(n)}(x_0, \cdot) = Pr\left\{X_{t+n} \in \cdot \middle| X_t = x_0\right\} \approx \pi(\cdot)$$

What we can expect ....

... from a (homogeneous) Markov chain with

1. a kernel $K(\cdot,\cdot)$ for which $\pi(\cdot)$ is invariant
2. initial distribution $\mu(\cdot) = \pi(\cdot)$

- the stochastic process can be stationary (if it starts from an appropriate $\mu(\cdot)$), or approximately stationary after a while (under suitable regularity condition).
- in particular if $\mu(\cdot) = \pi(\cdot)$, $X_i \sim \pi(\cdot)$ for all $i = 0, 1, 2, \ldots, t, \ldots$

The marginal distribution is not enough to completely characterize the limiting behavior of

$$\hat{I}_t = \frac{1}{t} \sum_{i=1}^{t} h(\theta_i)$$

It suffices to derive what happens on average ....
.. but we must be careful that the randomness of $\hat{I}_t$ will also depend on the dependence structure of the process ....

We will need suitable conditions to ensure that for our homogeneous Markov Chain

- the invariant distribution is unique
- it will admit a suitable asymptotic behaviour of the corresponding empirical means (a.s. convergence to a degenerate distribution, asymptotic normality, ...) [ergodicity]

Nothing is automatic/obvious: counterexample (see code `2025-W-12-Markov-Chain-simulations.R`)

Can you answer the following question?

- Can a homogeneous Markov Chain [characterized the initial distribution $\mu(\cdot)$ and the kernel $K(\cdot, \cdot)$] be a strictly stationary stochastic process?

Before providing details on the ergodic theory of Markov chains on a general state space we first look at our first goal:

let us find out how to build up a Markov kernel which allows us to get our target π as a stationary/invariant distribution.

We will look at two basic algorithms:

- Gibbs Sampling (GS)
- Metropolis-Hasting (MH)

Before providing details on the ergodic theory of Markov chains on a general state space we first look at our first goal:

find out how one can build up a Markov chain
(i.e. find out the appropriate transition rule or kernel)
which allows
to get our target π as a invariant/stationary distribution w.r.t the kernel

We will look at two basic algorithms:

- Gibbs Sampling (GS)
- Metropolis-Hasting (MH)

Gibbs Sampling

In the most simple case where $\boldsymbol{\theta} = (\theta_1, \theta_2) \in \Theta \subseteq \mathbb{R}^2$
Suppose that

1. our target is $\pi(\boldsymbol{\theta}) = \pi(\theta_1, \theta_2)$;
2. but we do not know how to simulate from π (2-dim);
3. but we know how to simulate from the so called (full conditionals) (1-dim):

$$\pi_1(\theta_1|\theta_2) \ ; \ \pi_2(\theta_2|\theta_1)$$

GS → we get the transition rule (i.e. kernel) as a transition density $q(\boldsymbol{x}, \boldsymbol{\theta}) = f(\boldsymbol{\theta}|\boldsymbol{x})$ by following all the steps within the Gibbs cycle

$$f(\boldsymbol{\theta}|\boldsymbol{x}) = f((\theta_1, \theta_2)|(x_1, x_2)) = f(\theta_1, \theta_2|x_1, x_2) = \pi_2(\theta_2|\theta_1)\pi_1(\theta_1|x_2)$$

We verify that π is invariant for this GS kernel

$$\begin{aligned}\int_{\mathbb{R}^2} q(\boldsymbol{x},\boldsymbol{\theta})\pi(\boldsymbol{x})dx &= \pi(\boldsymbol{\theta}) \quad \forall \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^2 \\ \int_{\mathbb{R}^2} f(\boldsymbol{\theta}|\boldsymbol{x})\pi(\boldsymbol{x})dx & \quad \pi(\boldsymbol{\theta}) \quad \forall \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^2\end{aligned}$$

In fact

$$\begin{aligned}\int_{\mathbb{R}}\int_{\mathbb{R}} f(\theta_1,\theta_2|x_1,x_2)\pi(x_1,x_2)dx_1dx_2 &= \\ \int_{\mathbb{R}}\int_{\mathbb{R}} \pi_2(\theta_2|\theta_1)\pi_1(\theta_1|x_2)\pi(x_1,x_2)dx_1dx_2 &= \\ \pi_2(\theta_2|\theta_1)\int_{\mathbb{R}} \pi_1(\theta_1|x_2)\pi_2(x_2)dx_2 &= \\ \pi_2(\theta_2|\theta_1)\int_{\mathbb{R}} \pi(\theta_1,x_2)dx_2 &= \\ \pi_2(\theta_2|\theta_1)\pi_1(\theta_1) &= \pi(\theta_1,\theta_2)\end{aligned}$$

Gibbs Sampling: general case

Some preliminary notation when $\theta = (\theta_1, ..., \theta_i, ..., \theta_k) \in \Theta \subseteq \mathbb{R}^k$. Given a target $\pi(\theta) = \pi(\theta_1, ..., \theta_k)$ the corresponding full conditionals will be denoted with

$$\pi(\theta_i|\theta_{(i)}) = \pi(\theta_i|\theta_1, \theta_2, ..., \theta_{i-1}, \theta_{i+1}, ..., \theta_k)$$

GS → We start at time $t = 0$ with a fixed initial point $\theta^0 = x = (x_1, ..., x_k)$ or, equivalently, with initial distribution $\mu = \delta_{\{x\}}$ In the nest time $t = 1$ the state θ^1 will be simulated through the following sequence of (typically univariate, but not necessarily) simulations ...

$$\begin{array}{lllll}
\theta_1^1 & \sim & \pi(\theta_1|x_2,\ldots,x_k) & \rightarrow & \theta_1^1 = \bar{\theta}_1^1 \\
\theta_2^1 & \sim & \pi(\theta_2|\bar{\theta}_1^1, x_3,\ldots,x_k) & \rightarrow & \theta_2^1 = \bar{\theta}_2^1 \\
\theta_3^1 & \sim & \pi(\theta_3|\bar{\theta}_1^1, \bar{\theta}_2^1, x_4,\ldots,x_k) & \rightarrow & \theta_3^1 = \bar{\theta}_3^1 \\
.. & .. & .. & .. & .. \\
\theta_k^1 & \sim & \pi(\theta_k|\bar{\theta}_1^1,\ldots,\bar{\theta}_{k-1}^1) & \rightarrow & \theta_k^1 = \bar{\theta}_k^1
\end{array}$$

same sequential procedure (with k intermediate steps) for the next times $t = 2, 3, \ldots$.

In summary the Markov transition kernel corresponding to the Gibbs Sampling is specified in terms of a collection of k full conditionals. More precisely

$$\begin{array}{lllll}
\theta_1^{t+1} & \sim & \pi(\theta_1|\theta_{(1)}) & = & \pi(\theta_1|\theta_2^t, \ldots \theta_k^t) = \\
\theta_2^{t+1} & \sim & \pi(\theta_2|\theta_{(2)}) & = & \pi(\theta_2|\theta_1^{t+1}, \theta_3^t \ldots \theta_k^t) \\
.. & .. & .. & .. & .. \\
\theta_k^{t+1} & \sim & \pi(\theta_k|\theta_{(k)}) & = & \pi(\theta_k|\theta_1^{t+1}, \theta_2^{t+1}, \ldots, \theta_{k-1}^{t+1})
\end{array}$$

If we had to write it more explicitly .... it would be .... like a transition kernel in k steps of a non-homogeneous Markov chain ...

Some advantages:

1. GS requires the simulation of a sequence of distributions (full-conditionals [f.c.]) which are defined over a lower-dimensional spaces rather than on the k-dimensional Θ; usually f.c. are univariate although they need not to be.
2. In Bayesian inference one can easily determine the full conditionals looking at the functional form of the posterior $\pi(\theta|Data)$ simply regarded as a function of one component

$$\pi(\theta_i|\theta_{(i)}, Data) \propto \pi(\theta|Data) \propto \pi(\theta)\ell(\theta; Data)$$

Sometimes one can recognize that the functional form of the full-conditional corresponds to some well-known parametric family of distributions which can be easily simulated from

Multivariate gaussian distribution

$$x = (x_1, ..., x_k)^T;\ \mu = (\mu_1, ..., \mu_k)^T,\ \Sigma = (\sigma_{ij})$$

$$f(x|\mu, \Sigma) = (2\pi)^{-\frac{k}{2}} |\Sigma|^{-\frac{1}{2}} exp\left\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right\}$$

Multivariate normal model

$$p_{\boldsymbol{X}}(x_1, \dots, x_p | \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{p}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} exp\left\{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})\right\}$$

Meaning of parameters. Restrictions and dimensionality of parameter space. Brief list of useful properties[12]:

- linear transformations with $k \times p$ full-rank matrix $\boldsymbol{L}$ with $k \leq p$:

  $$\boldsymbol{Lx} \sim N_k(\boldsymbol{L\mu}, \boldsymbol{L\Sigma L}^T)$$

  (indeed, this is a characterizing property when univariate linear transforms (i.e. $k = 1$) are considered)
- marginal distributions
- conditional distributions of partitioned sub vectors using two disjoint integer index sets $\boldsymbol{a}$ and $\boldsymbol{b}$ such that $\boldsymbol{a} \cup \boldsymbol{b} = \{1, 2, \dots, p\}$:

  $$\boldsymbol{x}_{[\boldsymbol{b}]} | \boldsymbol{x}_{[\boldsymbol{a}]}, \boldsymbol{\mu}, \boldsymbol{\Sigma} \sim MVN_b(\boldsymbol{\mu}_{\boldsymbol{b}|\boldsymbol{a}}, \boldsymbol{\Sigma}_{\boldsymbol{b}|\boldsymbol{a}})$$

[12] see K. Mardia, J.T. Kent, J.M. Bibby (1979) Multivariate analysis. Academic Press

Multivariate normal model

... where

$$\boldsymbol{\mu}_{\boldsymbol{b}|\boldsymbol{a}} = \boldsymbol{\mu}_{[\boldsymbol{b}]} + \boldsymbol{\Sigma}_{[\boldsymbol{b},\boldsymbol{a}]}(\boldsymbol{\Sigma}_{[\boldsymbol{a},\boldsymbol{a}]})^{-1}\left(\boldsymbol{x}_{[\boldsymbol{a}]} - \boldsymbol{\mu}_{[\boldsymbol{a}]}\right)$$

and

$$\boldsymbol{\Sigma}_{\boldsymbol{b}|\boldsymbol{a}} = \boldsymbol{\Sigma}_{[\boldsymbol{bb}]} - \boldsymbol{\Sigma}_{[\boldsymbol{ba}]}\boldsymbol{\Sigma}_{[\boldsymbol{aa}]}^{-1}\boldsymbol{\Sigma}_{[\boldsymbol{ab}]}$$

Bivariate gaussian distribution

We will use $\sigma_{12} = \rho\sigma_1\sigma_2$. Moreover we consider the vector $\theta = (\theta_1, \theta_2)$ instead of $x = (x_1, x_2)$. Hence

$$f(\theta_1, \theta_2; \mu_1, \mu_1, \sigma_1^2, \sigma_2^2, \sigma_{12}) =$$

$$\frac{1}{2\pi\sqrt{\sigma_1^2\sigma_2^2(1-\rho^2)}} exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(\theta_1-\mu_1)^2}{\sigma_1^2} + \frac{(\theta_2-\mu_2)^2}{\sigma_2^2} - 2\rho\frac{(\theta_1-\mu_1)(\theta_2-\mu_2)}{\sigma_1\sigma_2}\right]\right\}$$

Toy Example (GS1) - Correlated bivariate normal

```
gibbs.biv.norm<-function(nsim,theta1,theta2,rho){
  theta1vec <- rep(NA,nsim+1)
  theta2vec <- rep(NA,nsim+1)
  theta1vec[1] <- theta1
  theta2vec[1] <- theta2
  for(t in 1:nsim){
    theta1vec[t+1]<-rnorm(1,mean=rho*theta2vec[t],
                            sd=sqrt(1-rho^2))
    theta2vec[t+1]<-rnorm(1,mean=rho*theta1vec[t+1],
                            sd=sqrt(1-rho^2))
  }
  gibbssample<-cbind(theta1vec,theta2vec)
  return(gibbssample)
}
```

Toy Example (GS1) - Correlated bivariate normal

Although this is a toy example (you can simulate directly from a multivariate Normal [how-to]) it can be very instructive for beginners:

- learn how to single out f.c. from the joint
- the starting point
- shape of the joint
- simplicity and rigidity of the f.c. scheme
- effect of ρ on the previous items
- trace plots
- acf

Example (GS2) - Poisson counts with one change-point

(Carlin, Gelfand e Smith, 1992 Applied Statistics, **41**, 389-405)
$(Y_1, ..., Y_{m-1}, Y_m, Y_{m+1}, ... Y_n)$

$$\begin{aligned} Y_i &\sim Poi(\lambda) \;\; i = 1, 2, ..., m \\ Y_j &\sim Poi(\phi) \;\; j = m+1, m+2, ..., n \end{aligned}$$

$$\begin{aligned} \lambda &\sim Gamma(rate = \alpha, shape = \beta) \\ \phi &\sim Gamma(rate = a, shape = b) \\ m &\sim Unif\{1, 2, ..., n-1\} \end{aligned}$$

Prior on $\Lambda \times \Phi \times M = (0,\infty) \times (0,\infty) \times \{1, 2, ..., n-1\}$

$$\begin{aligned}
\pi(\lambda, \phi, m) &= f_{Gamma(\alpha,\beta)}(\lambda) f_{Gamma(a,b)}(\phi) f_{Unif\{1,\ldots,n-1\}}(m) \\
&= \frac{\alpha^\beta}{\Gamma(\beta)} e^{-\alpha\lambda} \lambda^{\beta-1} \; \frac{a^b}{\Gamma(b)} e^{-a\phi} \phi^{b-1} \; \frac{1}{n-1} I_{\{1,\ldots,n-1\}}(m) \\
&\propto e^{-\alpha\lambda} \lambda^{\beta-1} \quad e^{-a\phi} \phi^{b-1} \quad I_{\{1,\ldots,n-1\}}(m) I_{(0,\infty)}(\lambda) I_{(0,\infty)}(\phi)
\end{aligned}$$

Likelihood

$$\begin{aligned}
L_{\mathbf{y}}(\lambda, \phi, m) &= \prod_{i=1}^{m} \frac{e^{-\lambda} \lambda^{y_i}}{y_i!} \prod_{j=m+1}^{n} \frac{e^{-\phi} \phi^{y_j}}{y_j!} I_{\{1,\ldots,n-1\}}(m) I_{(0,\infty)}(\lambda) I_{(0,\infty)}(\phi) \\
&\propto e^{-m\lambda-(n-m)\phi} \lambda^{\sum_{i=1}^{m} y_i} \phi^{\sum_{j=m+1}^{n} y_i} I_{\{1,\ldots,n-1\}}(m) I_{(0,\infty)}(\lambda) I_{(0,\infty)}(\phi)
\end{aligned}$$

Posterior $\pi(\lambda, \phi, m|\mathbf{y}) \propto \;$???

Posterior is no longer factorizable in 3 terms as the prior and it has no standard functional form ....

$$\pi(\lambda, \phi, m|\boldsymbol{y}) \quad \propto$$

... but full conditionals are easily recognized as standard distributions and are easy [13] to simulate from

$$
\begin{aligned}
\pi(\lambda|\phi, m, \boldsymbol{y}) &\propto \\
\pi(\phi|\lambda, m, \boldsymbol{y}) &\propto \\
\pi(m|\lambda, \phi, \boldsymbol{y}) &\propto e^{(\phi-\lambda)m}\lambda^{\sum_{i=1}^{m} y_i}\phi^{\sum_{j=m+1}^{n} y_i}
\end{aligned}
$$

[13] some care is needed to avoid possible overflow problems when computing the masses of the discrete distribution for the integer-valued parameter *m*: use logarithmic scale and subtract a constant corresponding to the maximum mass so that it correspond to a multiplicative constant on the original scale and when the original scale is resumed the maximum mass is equal to 1 (see `max(logci)` in the `Gibbs-Sampling-Changepoint.R` code

Let us think about how we could write a clean code to implement our Gibbs sampling.

In R we can exploit the function cumsum(y) to store all the following quantities

```
stat.y.firstperiod = cumsum(y[1:(n-1)])
stat.y.secondoperiod = sum(y) - cumsum(y[1:(n-1)])
```

so that

`stat.y.firstperiod[m]` $= \sum_{i=1}^{m} y_i$ with $m = 1, 2, ..., n-1$

and, similarly,

`stat.y.secondoperiod[m]` $= \sum_{j=m+1}^{n} y_j = \sum_{k=1}^{n} y_k - \sum_{i=1}^{m} y_i$

MC $\longrightarrow$ MCMC

What shall we do with the simulated values from this Markov Chain?

$$\hat{I} = \frac{1}{t} \sum_{i=T_0+1}^{T_0+t} h(\theta_i) \xrightarrow{a.s.} E_\pi[h(\theta)] = I \qquad t \to \infty$$

We **announce** some relevant questions for MCMC implementation which will be considered in more details later on:

- What is the approximation error?
- How can we compute/approximate it?
- How large should we take t so that we get a good approximations?
- How large should we take T_0 so that the arbitrarily fixed initial state $\theta_0 = x$ does not affect $\hat{I}$ (with a substantial bias)?

Variance of the empirical means

Let $\sigma^2 = Var_\pi\left[h(X_i)\right]$, $\gamma_k = Cov_\pi\left[h(X_0), h(X_k)\right]$ and $\rho_k = \frac{\gamma_k}{\gamma_0} = \frac{\gamma_k}{\sigma^2}$

$$\sigma^2_{\hat{I}_t} = Var_\pi\left[\frac{1}{t}\sum_{i=1}^{t} h(X_i)\right] = \frac{\sigma^2}{t}\left(1 + 2\sum_{k=1}^{t-1}\frac{t-k}{t}\rho_k\right)$$

$$Var\left[\sqrt{t}\ \hat{I}_t\right] = t\sigma^2_{\hat{I}_t} \to \sigma^2\left(1 + 2\sum_{k=1}^{\infty}\rho_k\right) = \tau^2$$

$$\widehat{\sigma}^2_{\hat{I}} = ???$$

The factor between parentheses (which is > 1) is an inefficiency factor (as long as $\rho_k > 0$ prevails) and is used to compute the so-called effective sample size ESS

$$t_{eff} = \frac{t}{1 + 2\sum_{k=1}^{\infty}\rho_k}$$

Effective sample size

The factor between parentheses (which is > 1) is an inefficiency factor and is used to compute the so-called effective sample size ESS

$$t_{eff} = \frac{t}{1 + 2\sum_{k=1}^{\infty} \rho_k}$$

in fact if we used the usual variance formula for the i.i.d. case with the number of simulations n replaced by the expression of t_{eff} we would get the correct evaluation of the variance of the empirical mean

$$\sigma^2_{\hat{I}_t} = Var\left[\hat{I}_t\right] = \frac{Var_\pi[h(X_1)]}{t_{eff}} = \left(1 + 2\sum_{k=1}^{\infty} \rho_k\right)\frac{\sigma^2}{t}$$

Effective sample size

All in all we can interpret ESS as the number of independent identically distributed samples from the target that would produce an approximation through its empirical average $\hat{I}_t$ with the same variance as the empirical average $\hat{I}_t$ computed on the t correlated MCMC simulations.

Asymptotic variance of the empirical means

Under suitable regularity conditions

$$asy\mathbb{V}\left[\frac{1}{T}\sum_{t=1}^{T} X_t\right] = \lim_{T\to\infty} T \cdot \mathbb{V}\left[\frac{1}{T}\sum_{t=1}^{T} X_t\right] = \mathbb{V}_\pi[X] \sum_{t=-\infty}^{\infty} \rho_t$$

with $\rho_t = Cor_\pi(X_0, X_t)$

Metropolis-Hastings Algorithm

Let π be the target density. Let $X_t = x$ be the current state of the chain. Let $p_x(y)$ a (conditional) density named (proposal distribution). $p_x(y)$ may depend on the value x of the current state of the chian. Use the following rule for the transition from the state at time t to the state at time $(t+1)$.

1. Draw a candidate $Y_{t+1} \sim p_x(y)$. Let us denote with y the realized candidate $Y_{t+1} = y$
2. Decide whether or not the candidate is accepted as the next state of the chain at time $t+1$ otherwise set the next state equal to the current state x of the chain

$$X_{t+1} = \begin{cases} y & \text{with probability } \alpha(x,y) \\ x & \text{with probability } 1 - \alpha(x,y) \end{cases}$$

where

$$\alpha(x,y) = \min\left\{ \frac{\pi(y)}{\pi(x)} \frac{p_y(x)}{p_x(y)}, 1 \right\}$$

Metropolis-Hastings Algorithm

$$X_t = x \Longrightarrow$$

$$\Longrightarrow \quad Y_{t+1}|X_t = x \sim p_x(\cdot) \Longrightarrow Y_{t+1} = y$$

$$\Longrightarrow \quad E_{t+1}|Y_{t+1} = y, X_t = x \sim Bernoulli(\alpha(x, y))$$

$$E_{t+1} = e_{t+1} = \begin{cases} 1 = ACCEPT \\ 0 = REJECT \end{cases}$$

$$\Longrightarrow \quad X_{t+1}|E_{t+1} = e_{t+1}, Y_{t+1} = y, X_t = x = \begin{cases} y & \text{if } e_{t+1} = 1 = ACCEPT \\ x & \text{if } e_{t+1} = 0 = REJECT \end{cases}$$

Let us highlight similarities and differences between the MH algorithm in the MCMC perspective and the A/R simulation scheme we have seen in the MC world:

- MH $\Longrightarrow$ stochastich process of $X_1, ..., X_t, ...$ with Markovian dependence
- proposal distribution need not be "dominating" the target $\pi(\cdot)$ up to a known proportionality constant
- we only need the target computable up to a proportionality constant
- at each time t I do have a "new" current simulation but it may be not the proposed one

- When the proposal enjoys the following symmetry $p_y(x) = p_x(y)$ then we get the following simplified acceptance probability
$$\alpha(x, y) = \min\left\{\frac{\pi(y)}{\pi(x)}, 1\right\}$$
(was already in Metropolis et al., 1953)
- We assume that we start the chain at x with $\pi(x) > 0$. By convention we consider null the ratio $\frac{\pi(y)}{\pi(x)}$ whenever both numerator and denominator were zero
- again, for the time being, we will not study the ergodic properties of the MH Markov Chain but we limit ourselves to determine the transition kernel $K(\cdot, \cdot)$ and verify that the target $\pi(\cdot)$ is invariant with respect to the kernel

We start from considering the fact that there is a non null (positive) probability that the chain which is in x at the current time t will be still in x in the next time $(t+1)$. We can compute the probability of the complementary event:

$$\begin{aligned} a(x) &= Pr\left\{E_{t+1} = 1 | X_t = x\right\} = \\ &= \int_{\mathcal{S}} Pr\left\{E_{t+1} = 1, Y_{t+1} = y | X_t = x\right\} dy = \\ &= \int_{\mathcal{S}} \alpha(x,y) p_x(y) dy \end{aligned}$$

i.e. the probability of accepting the proposal Y_{t+1} [which may take any value y] conditionally on the current state x of the chain at the present time t

$\Longrightarrow$

The distribution of moving from the current x at time t to the next state y at time $t+1$ will contain a **discrete** component at x corresponding to the mass $1 - a(x)$.

$$\begin{aligned}
K(x, A) &= Pr\left\{X_{t+1} \in A | X_t = x\right\} = \\
&= Pr\left\{E_{t+1} = 1, X_{t+1} \in A | X_t = x\right\} + Pr\left\{E_{t+1} = 0, X_{t+1} \in A | X_t = x\right\} =
\end{aligned}$$

$$\begin{aligned}
&= Pr\left\{E_{t+1} = 1, Y_{t+1} = X_{t+1}, X_{t+1} \in A | X_t = x\right\} + Pr\left\{E_{t+1} = 0, X_{t+1} \in A | X_t = x\right\} = \\
&= Pr\left\{E_{t+1} = 1, Y_{t+1} = X_{t+1}, Y_{t+1} \in A | X_t = x\right\} + Pr\left\{E_{t+1} = 0, X_{t+1} \in A | X_t = x\right\} = \\
&= Pr\left\{Y_{t+1} \in A, E_{t+1} = 1 | X_t = x\right\} + \\
&\quad + Pr\left\{E_{t+1} = 0 | X_t = x\right\} Pr\left\{X_t \in A | E_{t+1} = 0, X_t = x\right\} = \\
&= \int_A \alpha(x, y) p_x(y) dy + (1 - a(x)) \delta_{\{x\}}(A)
\end{aligned}$$

The joint density for

$$j(Y_{t+1} = y, E_{t+1} = 1 | X_t = x)$$

can be computed (factorized) as the conditional Bernoulli probability of E_{t+1} given y by the marginal density of $Y_{t+1} = y$ (both conditionally on $X_t = x$)

$$j(E_{t+1} = 1 | Y_{t+1} = y, X_t = x) \cdot j(Y_{t+1} = y | X_t = x)$$

I need to integrate this joint density over the subset $A \times \{1\}$

I have used the following argument:

$$E_{t+1} = 1 = ACCEPT \qquad \subseteq \qquad X_{t+1} = Y_{t+1}$$

hence the intersection event

$$\{E_{t+1} = 1 \cap X_{t+1} = Y_{t+1}\} = \{E_{t+1} = 1\}$$

hence

$$\begin{aligned}\{E_{t+1} = 1 \cap X_{t+1} \in A\} &= \{E_{t+1} = 1 \cap X_{t+1} \in A \cap X_{t+1} = Y_{t+1}\} \\ &= \{E_{t+1} = 1 \cap Y_{t+1} \in A \cap X_{t+1} = Y_{t+1}\} =\end{aligned}$$

$$= \{Y_{t+1} \in A \cap E_{t+1} = 1\}$$

I can also rewrite $K(x, A)$ as follows:

$$
\begin{aligned}
K(x, A) &= \int_A \alpha(x, y) p_x(y) dy + (1 - a(x)) \delta_{\{x\}}(A) \\
&= a(x) \int_A \frac{\alpha(x, y) p_x(y)}{a(x)} dy + (1 - a(x)) \delta_{\{x\}}(A)
\end{aligned}
$$

where the ratio $\frac{\alpha(x,y) p_x(y)}{a(x)}$ represents the conditional density of X_{t+1} conditionally on $E_{t+1} = 1 = ACCEPT$

We will denote the transition with

$$q_{MH}(x,y) = \alpha(x,y)p_x(y)(1-\delta_{\{x\}}(y)) + (1-a(x))\delta_{\{x\}}(y)$$

the transition density corresponding to the (mixed-type) kernel $K(x,A)$

Notice that the dominating measure "$\underline{dy}$"(!) with respect to which the transition kernel (as a conditional distribution) is absolutely continuous is an unusual measure $\underline{\nu_x(dy)}$ which contains a discrete and an absolutely continuous (w.r.t. the usual Lebesgue measure) component!

$$K(x,\cdot) \ll \nu(\cdot)$$

the dominating measure is not the usual $\lambda(\cdot)$ (Lebesgue measure) but a suitable ν_x defined as follows

$$\nu(\cdot) = \lambda(\cdot) + \delta_{\{x\}}(\cdot)$$

which is obtained summing up the Lebesgue measure with a measure which assigns unit mass at a single point x

$$\delta_{\{x\}}(A) = \begin{cases} 1 & \text{se } x \in A \\ 0 & \text{se } x \notin A \end{cases}$$

Such a measure $\delta_{\{x\}}(\cdot)$ is usually referred to as Dirac measure at x.

We now introduce a notion and a result which is indeed slightly more than we need to verify that our target π is invariant with respect to the MH Kernel

Detailed Balance Condition - (DBC) - A transition kernel $K(\cdot,\cdot)$ and its corresponding transition density $q(x,y)$ is said to satisfy the detailed balance condition (DBC) with π if the following holds

$$q(x,y)\pi(x) = q(y,x)\pi(y) \qquad\qquad \forall (x,y) \in \mathcal{S}\times\mathcal{S}$$

Theorem - If a homogeneous Markov Chain corresponding to the kernel $K(\cdot,\cdot) \longleftrightarrow q(x,y)$ satisfies the DBC condition with π then

1. π is invariant for $K(\cdot,\cdot)$
2. the chain is reversible with respect to π

Reversible Markov Chain

A Markov Chain is said to be reversible with respect to π when

$$\overrightarrow{Pr}_{\pi}\left\{X_t \in A, X_{t+1} \in B\right\} = \overleftarrow{Pr}_{\pi}\left\{X_{t+1} \in A, X_t \in B\right\}$$

or,equivalently, using the corresponding densities

$$\pi(x)q(x,y) = \pi(y)q(y,x)$$

Backward transition

If we consider the time elapsing in the oppposite direction how should we write the backward transition density

$$\tilde{q}(x,y) = ?$$

N.B. $\pi(x)q(x,y)$ is a joint density with respect to the (corresponding/suitable) dominating $\nu_x(dy)\lambda(dx)$ while $\pi(y)q(y,x)$ is a joint density with respect to the (corresponding/suitable) dominating $\nu_y(dx)\lambda(dy)$.

Backward transition

If we consider the time elapsing in the oppposite direction how should we write the backward transition density

$$\tilde{q}(x,y) = \frac{\pi(y)q(y,x)}{\pi(x)}$$

The numerator correspond to the joint density of the stationary Markov chain with invariant $\pi(\cdot)$ where y corresponds to the present time t and x corresponds to time $t+1$. The backward transition density $\tilde{q}(x,y)$ corresponds to the conditional density of y (at time t) given x (at time $t+1$).

DBC condition for MH

for the first term in $q_{MH}(x,y)$, we have trivial identity for $x = y$.

For $x \neq y$ we can argue the case where the ratio in the definition of $\alpha(x,y)$ is less than 1 (which means that, symmetrically) the ratio in the definition of $\alpha(y,x)$ is greater than 1 (hence $\alpha(y,x) = 1$)

$$\frac{\pi(y)p_y(x)}{\pi(x)p_x(y)} < 1$$

DBC condition for MH

first term: case $\frac{\pi(y)p_y(x)}{\pi(x)p_x(y)} < 1$ hence $\alpha(y,x) = 1$

$$
\begin{aligned}
&\pi(x)\left[\alpha(x,y)p_x(y)(1-\delta_{\{x\}}(y))\right] + 0 \\
&\pi(x)\frac{\pi(y)p_y(x)}{\pi(x)p_x(y)}p_x(y) + 0 \\
&\pi(y)\cdot 1 \cdot p_y(x) + 0 \\
&\pi(y)\alpha(y,x)p_y(x) + 0 \\
&\pi(y)\left[\alpha(y,x)p_y(x)(1-\delta_{\{y\}}(x))\right] + 0
\end{aligned}
$$

DBC condition for MH

Moreover, for the second term in $q(x, y)$, we have

$$(1 - a(x))\delta_{\{x\}}(y)\pi(x) = (1 - a(y))\delta_{\{y\}}(x)\pi(y)$$

holding in both cases where $x = y$ and $x \neq y$.

...

- In theory there is a huge flexibility in the choice of $p_x(y)$
- We only need to be able to compute π up to a proportionality constant (since the constant cancels up in the acceptance probability ratio)
- What matters in the choice of $p_x(y)$ is how frequent is the event of rejecting the proposal in the Markov chain simulation since it tends to produce long subsequences of equal values which indeed correspond to a slow reweighting mechanism $\leftrightarrow$ it affects the speed of convergence
- Sometimes it may be difficult to calibrate suitable proposals

Some particular cases (1)

Random Walk Metropolis-Hastings

We can consider the following (<u>proposal</u>) density $p_x(y)$ for implementing the MH algorithm

$$p_x(y) = s(y - x)$$

where $s(\cdot)$ is a fixed density on $\mathbb{R}$. If $s(\cdot)$ symmetric about 0 → the implementation of the algorithm and the corresponding acceptance probability $\alpha(x, y)$ is simplified as follows

We propose $Y_t = X_t + U$ where $U \sim s(\cdot)$ (hence the RW terminology)

$$p_x(y) = s(y - x) = s(x - y) = p_y(x)$$

and

$$\alpha(x, y) = \min\left\{\frac{\pi(y)}{\pi(x)}, 1\right\}$$

Some particular cases (1)

Random Walk Metropolis-Hastings

Indeed, the most common implementation of RW-MH is through a random walk with symmetric random steps (innovations).

Some particular cases (2)

Independent proposal

$$p_x(y) = h(y)$$

h it is $\underline{\text{very}}$ similar to the spirit of the A/R scheme whith some relevant differences

$$\alpha(x,y) = \min\left\{1, \frac{\pi(y)}{\pi(x)}\frac{p_y(x)}{p_x(y)} = \frac{\pi(y)}{\pi(x)}\frac{h(x)}{h(y)} = \frac{\frac{\pi(y)}{h(y)}}{\frac{\pi(x)}{h(x)}} = \frac{w(y)}{w(x)}\right\}$$

Notice that:

- $w(\cdot)$ is close to 1 when $h(x) \approx \pi(x)$
- the proposal is independent from the past but .... the acceptance probability is not ....

hybrid Algorithms

Let X_t a Markov chain with kernel $K(\cdot,\cdot)$. Let us admit that π invariant for $K(\cdot,\cdot)$

$$\pi(A) = \int_{\mathcal{S}} K(x,A)\pi(dx) \qquad \forall A \in \sigma(\mathcal{S})$$

In the finite state space case we could write $\pi = \pi K$ (with π denoting a row vector) [or, equivalently, $\pi = t(K)\pi$ with the usual column vector notation.]

Let Y_t be another Markov chain with a different kernel H. Let us admit that π invariant for H as well

$$\pi(A) = \int_{\mathcal{S}} H(x,A)\pi(dx) \qquad \forall A \in \sigma(\mathcal{S})$$

Again in the finite state case this corresponds to $\pi^T = \pi^T H$.

hybrid Algorithms

Now with both kernels K and H we can define (*) a new kernel $N = HK$ called composition kernel or product kernel namely

$$N(x, A) = \int_{\mathcal{S}} K(y, A) H(x, dy)$$

We can then prove that the Markov chain Z_t defined by means of $N = HK$ has the same π as invariant distribution

Gibbs Sampling as a hybrid kernel

Each full-conditional step can be regarded as a single kernel and the GS is the composition of k kernels:

$$\begin{array}{lllllll}
\theta_1^{t+1} & \sim & \pi(\theta_1|\theta_{(1)}) & = & \pi(\theta_1|\theta_2^t, ...\theta_k^t) & \rightarrow & K_1 \\
\theta_2^{t+1} & \sim & \pi(\theta_2|\theta_{(2)}) & = & \pi(\theta_2|\theta_1^{t+1}, \theta_3^t...\theta_k^t) & \rightarrow & K_2 \\
.. & .. & .. & .. & .. & & \\
\theta_k^{t+1} & \sim & \pi(\theta_k|\theta_{(k)}) & = & \pi(\theta_k|\theta_1^{t+1}, \theta_2^{t+1}, ..., \theta_{k-1}^{t+1}) & \rightarrow & K_k
\end{array}$$

If we consider a measurable subset $A \subseteq \mathbb{R}^k$ as $A = A_1 \times A_2 \times ... \times A_{i-1} \times A_i \times A_{i+1} \times ... \times A_k = A_{(i)} \times A_i$ and we understand that $x \leftrightarrow (x_{(i)}, x_i)$ each kernel is given by K_i

$$K_i(x, A) = \delta_{x_{(i)}}(A_{(i)}) \int_{A_i} \pi(y_i|x_{(i)})dy_i$$

hence it is easy to see that

$$\begin{aligned}
\pi(A) &= \int_{\mathbb{R}^k} K_i(x, A)\pi(x)dx \\
&= \int_{\mathbb{R}^k} \left[\delta_{x_{(i)}}(A_{(i)}) \int_{A_i} \pi(y_i|x_{(i)})dy_i\right]\pi(x)dx \\
&= \int_{\mathbb{R}^{k-1}} I_{A_{(i)}}(x_{(i)}) \int_{\mathbb{R}} \left[\int_{A_i} \pi(y_i|x_{(i)})dy_i\right] \pi(x_i|x_{(i)})dx_i \pi(x_{(i)})dx_{(i)} \\
&= \int_{A_{(i)} \times A_i} \pi(y_i|x_{(i)})\; dy_i\; \pi(x_{(i)})dx_{(i)} = \int_A \pi(x)dx
\end{aligned}$$

The whole GS transition, rewritten as follows: $K = K_1 K_2 ... K_k$, can now be easily argued to have $\pi(\cdot)$ as invariant distribution.

Remind that we are considering a measurable subset $A \subseteq \mathbb{R}^k$ which is a cartesian product and it is represented as

$$A = A_1 \times A_2 \times \ldots \times A_{i-1} \times A_i \times A_{i+1} \times \ldots \times A_k = A_{(i)} \times A_i$$

Moreover remind that

$$\int_{\mathbb{R}} \pi(x_i | x_{(i)}) dx_i = 1 \qquad\qquad \forall x_{(i)} \in \mathbb{R}^{k-1}$$

Gibbs Sampling as a Metropolis-Hastings

If we look at each single K_i, we can write (think of) the proposal density as follows

$$p_x(y) = \delta_{x_{(i)}}(y_{(i)})\pi(y_i|x_{(i)}) = \delta_{x_{(i)}}(y_{(i)})\pi(y_i|y_{(i)})$$

Hence

$$\begin{aligned}\alpha(x,y) &= \min\left\{1, \frac{\pi(y)}{\pi(x)}\frac{p_y(x)}{p_x(y)}\right\} \\ &= \min\left\{1, \frac{\pi(y_i|y_{(i)})\pi(y_{(i)})}{\pi(x_i|x_{(i)})\pi(x_{(i)})}\frac{\delta_{y_{(i)}}(x_{(i)})\pi(x_i|x_{(i)})}{\delta_{x_{(i)}}(y_{(i)})\pi(y_i|y_{(i)})}\right\} \\ &= 1\end{aligned}$$

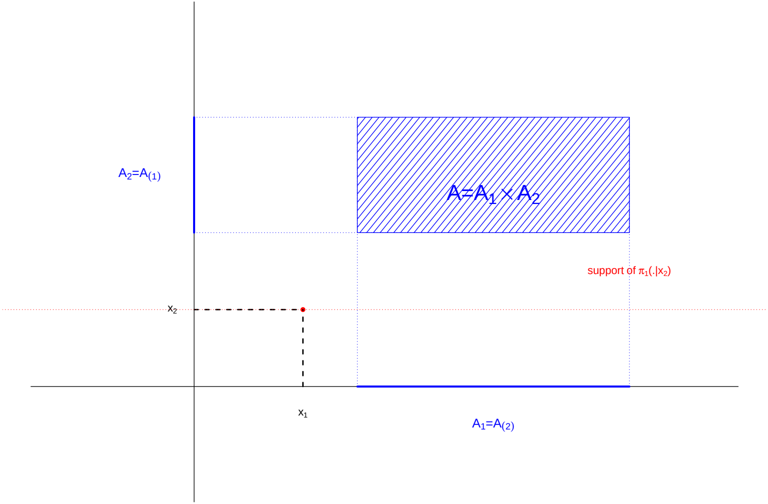
N.B. actually Gibbs Sampling could be regarded as a composition of Metropolis-Hastings kernels

Metropolis-within-Gibbs or Gibbs-Within-Metropolis

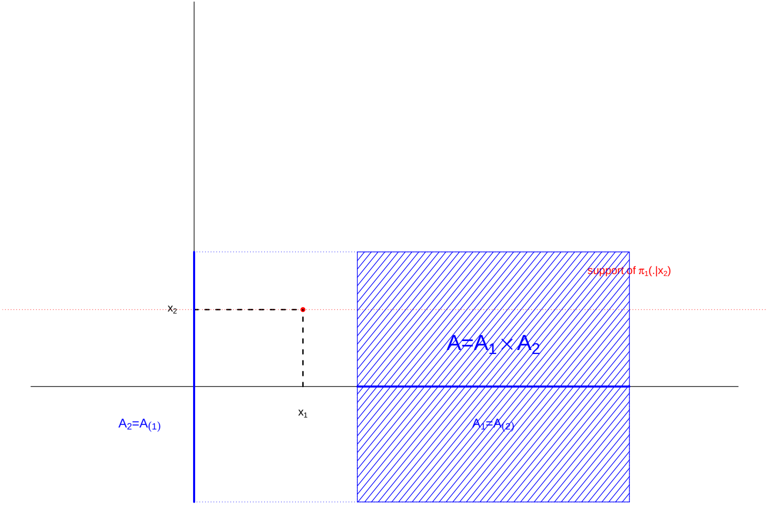
Now that we can think of Gibbs Sampling as a hybrid of k kernels $K = K_1 K_2 ... K_k$ where the i-th kernel $K_i(x, A) = K_i^{FC_i}(x, A)$ [actually the i-th full-conditional (FC_i) $\pi(x_i | x_{(i)})$] with invariant π it is immediate to realize that we can keep the same invariant distribution π if we replace $K_i(x, A)$ with another kernel $\tilde{K}_i^{MET_i}(x, A)$ related to the i-th coordinate so that $\pi_i(A_i | x_{(i)}) \delta_{x_{(i)}}(A_{(i)})$ is the invariant target for $\tilde{K}_i^{MET_i}(x, A)$. In that case we can draw the i-th coordinate from a proposal $p_{x_i}(y_i)$ and then accept it with probability

$$\alpha(x_i, y_i) = \min\left\{1, \frac{\pi(y_i | x_{(i)})}{\pi(x_i | x_{(i)})} \frac{p_{y_i}(x_i)}{p_{x_i}(y_i)}\right\} = \min\left\{1, \frac{\frac{\pi(y_i | x_{(i)})}{p_{x_i}(y_i)}}{\frac{\pi(x_i | x_{(i)})}{p_{y_i}(x_i)}}\right\}$$

Visual illustration of the features of $K_1(x, A)$



Visual illustration of the features of $K_1(x, A)$



Metropolis-within-Gibbs or Gibbs-Within-Metropolis

More formally we have already seen that

$$\begin{aligned}
\pi(A) &= \int_{\mathbb{R}^k} K_i(x, A)\pi(x)dx \\
&= \int_{\mathbb{R}^k} \left[I_{A_{(i)}}(x_{(i)}) \int_{A_i} \pi(y_i|x_{(i)})dy_i \right] \pi(x)dx \\
&= \int_{\mathbb{R}^{k-1}} I_{A_{(i)}}(x_{(i)}) \int_{\mathbb{R}} \left[\int_{A_i} \pi(y_i|x_{(i)})dy_i \right] \pi(x_i|x_{(i)})dx_i \pi(x_{(i)})dx_{(i)} \\
&= \int_{A_{(i)}} \int_{A_i} \pi(y_i|x_{(i)})\ dy_i\ \pi(x_{(i)})dx_{(i)} \\
&= \int_A \pi(x)dx
\end{aligned}$$

Now, let us focus only on the updating of the i-th component and assume that, instead of using the full-conditional $\pi(y_i|x_{(i)})$, or, equivalently, $K_{FC}(x_i, A_i) = \int_{A_i} \pi(y_i|x_{(i)})dy_i$, we are using a Metropolis updating $\tilde{K}_i^{MET}(x_i, A)$ that has $\pi(y_i|x_{(i)})$ as its invariant distribution so that

$$\int_{\mathcal{S}} \tilde{K}_i^{MET}(x_i, A_i)\pi(x_i|x_{(i)})dx_i = \int_{A_i} \pi(y_i|x_{(i)})dy_i$$

If $\mathcal{S} = \mathbb{R}$ we can write

$$\int_{\mathbb{R}} \tilde{K}_i^{MET}(x_i, A_i)\pi(x_i|x_{(i)})dx_i = \int_{A_i} \pi(y_i|x_{(i)})dy_i$$

We can then embed this updating for all the components similarly to the Gibss sampling as follows

$$\tilde{K}^{MET}(x, A) = I_{A_{(i)}}(x_{(i)})K_i^{MET}(x_i, A_i)$$

and we can then repeat the same arguments as in the previous proof for the GS intermediate kernel/step K_i

Metropolis-within-Gibbs or Gibbs-Within-Metropolis

$$
\begin{aligned}
\pi(A) &= \int_{\mathbb{R}^k} \tilde{K}^{MET}(x, A)\pi(x)dx \\
&= \int_{\mathbb{R}^k} \left[I_{A_{(i)}}(x_{(i)}) K_i^{MET}(x_i, A_i) \right] \pi(x)dx \\
&= \int_{\mathbb{R}^{k-1}} I_{A_{(i)}}(x_{(i)}) \int_{\mathbb{R}} K_i^{MET}(x_i, A_i)\pi(x_i|x_{(i)})dx_i \pi(x_{(i)})dx_{(i)} \\
&= \int_{A_{(i)}} \int_{A_i} \pi(y_i|x_{(i)})\ dy_i\ \pi(x_{(i)})dx_{(i)} \\
&= \int_A \pi(x)dx
\end{aligned}
$$

Dugong example

On 27 seacows (dugongs) have been reported: length (Y_i) [meters] and age (x_i) [years]. We are interested in the following non-linear regression

$$\begin{aligned} Y_i &\sim N(\mu_i, \tau^2) \\ \mu_i = f(x_i) &= \alpha - \beta\gamma^{x_i} \end{aligned}$$

Length:	1.80	1.85	1.87	...	2.70	2.72	2.57
Age:	1.00	1.50	1.50	...	22.50	29.00	31.50

* D.A. Ratkowsky. Nonlinear Regression Modeling. Dekker, 1983.

Four model parameters:
$\alpha \in (1, \infty)$ (!)
$\beta \in (1, \infty)$ (!)
$\gamma \in (0, 1)$
$\tau^2 \in (0, \infty)$
A priori

$$\begin{aligned} \alpha &\sim N(0, 10000)(!) \\ \beta &\sim N(0, 10000)(!) \\ \gamma &\sim Unif(0, 1) \\ \tau^2 &\sim I\Gamma(0.001, 0.001) \end{aligned}$$

Gibbs Sampling troubles

There are circumstances where the parameterization of the target distribution and the corresponding related full-conditionals do **not** allow a fast exploration of the main bulk of the support of the target (stationary region):

1. strong dependence (bivariate normal example)
2. multimodality

Gibbs sampling troubles

Possible remedies:

1. reparameterizing is an option for the former but not for the latter
2. combining the current kernel with a Metropolis step: $K^* = K_1 \circ ... \circ K_p$ (composition)
3. combining the current kernel with a Metropolis step: $\tilde{K} = w_1 K_1 + ... + w_p K_p$ (mixing)

Gibbs Sampling variants

- Gibbs sampling with systematic order of full-conditionals [non-reversibile Markov chain]
- Gibbs sampling with symmetric (double) scan of full-conditionals [reversibile Markov chain]
- Gibbs sampling with random (symmetric, exchangeable) order of full-conditionals [reversibile Markov chain]

Ergodic theorems

There are many results which can be considered part of the ergodic theory of Markov chains. We will be mainly interested in the so-called *h*-ergodicity i.e. under which conditions we can guarantee that

$$\frac{1}{t}\sum_{i=1}^{t} h(X_i) \overset{t\to\infty}{\longrightarrow} I$$

We will not have time to cover this part of the MCMC theory. We limit ourselves to list the main results and the properties involved. The interested students can look at the book by Robert & Casella.

Definitions and ergodic theorems

Outline:

We will need the following definitions

- Irreducibility (ϕ)
- Periodicity and Aperiodiciy
- Harris recurrence

to show

- Unicity of the limiting distribution
- Convergence of empirical average
- Rate of convergence and uniform ergodicity
- Central Limit Theorem for Markov chains
- Variance of empirical average and asymptotic variance
- Consistent estimation of the variance of the empirical average

WINBUGS/JAGS/R2jags

- `http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtm`
- `https://sourceforge.net/projects/mcmc-jags/files/latest`
- Download → Install
- Documentation
- Direct interactive use
- Indirect use through `R`: `R2WinBUGS` package or `rjags` package or `R2ags` package
- Output in R and convergence monitoring tools

JAGS and the like

Pros

+ Automatic MCMC implementation
+ Flexibility

Cons

- Computing time
- Requires some expertise
  - critical inputs (initial values, a-priori, convergence monitoring)
  - reparameterization, ordering, blocking

Fundamental steps

1. Model specification (Model)
2. Observed data (and other fixed quantities) input (Load Data)
3. Model compiling (Compile)
4. Initial values (Simulated or fixed in advance [input])
5. Updating → Simulations → Output
6. Visual inspection of simulated coordinates/components and summary values of the simulated components

We can embed the power of WinBUGS within `R` using an add-on package named `R2WinBUGS` [Windows user]. Alternatively, a (unix-native, mac compatible) similar tool is `JAGS` and its R-port `rjags`.

Analyzing Dugongs data with WINBUGS

```
# Model specification for DUGONGS DATA
# to be recorded in a separate ASCII file named winbugs-model-dugong.tx
model
    {
        for( i in 1 : N ) {         # MODEL DESCRIPTION
            Y[i] ~ dnorm(mu[i], tau)
                            # N.B. tau is conceived as precision parame
            mu[i] <- alpha - beta * pow(gamma,x[i])                 }
                            # A PRIORI
 alpha ~ dnorm(0.0, 1.0E-3)I(1.0,)
 beta ~ dnorm(0.0, 1.0E-3)I(1.0,)
        gamma   ~ dunif(0.5, 1.0)
        tau     ~ dgamma(0.001, 0.001)

        sigma <- 1 / sqrt(tau)
        U3 <- logit(gamma)
    }
```

```
# OBSERVED DATA (e quantita' costanti) --> Data

list(x = c( 1.0,  1.5,  1.5,  1.5, 2.5,   4.0,  5.0,  5.0,  7.0,
                8.0,  8.5,  9.0,  9.5, 9.5,  10.0, 12.0, 12.0, 13.0,
               13.0, 14.5, 15.5, 15.5, 16.5, 17.0, 22.5, 29.0, 31.5),
         Y = c(1.80, 1.85, 1.87, 1.77, 2.02, 2.27, 2.15, 2.26, 2.47,
               2.19, 2.26, 2.40, 2.39, 2.41, 2.50, 2.32, 2.32, 2.43,
               2.47, 2.56, 2.65, 2.47, 2.64, 2.56, 2.70, 2.72, 2.57), N


# INITIAL VALUES  --> Inits

list(alpha = 1, beta = 1, tau = 1, gamma = 0.9)
```

DIC

DIC: Deviance Information Criterion

An alternative (Bayesian) criterion for model comparison

Deviance $D(\theta) = -2 \log L(\theta)$

(N.B. there can be additional costant offset due to proportionality constants in the likelihood evaluation)

$$DIC = D(\bar{\theta}) + 2 \cdot p_D$$

where $p_D = \bar{D} - D(\bar{\theta})$ is the effective number of parameters of the model. Alternatively, Gelman proposed to set $p_D = \frac{1}{2} Var[D(\theta)]$.

MCMC Output and convergence diagnostics

- Graphical diagnostics
- Formal tests

Graphical diagnostics

- traceplot of simulations of each component $\theta_{i,t}$ of the parameter vector (marginal)
- traceplot of running means $\hat{I}_t$
- crossing of overlapped traceplots of multiple chains started from different (and suitably scattered) starting points

Some of these diagnostics are already automatically output from some packages like `ggmcmc` (a single pdf file with histograms, density plots, traceplots, running means, comparison of partial and full chain, autocorrelation plots, crosscorrelation plot, caterpillar plot together with some formal diagnostics)

More interactive diagnostics through the `coda` packages (try `codamenu()`). It can use BUGS output files as well as `mcmc`-class objects

MCMC Output and convergence diagnostics

- Graphical diagnostics
- Formal tests

Graphical diagnostics

- traceplot of simulations of each component $\theta_{i,t}$ of the parameter vector (marginal)
- traceplot of running means $\hat{I}_t$
- crossing of overlapped traceplots of multiple chains started from different (and suitably scattered) starting points

Some of these diagnostics are already automatically output from some packages like `ggmcmc` (a single pdf file with histograms, density plots, traceplots, running means, comparison of partial and full chain, autocorrelation plots, crosscorrelation plot, caterpillar plot together with some formal diagnostics)

More interactive diagnostics through the coda packages (try `codamenu()`). It can use BUGS output files as well as `mcmc`-class objects

Fundamental features to monitor

- Has stationarity been achieved?
- Speed of exploration of target support
- Correlation
- Overall convergence of the empirical distribution of simulations to the target
- Convergence of each individual empirical mean
- Proximity to i.i.d. simulation

Convergence diagnostics

- In theory everything works ... but in practice? What kind of troubles lurks in the background?
  1. *non complete* exploration of the support of the target distribution
  2. approximation error accumulation
- Ideally we would like to have a stopping rule which decides when a prescribed level of precision is met. Unfortunately this is possible only for oversimplified problems and computations may be not straightforward

Diagnostic checks must be considered with a pessimistic attitude: their role is to highlight problems; they never provide absolute warranties

3 types of convergence monitoring

1. Global convergence of the empirical distribution at time t to the stationary (target) distribution

$$P^t_{\theta_0}(\cdot) \approx \pi(\cdot)$$

   1.1 Homogeneity test of two adjacent *subsamples* (Kolmogorov-Smirnov test)

   1.2 Unbiased estimators of the distance $d(P^t_{\theta_0}, \pi)$ (e.g. total variation distance)

2. Convergence of $\hat{I}_t$ to I

   2.1 CUSUM

   2.2 Multiple estimates

3. Proximity to i.i.d.

   3.1 *Batching* or *Sub-sampling*

   3.2 Renewal theory and the like (perfect sampling)

Single or multiple?

- Single chain
- Multiple (parallel) chains

Multimodality (?)

robustness of diagnostics with respect to the starting point

direct comparison among chains

non self-referential

the slowest chain might be taken as reference

difficulties in the choice of initial points (not too far though ...)

Gelman & Rubin

[Gelman & Rubin (1992)] Based on M simulated chains starting from suitably scattered initial values[*]. Denoting the m-th chain with $\theta_1^{(m)}, \theta_2^{(m)}, \ldots, \theta_t^{(m)}, \ldots, \theta_T^{(m)}$ and the corresponding estimator with $\hat{I}_T^{(m)} = \frac{1}{T}\sum_{i=1}^{T} h(\theta_i^{(m)})$ and with $\hat{I}_T$ the overall estimator which uses all the $M \times T$ simulations this diagnostics is based on

$$\begin{aligned} B_T &= \frac{1}{M}\sum_{m=1}^{M}\left(\hat{I}^{(m)} - \hat{I}_T\right)^2 \qquad (\underline{\text{Between}}) \\ W_T &= \frac{1}{M}\sum_{m=1}^{M}\left[\frac{1}{T}\sum_{i=1}^{T}\left(h(\theta_i^{(m)}) - \hat{I}_T^{(m)}\right)^2\right] \qquad (\underline{\text{Within}}) \end{aligned}$$

Gelman & Rubin (2)

It is a formal diagnostics based on monitoring the so-called Potential Scale Reduction Factor (PSRF)

$$R_T = \frac{\frac{T-1}{T} W_T + \frac{M+1}{M} B_T}{W_T} \cdot \frac{\nu_T}{\nu_T - 2}$$

with a suitable expression of $\nu_T = ...$ based on the asymptotic behaviour of the ratio of between and within variance as in the ANOVA analysis under normality assumptions

When $T \to \infty$ R_T is expected to decrease to 1 under the assumption of homogeneity of means (which may be different when strongly dependent on starting values if convergence is slow).

For more details see also here

Geweke

Geweke (1992) proposed a convergence diagnostic for Markov chains based on a test for equality of the means of the first and last part of a Markov chain (by default the first 10% and the last 50%).

If the samples are drawn from the stationary distribution of the chain, the two means are equal and Geweke's statistic has an asymptotically standard normal distribution.

Heidelberger & Welch

The convergence test uses the Cramer-von-Mises statistic to test the null hypothesis that the sampled values come from a stationary distribution.

- The test is successively applied, firstly to the whole chain, then after discarding the first 10%, 20%, of the chain until either the null hypothesis is accepted, or 50% of the chain has been discarded.
- The latter outcome constitutes "failure" of the stationarity test and indicates that a longer MCMC run is needed.
- If the stationarity test is passed, the number of iterations to keep and the number to discard (burn-in) are reported.

Raftery & Lewis

It is based on the "binarization" of the chain θ_t using the following transform

$$Z_t = I_{\theta \leq \bar{\theta}}(\theta_t)$$

where $\bar{\theta}$ is a value for which we could be interested in (e.g. a quantile of the posterior distribution of θ).

If Z_t were a binary Markov chain then there would be the possibility of explicit computations about bounding the distance between the empirical distribution (relative frequencies) and the stationary distribution π which puts mass $\frac{\alpha}{\alpha+\beta}$ on the outcome $Z = 1$ where $\alpha = P(Z_{t+1} = 1|Z_t = 0)$ and $\beta = P(Z_{t+1} = 0|Z_t = 1)$. We can then provide a lower bound on T such that

$$P\left(\left|\hat{I}_T - \frac{\alpha}{\alpha+\beta}\right| < \delta\right) \geq \epsilon$$

Indeed the limit depends on α e β which are unknown → (we could either estimate them or use an alternative definition of T_{min} based on independence)

Raftery & Lewis (2)

$$T_{min} \geq \Phi^{-1}\left(\frac{\epsilon+1}{2}\right)^2 \frac{\alpha\beta}{(\alpha+\beta)^2}\delta^{-1}$$

Some issues:

- Approxiamtion of α β
- unidimensional → does not take into account global convergece of all parameters
- problems when $\bar{\theta}$ is chosen in the tails

Raftery and Lewis

In practice, suppose we want to measure some posterior quantile of interest q.

If we define some acceptable tolerance r for q and a probability s of being within that tolerance, the Raftery and Lewis diagnostic will calculate the number of iterations $N = T_{min}$ and the number of burn-ins M necessary to satisfy the specified conditions.

The diagnostic was designed to test the number of iterations and burn-in needed by first running and testing shorter pilot chain. In practice, we can also just test our normal chain to see if it satisfies the results that the diagnostic suggests.