PRIOR ELICITATION

Alessandra Guglielmi

Politecnico di Milano
Dipartimento di Matematica
Milano, Italia
e-mail: alessandra.guglielmi@polimi.it

1 October 2020

Take-home messages:
- how can we construct for the natural exp. family the family of conjugate priors
- Jeffrey's prior (= non informative prior)

The choice of the prior

It is the most controversial aspect of the Bayesian approach.

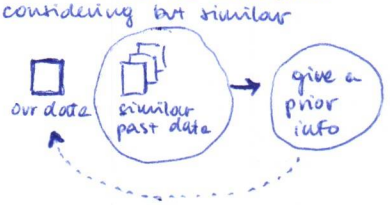
The statistician should choose a subjective prior according to one's prior knowledge and belief; some form of prior information is always available!

- if other similar data are available, under the exchangeability assumption, it is possible to use information on the empirical distribution of these other data when the size is large (e.g. empirical mean and variance of 0-1 r.v.'s) in terms of π (mean and variance of π); compare with the Representation's theorem)
- when θ has an objective (i.e. "physical") meaning, that is, when θ represents a well-specified characteristic of the members of a statistical population, we could obtain information on the parameter to be translated into prior moments or quantiles from an expert

This came from de Finetti's representation theorem for events:

$\frac{\sum_{i=1}^{n} X_i}{n} \to \tilde{\theta} \sim F$

these are not the data we're considering but similar



The choice of the prior

- if prior information is available on the marginal distribution of X_1, and the likelihood has already been chosen, this information can be transformed into values of moments and quantiles of π

In general, available information is not so precise to generate one single prior for θ

Our goal is NOT to find the *perfect* prior distribution, but rather to incorporate salient features of available scientific knowledge into the data analysis through the prior → many priors consistent with available information

We must select one prior among the plausible distributions, or rather approximate the *true* prior

Prior Elicitation

Which prior shall I pick out? Today I am going to tell you about two possible choices:

- a mixture of conjugate distributions
- a non-informative prior (when there is no information or *vague* information)

In real problems (with hundreds of parameters, or more), we typically use a third alternative:

- we assume a prior that gives full-conditionals easy to generate from

Mixtures of conjugate or non-informative priors

Mixtures of conjugate priors: hierarchical models, i.e. the prior is given in a 2-level hierarchy, meaning that a hyper-prior distribution is given to the hyperparameters of the (conjugate) prior; in addition, a robustness analysis is required w.r.t. the last level hyperparameters varying in a *reasonable* set

since we're receiving informations gradually we have to verify what would change to receive these informations in other order (hoping that inference does not depend too much on the order)

In the total or partial absence of information,
NONINFORMATIVE priors (however being improper priors);
those are used mainly to automate the Bayesian Approach

CONIUGATE PRIORS

Informal def: a family $\mathcal{F}$ of distributions on Θ is conjugate (to the sampling model/distribution) if, for every $\pi \in \mathcal{F}$, the posterior distribution $\pi(\cdot|x) \in \mathcal{F}$.

Useful if the family $\mathcal{F}$ is parametric, since switching from prior to posterior is given by an update of the corresponding parameters.

However, for a subjectivist Bayesian, the use of conjugate priors is suspicious, since it is justified by their mathematical tractability rather then for fitting the available prior information.

Their role (of the conjugate prior) is justified since

- they provide a first approximation to the *true* prior (followed by a robustness analysis)
- the class of mixtures of conjugate priors is dense, in the Prohorov's metric, in $\mathbf{P}_{\mathbb{R}}$

the class of mixtures of conjugates priors is dense in the space of all probability measures over $\mathbb{R}$ ($\Rightarrow$ any distribution in $\mathbb{R}$ can be approximate by a finite mixture of conjugate distributions (under some conditions))

Exponential family - Robert (2007)

Definition

Let $\mathcal{X}$ be the sample space of observation X, Θ the parametric space; μ a σ-finite measure on $\mathcal{X}$. Let

$$C : \Theta \to \mathbb{R}^+ \quad h : \mathcal{X} \to \mathbb{R}^+ \quad R : \Theta \to \mathbb{R}^k \quad T : \mathcal{X} \to \mathbb{R}^k.$$

The family of distributions with densities, w.r.t. μ, given by

$$f(x|\theta) = h(x)e^{R(\theta)\cdot T(x)}C(\theta) = h(x)e^{R(\theta)\cdot T(x)-\Psi(\theta)},$$
$$\text{with } C(\theta) := e^{-\Psi(\theta)}$$

is called an exponential family of dimension k.

for instance the lebesgue measure / counting measure

Exponential family

If $\Theta \subset \mathbb{R}^k$ and $\mathcal{X} \subset \mathbb{R}^k$, the natural reparameterization of the exponential family is:

$$f(x|\theta) = h(x)e^{\theta\cdot x-\Psi(\theta)}.$$

natural exponential family

REM: the support of an exponential family does NOT depend on the parameter θ.

example: $\sim \mathcal{U}([0,\theta])$

REM: If $X_1, \ldots, X_n$ are iid from an exponential family of dimension k, the random vector $(X_1, \ldots, X_n)$ has density belonging to an exponential family of dimension k.
In fact

$$(X_1, \ldots, X_n) \sim \prod_1^n h(x_i)e^{R(\theta)\cdot\sum_1^n T(x_i)-n\Psi(\theta)}$$

CONJUGATE DISTRIBUTIONS

Define the natural parameter space of an exponential family as

$$N = \{\theta \in \Theta : \int_{\mathcal{X}} h(x)e^{\theta \cdot x}\mu(dx) < +\infty\}.$$

Theorem

Let $f(x|\theta) = h(x)e^{\theta \cdot x - \Psi(\theta)}$ be a (natural) exponential family. A conjugate family for $f(x|\theta)$ is

$$\pi(\theta; \mu, \lambda) = K(\mu, \lambda)e^{\theta \cdot \mu - \lambda\Psi(\theta)}, \quad \lambda > 0, \frac{\mu}{\lambda} \in \dot{N}.$$

The posterior for θ, given $x \sim f(x|\theta)$, is

$$\pi(\theta; \mu + x, \lambda + 1).$$

CONJUGATE DISTRIBUTIONS

If the data $x_1, \ldots, x_n$ are an iid sample (condit. to θ) from $f(x|\theta)$, then the posterior of θ is

$$\pi(\theta; \mu + \sum_1^n x_i, \lambda + n).$$

CONJUGATE DISTRIBUTIONS

THEOREM. If $X_1, \ldots, X_n|\theta \overset{iid}{\sim} f(\cdot|\theta)$ natural exponential family and π is the conjugate prior, and $\xi(\theta)$ is the conditional mean of X_1, i.e. $\mathbb{E}(X_1|\theta) = \xi(\theta)$, then

$$\mathbb{E}_\pi[\xi(\theta)] = \frac{\mu}{\lambda}.$$

Therefore, since the posterior is $\pi(\theta; \mu + \sum_1^n x_i, \lambda + n)$, we have:

$$\mathbb{E}_\pi[\xi(\theta)|X_1, \ldots, X_n] = \frac{\mu + \sum_1^n X_i}{\lambda + n} = \frac{\lambda}{\lambda + n}\frac{\mu}{\lambda} + \frac{n}{\lambda + n}\bar{X}_n$$
$$= p_n\, \mathbb{E}_\pi[\xi(\theta)] + (1 - p_n)\bar{X}_n$$

convex combination of $\mathbb{E}_\pi[\xi(\theta)]$ and $\bar{X}_n$ (result that we already saw with Bernoulli and Gaussian)

where

$$p_n := \frac{\lambda}{\lambda + n}$$

Examples

CONJUGATE DISTRIBUTIONS: a comment

The Bayesian estimator (under the quadratic loss) of the parameter representing the conditional mean of data is a mixture between prior information (prior mean) and the frequentist estimator

When n is large, or when $\frac{1}{\lambda}$ is large, the Bayesian estimate will be almost equal to the frequentist estimate

MIXTURES of CONJUGATE DISTRIBUTIONS

THEOREM. For any fixed N, the family

$$\mathcal{F}_N = \{\sum_1^N w_i \pi(\theta; \mu_i, \lambda_i), \sum_1^N w_i = 1, w_i > 0, i = 1, \dots, N\}$$

is also a conjugate family to the natural exponential family. Moreover, if $\pi(\theta) = \sum_1^N w_i \pi(\theta; \mu_i, \lambda_i)$, then the posterior distribution $\pi(\theta|x)$ is

$$\pi(\theta|x) = \sum_1^N w_i' \pi(\theta; \mu_i + x, \lambda_i + 1)$$

$$\text{with } w_i' = \frac{w_i K(\mu_i, \lambda_i)/K(\mu_i + x, \lambda_i + 1)}{\sum_j w_j K(\mu_j, \lambda_j)/K(\mu_j + x, \lambda_j + 1)}.$$

MIXTURES of CONJUGATE DISTRIBUTIONS

THEOREM. If Θ is the natural parameter space for the exponential family $f(x|\theta)$ and π is a PRIOR on Θ, then, for any $\epsilon > 0$, there exist N and $\tilde{\pi}_N \in \mathcal{F}_N$ such that $d_P(\pi, \tilde{\pi}_N) < \epsilon$.

Whichever prior information is available (represented by the "true" prior π), it could be modeled through a mixture from $\mathcal{F}_N$ with N as small as possible.

ISSUES:

- how $\tilde{\pi}_N$ is really build?
- though we define $\tilde{\pi}_N \in \mathcal{F}_N$ such that $d_P(\pi, \tilde{\pi}_N) < \epsilon$, it is NOT guaranteed that

$$d_P(\pi(\cdot|x), \tilde{\pi}_N(\cdot|x))) < \epsilon.$$

ADVANTAGE: a mixture of conjugate distributions may represent the real combination of experts' prior beliefs.

NONINFORMATIVE PRIORS

If there is available (partial) information, ideally I should translate it according my subjective Bayesian belief.

Which prior should we adopt in case of no available information?

It is common to choose and use distributions called objective priors, or noninformative priors; indeed, it is difficult to give a definition of total absence of information! There is no prior which could represent *true* absence of information

NONINFORMATIVE PRIORS

Justification for their use:

- they yield posteriors which depend more on data than on the prior; e.g. the posterior does not depend on hyperparameters
- they should be interpreted as a device, as a conventional or reference prior, or as those priors yielding a minimal effect on the posterior, given the dataset; under those priors, posterior inference is *similar* to frequentist inference. They could be also understood as default prior, when there are no resources (time, money, etc) to accurately elicit the *true* prior

In this perspective, some noninformative priors could be more useful or more efficient than others, but we cannot state they are less informative than others.

Once that we have found that the conditional distr. for the data belongs to the natural exp. family and we found the conjugate prior we can consider the family of a finite mixture of conjugate priors. →

The family of finite mixtures of conjugate priors is dense in the space of all prior distributions for θ

(20 exam exercises)

Only important thing: the family of finite mixtures of conjugate priors are useful since they approximate any distribution.

they try to make the posterior not depending on the parameters

NONINFORMATIVE PRIORS

Why the uniform distribution for the absence of information? The uniform treat every outcome equally, so without infos like "this outcome is more probable than others" (= prior information) we assume the uniform distribution.

Historically, they were introduced as those priors representing absence of information in the discrete case

Laplace: discrete uniform probability (on a finite set) through the Principle of Insufficient Reason (all elementary events are equally probably unless there is a reason to believe differently)

Generalization to a continuous r.v. $\theta \in (0,1)$: $\mathcal{U}(0,1)$

When Θ is unbounded: $\pi(\theta) \propto c$, $c > 0 \Rightarrow \int_{\mathbb{R}} \pi(\theta) d\theta = +\infty$, that means we are using an improper distribution!

MAIN CRITICISM to improper noninformative priors

1. They are NOT probabilities:
   - however they have finitely additive counterparts (but useless in applications)
   - many scientists (and statisticians) use them, as long as the posterior is proper
2. it is NOT guaranteed they are invariant under reparameterization: if, for inst., there is no information on the proportion of male newborns on Oct 20, 2016, at Clinica Mangiagalli, I should neither have information on the odds parameter $p/(1-p)$. If $p \sim U(0,1)$ (density prop. to a constant), then

$$\rho = \frac{p}{1-p} \sim \pi(\rho) = \frac{1}{(1+\rho)^2}, \rho \in \mathbb{R}^+$$

which is NOT proportional to a constant on $\mathbb{R}^+$.

PRIORS INVARIANT under REPARAMETERIZATION

We want to find a criterion that translates the non informativeness which is preserved under reparametrization:

Let $\theta \in \Theta$ and π_Θ be a noninformative prior according to some criterion (e.g. Uniform prior, Principle of IR); let $\eta = g(\theta)$ (reparametrization of θ), g one-to-one, and let π_η be the noninformative prior for η according the same criterion. However, they are distributions, i.e. they should obey the *transformation rule*: if $\eta = g(\theta)$ then

$$\pi_\eta(\eta) = \pi_\theta(g^{-1}(\eta)) \left| \frac{d}{d\eta} g^{-1}(\eta) \right|$$

This is NOT true in general; for instance, if $\pi_\theta(\theta) = 1$, $\theta \in \Theta \Rightarrow$

$$\pi_\eta(\eta) = \left| \frac{d}{d\eta} g^{-1}(\eta) \right| \neq const$$

(as it is true for π_θ) when g is general.

JEFFREYS PRIOR for a real parameter

example.

a prior which somehow translates the absence of information

It is a noninformative prior, based on Fisher information $\mathcal{I}(\theta)$

$$\pi^*(\theta) \propto \sqrt{\mathcal{I}(\theta)}, \quad \theta \in \Theta \subset \mathbb{R},$$

where

$$\mathcal{I}(\theta) = \mathbb{E}\left[\left(\frac{\partial}{\partial \theta} \log f(X|\theta) \right)^2 \right]$$

Jeffrey's prior for a real parameter is a prior proportional to the Fisher information

Examples

JEFFREYS PRIOR for a multidimensional parameter

Let $\theta \in \Theta \subset \mathbb{R}^p$: $\mathcal{I}(\theta)$ Fisher information matrix $p \times p$

Jeffreys prior: $\pi^*(\theta) \propto \sqrt{det(\mathcal{I}(\theta))}$

Under regularity condition of the (conditional) density

$$(\mathcal{I}(\theta))_{ij} = -\mathbb{E}_\theta \left[\frac{\partial^2}{\partial\theta_i \partial\theta_j} \log f(X|\theta) \right]$$

Example

BEWARE of IMPROPER PRIORS

priors s.t. the integral over all the space is not finite

We will NEVER use improper priors in applications!!!

why some bayesians do? Because as long as the posterior is proper, it'll be independent of hyperparameters, therefore the posterior let the data speaks for themselves.

When information is vague, we will use vague priors, i.e. user-friendly priors with large variance!

prior with large variance → a posteriori the bayesian estimate gives more weights to the data than to the prior information

MERGING of the PRIORS

Two different priors, elicited by 2 different experts who may have distinct subjective prior belief: two distinct posterior distributions.

Which one is the best ??

If the sample size n is large, posteriors will merge: the total variation distance between the two probabilities converges to 0 as $n \to +\infty$

Reason: consistency of the posterior distribution; if θ_0 denotes the *true* value of θ, then the posterior will concentrate more and more mass in a neighborhood of θ_0 as n increases to $+\infty$.

the posteriors eventually will merge because somehow they both go with the true θ

Asymptotic Normality of Posterior Distribution

THEOREM. Let $X_1, \ldots, X_n|\theta \overset{\text{i.i.d.}}{\sim} f(\cdot|\theta)$; $\pi(\theta)$ prior density for $\theta \in \Theta \subset \mathbb{R}^p$. Let $\pi(\theta|\boldsymbol{X}_n)$ be the posterior, obtained by Bayes Theorem. Let

$\widetilde{\theta}_n =$ posterior mean, $V_n =$ posterior covariance matrix

$\tilde{\theta}_n =$ posterior mode, $\tilde{I}_n = \left[-\frac{\partial^2}{\partial\theta_i \partial\theta_j} log\pi(\theta|\boldsymbol{X}_n)|_{\tilde{\theta}_n} \right]_{ij}$

generalized observed Fisher information matrix

$\widehat{\theta}_n =$ MLE, $\widehat{I}_n = \left[-\frac{\partial^2}{\partial\theta_i \partial\theta_j} \log f(\boldsymbol{X}_n|\theta)|_{\hat{\theta}_n} \right]_{ij}$

observed Fisher information matrix

$$I_n(\theta) = E_\theta \left[-\frac{\partial^2}{\partial\theta_i \partial\theta_j} \log f(\boldsymbol{X}_n|\theta) \right]_{ij}$$

(expected) Fisher information matrix

Asymptotic Normality - Statement of the Theorem

Then, under suitable regularity conditions, for *large n*, the posterior distribution $\pi(\theta|\boldsymbol{X}_n)$ can be approximated by any of the following distributions:

1. $\mathcal{N}_p(\tilde{\theta}_n, V_n)$ (posterior mean and variance)
2. $\mathcal{N}_p(\tilde{\theta}_n, (\tilde{I}_n)^{-1})$ (posterior mode and generalized observed Fisher information matrix)
3. $\mathcal{N}_p(\hat{\theta}_n, (\hat{I}_n)^{-1})$ (MLE and observed Fisher information)
4. $\mathcal{N}_p(\hat{\theta}_n, (I_n(\hat{\theta}_n))^{-1})$ (MLE and Fisher information m. at MLE).

procedure:
- $f(x|\text{param}) = \dots$ ← exp. family
- $f(x|\text{param}) \to f(x|\theta)$
- $f(x|\theta) \to$ thm: $\pi(\theta)$
- $\pi(\theta) \to \Pi(p)$
- $\pi(p) \to$ kernel of the prior
- kernel → conjugate prior

* Examples of exponential family:

Ex. $X|p \sim Be(p)$ $\quad p \in (0,1)$

$$f(x|p) = p^x (1-p)^{1-x} \qquad x \in \{0,1\}$$
$$= e^{x \log(p) + (1-x)\log(1-p)}$$
$$= e^{x \log\left(\frac{p}{1-p}\right) - (-\log(1-p))}$$

$$\Rightarrow \vartheta = \log\left(\frac{p}{1-p}\right) \in \mathbb{R} = \Theta \quad \Rightarrow \quad p = \frac{e^{\vartheta}}{1+e^{\vartheta}} \quad \Rightarrow \quad 1-p = \frac{1}{1+e^{\vartheta}}$$

$$\psi(\vartheta) = -\log(1-p)$$
$$= +\log(1+e^{\vartheta})$$

$$f(x|p) = e^{x\vartheta}\, e^{-\log(1+e^{\vartheta})}$$

Conjugate prior for θ: $\quad \pi(\vartheta) \propto e^{\vartheta\mu - \lambda \log(1+e^{\vartheta})} \qquad \lambda > 0$

$$\propto e^{\mu(\log(\frac{p}{1-p})) + \lambda \log(1-p)}$$
$$\propto e^{\mu \log(p) + (\lambda-\mu)\log(1-p)}$$
$$\propto \underbrace{p^{\mu}(1-p)^{(\lambda-\mu)}\, \mathbb{1}_{(0,1)}(p)}_{\text{Kernel of a Beta distribution}}$$

$$\Rightarrow \pi(p) = \frac{1}{B(a,b)} p^{a-1}(1-p)^{b-1}\, \mathbb{1}_{(0,1)}(x)$$ ← conjugate prior for the original param

Ex. $X|\theta \sim N(0, \sigma_0^2) = N(0,1)$

$$f(x|\theta) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\vartheta)^2}{2}} \qquad \vartheta \in \mathbb{R} = \Theta$$
$$= \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\right) e^{x\vartheta - \frac{\vartheta^2}{2}} \qquad \psi(\vartheta) = \frac{\vartheta^2}{2}, \quad h(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

Conjugate prior: $\quad \pi(\vartheta) \propto e^{\vartheta\mu - \lambda\frac{\vartheta^2}{2}} \qquad \lambda > 0, \ \mu \in \mathbb{R}$

$$\propto \underbrace{e^{-\frac{\lambda}{2}(\vartheta-\mu)^2}}_{\text{kernel of a gaussian}}$$

$$\Rightarrow \pi(\vartheta): \ N(\eta, \tau^2) \quad \eta, \tau^2 \text{ fixed} \qquad \text{(we'll see how to fix them)}$$

✱ Examples of Jeffrey's prior

Ex. $X|p \sim Bi(n,p)$

$f(x,p) = \binom{n}{x} p^x (1-p)^{1-x}$ $\quad x = 0,1,\dots,n \quad p \in (0,1)$

Information matrix:

$\log f(x,p) = \log\binom{n}{x} + x\log(p) + (1-x)\log(1-p)$

$\frac{\partial}{\partial p}\log f(x,p) = \frac{x}{p} - \frac{n-x}{1-p}$, $\quad \frac{\partial^2}{\partial p^2}\log f(x,p) = x\left(-\frac{1}{p^2}\right) - \frac{n-x}{(1-p)^2}$

$$I(\theta) = \mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log f(X,\theta)\right)^2\right] = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\log f(X,\theta)\right]$$

$\Rightarrow I(p) = \mathbb{E}\left[\frac{X}{p^2} + \frac{n-X}{(1-p)^2}\right] = \frac{1}{p^2}\mathbb{E}[X] + \frac{n-\mathbb{E}[X]}{(1-p)^2} = p^2(np) + \frac{n-(np)}{(1-p)^2}$

$= [\dots] = \frac{n}{p(1-p)}$

$\Rightarrow$ Jeffrey's prior: $\pi^*(p) \propto \sqrt{\frac{n}{p(1-p)}}\, \mathbb{1}_{(0,1)}(p)$

$\propto p^{-\frac{1}{2}}(1-p)^{-\frac{1}{2}}\mathbb{1}_{(0,1)}(p)$

$\propto \underbrace{p^{\frac{1}{2}-1}(1-p)^{\frac{1}{2}-1}\mathbb{1}_{(0,1)}(p)}$

Beta distr. $\quad$ Beta$(\frac{1}{2},\frac{1}{2})$ $\quad$ (<u>not</u> improper)

But we know that Beta is conjugate!

$\pi^*(\theta) =$ Beta$(\frac{1}{2},\frac{1}{2})$

$\pi^*(\theta|X) =$ Beta$(\frac{1}{2}+x, \frac{1}{2}+(n-x))$

advantag vs. frequentists: in the case of $x=0$ (one datum) the frequentists would say that the estimation of p is 0 (unreasonable), while with this we get a reasonable distr. of p even if $x=0$

Ex. $X|\theta \sim N(\theta, \sigma_0^2) \overset{\text{w.l.o.g.}}{=} N(0,1)$

$f(x,\theta) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\theta)^2}{2}}$ $\quad \theta \in \mathbb{R} \quad x \in \mathbb{R}$

$\log f(x,\theta) \propto -\frac{(x-\theta)^2}{2} \Rightarrow \frac{\partial}{\partial\theta}(\dots) \propto \frac{2(x-\theta)}{2}$

$I(\theta) = \mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log f(X,\theta)\right)^2\right]$

$\Rightarrow I(\theta) = \mathbb{E}_\theta[(X-\theta)^2] = \mathrm{Var}(X|\theta) = 1 \quad (=\sigma^2) \quad =$ constant w.r.t. θ

$\Rightarrow$ Jeffrey's prior: $\pi^*(\theta) \propto \sqrt{1}$

$\propto c =$ constant $\quad c>0$

but θ lives in $\mathbb{R}$, and so this is not a density (improper distribution) 

However: (we apply Bayes' theorem for the posterior) holds even if the prior is an improper distribution !

$\pi^*(\theta|X) \propto e^{-\frac{(x-\theta)}{2}} \cdot 1 \quad \theta \in \mathbb{R}$

We interpret it as a function of θ: $\theta \mapsto \underbrace{e^{-\frac{(\theta-x)^2}{2}}}$

kernel of a gaussian: $N(x,1)$

$\Rightarrow$ the posterior is a proper density

$X_1,\dots,X_n \Rightarrow \pi^*(\theta|\underline{X}) = N(\bar{x}_n, \frac{1}{n})$ $\quad$ a posteriori $\not\exists$ influence of hyperparameters

Def. (Inverse Gamma distribution)

$X \sim$ inverse-gamma (α, β) $\quad \alpha, \beta > 0$ if $Y = \frac{1}{X} \sim$ gamma (α, β).

Density:

$$\begin{cases} f_Y(y) = \dfrac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-\beta y} \mathbb{1}_{(0,+\infty)}(y) \\ X = \frac{1}{Y}, \quad Y = \frac{1}{x}, \quad \dfrac{\partial Y}{\partial x} = -\frac{1}{x^2} \end{cases}$$

$$\Longrightarrow \quad f_X(x) = f_Y\left(\frac{1}{x}\right) \frac{1}{x^2} = \frac{\beta^\alpha}{\Gamma(\alpha)} \left(\frac{1}{x}\right)^{\alpha+2-1} e^{-\frac{\beta}{x}} \mathbb{1}_{(0,+\infty)}(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} \left(\frac{1}{x}\right)^{\alpha+1} e^{-\frac{\beta}{x}} \mathbb{1}_{(0,+\infty)}(x)$$

$X \sim$ inv-gamma (α, β) :

$$\begin{cases} \mathbb{E}[X] = \dfrac{\beta}{\alpha-1} & \text{if } \alpha > 1 \quad (\text{otherwise } \nexists) \\ \text{Var}(X) = \dfrac{\beta^2}{(\alpha-1)^2(\alpha-2)} & \text{if } \alpha > 2 \quad (\text{otherwise } +\infty) \end{cases}$$

just to see how it works:

$$\mathbb{E}[X] = \mathbb{E}\left[\frac{1}{Y}\right] = \int_0^\infty \frac{1}{y} \frac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-\beta y} dy = \frac{\beta^\alpha}{\Gamma(\alpha)} \boxed{\int_0^\infty y^{\alpha-2} e^{-\beta y} dy}$$

(almost) a gamma $(\alpha-1, \beta)$

$$= \frac{\beta^\alpha}{\Gamma(\alpha)} \cdot \frac{\Gamma(\alpha-1)}{\beta^{\alpha-1}} = \frac{\beta}{\alpha-1}$$

$\Gamma(\alpha) = (\alpha-1)\,\Gamma(\alpha-1)$

3rd well known conjugate bayesian model

Ex. Bayesian model: iid Gaussian sample (unknown mean and variance), conjugate prior (Normal - Inv-gamma)

$$\begin{cases} X_1, \dots, X_n \mid \mu, \sigma^2 \overset{iid}{\sim} N(\mu, \sigma^2) \\ \boxed{\begin{aligned} \mu \mid \sigma^2 &\sim N\left(\mu_0, \frac{\sigma^2}{n_0}\right) \\ \sigma^2 &\sim \text{inv-gamma}\left(\frac{\nu_0}{2}, \frac{\nu_0 \sigma_0^2}{2}\right) \end{aligned}} \end{cases}$$

$\mu_0 \in \mathbb{R}$, $n_0 > 0$, $\nu_0, \sigma_0^2 > 0$

normal inv.-gamma $(\mu_0, n_0, \nu_0, \sigma_0^2)$

every time we got a bivariate r.v. we can express the joint distribution as:

$$\pi(\mu, \sigma^2) = \pi(\mu \mid \sigma^2)\, \pi(\sigma^2)$$

- What do these parameters mean? (hyperparameters $\mu_0, n_0, \nu_0, \sigma_0^2$)

$$\mathbb{E}_\pi[\mu] = \mathbb{E}\left[\mathbb{E}[\mu \mid \sigma^2]\right] = \mathbb{E}[\mu_0] = \mu_0$$

the first hyperparameter of this prior (normal inv-gamma) represent the apriori expectation of μ.

Ex. $\mathrm{Var}_\pi(\mu) = \mathbb{E}[\mathrm{var}(\mu|\sigma^2)] + \mathrm{Var}(\mathbb{E}[\mu|\sigma^2])$

$= \mathbb{E}\left[\frac{\sigma^2}{n_0}\right] + \mathrm{var}(\mu_0)$

$= \frac{1}{n_0}\mathbb{E}[\sigma^2]$

$\mathbb{E}[\sigma^2] = \frac{\nu_0 \frac{\sigma_0^2}{2}}{\frac{\nu_0}{2} - 1} = \frac{\nu_0 \sigma_0^2}{\nu_0 - 2}$ if $\nu_0 > 2$

$\mathrm{Var}(\sigma^2) = \frac{4\nu_0\sigma_0^2}{(\nu_0-2)^2(\nu_0-4)}$ if $\nu_0 > 4$

} from the inverse-gamma distribution

$\Longrightarrow \mathrm{Var}_\pi(\mu) = \frac{1}{n_0} \cdot \frac{\nu_0\sigma_0^2}{\nu_0 - 2}$ if $\nu_0 > 2$

To interpret the parameters of the inverse gamma distribution let us compute the mean and variance of $1/\sigma^2$ (which is the precision of our data):

σ^2 is the variance of our data

$1/\sigma^2$ is the precision of our data

$\mathbb{E}\left[\frac{1}{\sigma^2}\right] = \frac{\nu_0/2}{\nu_0\sigma_0^2/2} = \frac{1}{\sigma_0^2}$

the way to fix σ_0^2 is not thinking about the variance but, instead, about the precision (that we have): $1/\sigma^2$

$\mathrm{Var}\left(\frac{1}{\sigma^2}\right) = \frac{2}{\nu_0} \cdot \frac{1}{\sigma_0^4}$ = uncertainty of the precision w.r.t. its expectation

What about n_0?

Usually we fix it s.t. $\frac{1}{n_0} > 1$ (e.g. $\frac{1}{n_0} = 10$)

so that, a priori the conditional distribution of μ is gaussian with a variance that is 10 times σ^2 (actually this is more for computational reasons)

- Let's now check that this distribution is conjugate.

First we compute the likelihood:

$$L(\mu, \sigma^2, x_1, \dots, x_n) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

$$= \left(\frac{1}{\sqrt{2\pi}}\right)^n \left(\frac{1}{\sigma^2}\right)^{n/2} e^{-\frac{\sum_{i=1}^n (x_i-\mu)^2}{2\sigma^2}}$$

$$\sum_{i=1}^n (x_i-\mu)^2 = \sum_{i=1}^n = (x_i \pm \bar{x} - \mu)^2$$

$$= (n-1)s^2 + n(\mu - \bar{x})^2$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})^2$$

posterior $\propto$ likelihood · prior

$$\pi(\mu, \sigma^2|\underline{x}) \propto \left[\left(\frac{1}{\sigma^2}\right)^{n/2} e^{-\frac{(n-1)s^2}{2\sigma^2}} e^{-\frac{n}{2\sigma^2}(\mu-\bar{x})^2}\right] \cdot \left[\underbrace{\frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{\frac{\sigma^2}{n_0}}} e^{-\frac{n_0(\mu-\mu_0)^2}{2\sigma^2}}}_{\mu|\sigma^2} \underbrace{\left(\frac{1}{\sigma^2}\right)^{\frac{\nu_0+1}{2}} e^{-\frac{\nu_0\sigma_0^2}{2\sigma^2}}}_{\sigma^2}\right]$$

Ex. $\pi(\mu,\sigma^2|\underline{x}) \propto \left(\frac{1}{\sigma^2}\right)^{\frac{1}{2}} e^{-\frac{n(\mu-\bar{x})^2}{2\sigma^2}} e^{-\frac{n_0(\mu-\mu_0)^2}{2\sigma^2}} \left(\frac{1}{\sigma^2}\right)^{\frac{n+\nu_0}{2}+1} e^{-\frac{1}{\sigma^2}\left[\frac{(n-1)s^2}{2}+\frac{\nu_0\sigma_0^2}{2}\right]} \mathbb{1}_{(0,\infty)}(\sigma^2)$

$$: -\frac{1}{2\sigma^2}\left[n(\mu-\bar{x})^2 + n_0(\mu-\mu_0)^2\right] \overset{*}{=} -\frac{1}{2\sigma^2}\left[(n_0+n)(\mu-\mu_1)^2 + \frac{n_0 n}{n_0+n}(\mu_0-\bar{x})^2\right]$$

$$\mu_1 = \frac{n_0\mu_0 + n\bar{x}}{n_0+n}$$

$$* \; d_1(z-c_1)^2 + d_2(z-c_2)^2 = (d_1+d_2)(z-c)^2 + \frac{d_1 d_2}{d_1+d_2}(c_1-c_2)^2$$

$$\Rightarrow \pi(\mu,\sigma^2|\underline{x}) \propto \underbrace{\left(\frac{1}{\sigma^2}\right)^{\frac{1}{2}} e^{-\frac{(n_0+n)(\mu-\mu_1)^2}{2\sigma^2}}}_{\mu|\sigma^2} \underbrace{\left(\frac{1}{\sigma^2}\right)^{\frac{n+\nu_0}{2}+1} e^{-\frac{1}{\sigma^2}\left[\frac{n_0 n}{2(n_0+n)}(\mu-\bar{x}) + \frac{(n-1)s^2}{2} + \frac{\nu_0\sigma_0^2}{2}\right]}}_{\sigma^2} \mathbb{1}_{(0,\infty)}(\sigma^2)$$

$$\Rightarrow \begin{cases} \mu|\sigma^2,\underline{x} \sim N\left(\mu_1, \frac{\sigma^2}{n+n_0}\right) \\ \sigma^2|\underline{x} \sim \text{inv-gamma}\left(\frac{\nu_0+n}{2}, \frac{n_0 n}{2(n_0+n)}(\mu-\bar{x}) + \frac{(n-1)s^2}{2} + \frac{\nu_0\sigma_0^2}{2}\right) \end{cases}$$

9/10

Prior:

$$\left.\begin{aligned} X_1,\dots,X_n|\mu,\sigma^2 &\sim N(\mu,\sigma^2) \\ \mu|\sigma^2 &\sim N\left(\mu_0, \frac{\sigma^2}{n_0}\right) \\ \sigma^2 &\sim \text{inv-gamma}\left(\frac{\nu_0}{2}, \frac{\nu_0\sigma_0^2}{2}\right) \end{aligned}\right] (\mu,\sigma^2) \sim \text{normal-inv-gamma}(\mu_0, n_0, \nu_0, \sigma_0^2)$$

$\mu_0 \in \mathbb{R}$, $n_0, \nu_0, \sigma_0^2 > 0$

Posterior:

$(\mu,\sigma^2)|x_1,\dots,x_n \sim \text{normal-inv-gamma}(\mu_1, n_1, \nu_1, \sigma_1^2)$ → conjugate

$$\begin{cases} \mu_1 = \frac{n_0\mu_0 + n\bar{x}}{n_0+n} \\ n_1 = n_0 + n \\ \nu_1 = \nu_0 + n \\ \sigma_1^2 = \frac{\frac{n_0 n}{(n_0+n)}(\mu-\bar{x}) + (n-1)s^2 + \nu_0\sigma_0^2}{\nu_0+n} \end{cases}$$

- We want now to compute the prior distribution of μ (marginal).

(The posterior marginal will have the same form)

We use: $\begin{cases} \mu|\sigma^2 \sim N\left(\mu_0, \frac{\sigma^2}{n_0}\right) \\ \sigma^2 \sim \text{inv-gamma}\left(\frac{\nu_0}{2}, \frac{\nu_0\sigma_0^2}{2}\right) \end{cases}$

$$\pi(\mu) = \int_0^\infty \text{joint distr}(\mu,\sigma^2)\,d\sigma^2 = \int_0^\infty \underbrace{\text{conditional}(\mu|\sigma^2)}\;\underbrace{\text{marginal}(\sigma^2)}\,d\sigma^2$$

$$\pi(\mu) = \int_0^\infty \underbrace{\frac{1}{\sqrt{2\pi}}\left(\frac{n_0}{\sigma^2}\right)^{\frac{1}{2}} e^{-\frac{(\mu-\mu_0)^2}{2\sigma^2/n_0}}}\;\underbrace{\frac{\left(\frac{\nu_0\sigma_0^2}{2}\right)^{\frac{\nu_0}{2}}}{\Gamma\left(\frac{\nu_0}{2}\right)}\left(\frac{1}{\sigma^2}\right)^{\frac{\nu_0}{2}+1} e^{-\frac{\nu_0}{2}\frac{\sigma_0^2}{\sigma^2}}}\,d\sigma^2$$

$$= \sqrt{\frac{n_0}{2\pi}}\left(\frac{\nu_0\sigma_0^2}{2}\right)^{\frac{\nu_0}{2}}\frac{1}{\Gamma\left(\frac{\nu_0}{2}\right)} \cdot \int_0^\infty x^{\frac{1}{2}+\frac{\nu_0}{2}+1-2} e^{-x\left(\frac{n_0}{2}(\mu-\mu_0)^2 + \frac{\nu_0\sigma_0^2}{2}\right)} dx$$

$$x = \frac{1}{\sigma^2}, \quad \sigma^2 = \frac{1}{x}, \quad d\sigma^2 = \left(-\frac{1}{x^2}\right)dx$$

Ex $\pi(\mu) = \sqrt{\frac{n_0}{2\pi}} \left(\frac{\nu_0\sigma_0^2}{2}\right)^{\frac{\nu_0}{2}} \frac{1}{\Gamma(\frac{\nu_0}{2})} \cdot \int_0^\infty x^{\frac{\nu_0+1}{2}} e^{-x\left(\frac{n_0}{2}(\mu-\mu_0)^2 + \frac{\nu_0\sigma_0^2}{2}\right)} dx$

kernel of a gamma $\left(\frac{\nu_0+1}{2}, \frac{n_0}{2}(\mu-\mu_0)^2 + \frac{\nu_0\sigma_0^2}{2}\right)$

$= \sqrt{\frac{n_0}{2\pi}} \left(\frac{\nu_0\sigma_0^2}{2}\right)^{\frac{\nu_0}{2}} \frac{1}{\Gamma(\frac{\nu_0}{2})} \cdot \frac{\Gamma\left(\frac{\nu_0+1}{2}\right)}{\left(\frac{n_0}{2}(\mu-\mu_0)^2 + \frac{\nu_0\sigma_0^2}{2}\right)^{\frac{\nu_0+1}{2}}}$

$= \sqrt{\frac{n_0}{2\pi}} \left(\frac{\nu_0\sigma_0^2}{2}\right)^{\frac{\nu_0}{2}} \frac{\Gamma\left(\frac{\nu_0+1}{2}\right)}{\Gamma\left(\frac{\nu_0}{2}\right)} \frac{1}{\left(\frac{\nu_0\sigma_0^2}{2}\right)^{\frac{\nu_0+1}{2}} \left[1 + \frac{n_0}{\nu_0}\left(\frac{\mu-\mu_0}{\sigma_0}\right)^2\right]^{\frac{\nu_0+1}{2}}}$

$= \frac{\Gamma\left(\frac{\nu_0+1}{2}\right)}{\Gamma\left(\frac{\nu_0}{2}\right)} \sqrt{\frac{n_0}{\pi\nu_0\sigma_0^2}} \left[1 + \frac{1}{\nu_0}\left(\frac{\sqrt{n_0}(\mu-\mu_0)}{\sigma_0}\right)^2\right]^{-\frac{\nu_0+1}{2}}$

t-student ; ν_0 degree of freedom
μ_0 location
$\sigma_0/\sqrt{n_0}$ scale

Notice : the posterior for μ $(\mu|\underline{x})$ will be again a t-student with parameters $\nu_1, \mu_1, \frac{\sigma_1}{\sqrt{n_1}}$

$X \sim t_\nu(\mu; \sigma^2)$ if $f(x) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})} \frac{1}{\sqrt{\pi\nu\sigma^2}} \left[1 + \frac{1}{\nu}\left(\frac{x-\mu}{\sigma}\right)^2\right]^{-\frac{\nu+1}{2}}$

ν → degrees of freedom; μ → location; σ^2 → scale

if : $\nu > 1 \Rightarrow \exists\, \mathbb{E}[X] : \mathbb{E}[X] = \mu$

$\nu > 2 \Rightarrow \mathrm{Var}(X) = \frac{\nu}{\nu-2}\sigma^2$

Note:
As $\nu \uparrow$ there exists more and more moments ($\nu>1$ ∃ first, $\nu>2$ ∃ second, ..) and so as $\nu \uparrow$ tails become thinner

suppose now that we have observed n data.

- What is the predictive distribution? $\mathbb{P}(X_{n+1} \mid X_1 = x_1, \dots, X_n = x_n)$?

predictive density = ∫ density integrated w.r.t. the posterior density

$p(x \mid x_1, \dots, x_n) = \int_{\mathbb{R}} \int_0^{+\infty} f(x|\mu,\sigma^2)\, \pi(\mu, \sigma^2 | x_1, \dots, x_n)\, d\mu\, d\sigma^2$

$f(x|\mu,\sigma^2)$: univariate normal density; $\pi(\mu,\sigma^2|x_1,\dots,x_n)$: normal inverse gamma with the updated parameters

$\propto \underbrace{\left[1 + \frac{1}{\nu_1}\left(\frac{x-\mu_1}{\sigma_1}\sqrt{\frac{n_1}{n_1+1}}\right)^2\right]^{-\frac{\nu_1+1}{2}}}_{t_{\nu_1}\left(\mu_1, \left(\sigma_1\sqrt{\frac{n_1+1}{n_1}}\right)^2\right)}$

- What happens if we use n improper priors? (Jeffrey's prior)

$\pi(\mu, \sigma^2) \propto \frac{1}{\sigma^2} \mathbb{1}_{(0,\infty)}(\sigma^2)$

$\Rightarrow \pi(\mu, \sigma^2 | x_1, \dots, x_n) \propto \frac{1}{\sigma^2} \underbrace{\left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}} e^{-\frac{n(\mu-\bar{x})^2}{2\sigma^2}} e^{-\frac{(n-1)s^2}{2\sigma^2}}}_{\text{likelihood}} \mathbb{1}_{(0,\infty)}(\sigma^2)$

$\propto \underbrace{\left(\frac{1}{\sigma^2}\right)^{\frac{1}{2}} e^{-\frac{(\mu-\bar{x})^2}{2\sigma^2/n}}}_{\mu|\sigma^2,\underline{x}} \underbrace{\left(\frac{1}{\sigma^2}\right)^{\frac{n+1}{2}+1-1} e^{-\frac{(n-1)s^2}{2\sigma^2}}}_{\sigma^2|\underline{x}} \mathbb{1}_{(0,\infty)}(\sigma^2)$

$\Rightarrow \begin{cases} \mu|\sigma^2,\underline{x} \sim N\left(\bar{x}, \frac{\sigma^2}{n}\right) \\ \sigma^2|\underline{x} \sim \text{inv-gamma}\left(\frac{n-1}{2}, \frac{(n-1)s^2}{2}\right) \end{cases}$

a posteriori they're all proper distributions which do not depend on hyperparams. (they depend only on n and data $(\bar{x}, s^2)$)

MONTE CARLO METHODS

Idea: all we want to know about a distribution can be achieved by simulating many (many) draws from it. Everything can be known from a (iid) sample of the distribution. This idea is based on the law of large numbers.

Context:

$\theta \sim \pi$ distribution (which we consider as posterior of the Bayesian model)

random object distributed as π

$\theta \in \Theta \subseteq \mathbb{R}^p$

Suppose to consider $h: \Theta \to \mathbb{R}$ s.t. $\mathbb{E}_\pi[|h(\theta)|] < +\infty$

We define: $\bar{h} := \int_\Theta h(\theta)\, \pi(d\theta)$.

We want to evaluate $\bar{h}$.

Theorem (Strong Law of Large Numbers)

1. $\theta^{(1)}, \theta^{(2)}, \theta^{(3)}, \dots$ iid from π, $\mathbb{E}_\pi[|h(\theta)|] < \infty$ $\left(h: \Theta \to \mathbb{R}\right)$

$$\Longrightarrow \boxed{\bar{h}^{(T)} = \frac{1}{T}\sum_{t=1}^{T} h(\theta^{(t)}) \xrightarrow[a.s.]{T\to\infty} \bar{h} = \int_\Theta h(\theta)\,\pi(d\theta)}$$

2. Fix $p \in (0,1)$, $q_p = p^{th}$ quantile of $h(\theta)$, $h: \Theta \to \mathbb{R}$.
$\theta^{(1)}, \theta^{(2)}, \theta^{(3)}, \dots$ iid from π.
Consider $h(\theta^{(1)}), h(\theta^{(2)}), h(\theta^{(3)}), \dots, h(\theta^{(T)})$, which are random variables, and consider the empirical quantiles corresponding to those random variables:

$q_p^{(T)} :=$ p^{th} empirical quantile of $h(\theta)$

$$\Longrightarrow q_p^{(T)} \xrightarrow[a.s.]{T\to\infty} q_p$$

p^{th} quantile of $h(\theta)$:
if $h(\theta)$ is absolutely continuous:
$F_{h(\theta)}(q_p) = p$
if $h(\theta)$ is not:
$\mathbb{P}(h(\theta) \le q_p) \ge p$
$\mathbb{P}(h(\theta) \ge q_p) \ge 1-p$

Basically this theorem says that:
the error: $err^{(T)} := \bar{h}^{(T)} - \bar{h} \xrightarrow[a.s.]{T\to\infty} 0$

How fast? We need to know it to know how to set up T.

If it go to 0 slowly $\Rightarrow$ T high
If it go to 0 fastly $\Rightarrow$ T low

Theorem (Central Limit Theorem)

Consider an infinite sequence $\theta^{(1)}, \theta^{(2)}, \dots$ of iid from π.
Consider $h: \Theta \to \mathbb{R}$, $\bar{h} := \mathbb{E}_\pi[h(\theta)]$, s.t. $0 < Var(h(\theta)) := \sigma^2 < +\infty$.
We define:

$$\sigma^{2(T)} = \frac{1}{T}\sum_{t=1}^{T}\left(h(\theta^{(t)}) - \bar{h}\right)^2 := \text{empirical variance (of the first } T \text{ random var.s)}$$

$$\Longrightarrow \begin{cases} \bullet\ \sqrt{T}\left(\bar{h}^{(T)} - \bar{h}\right) \xrightarrow{d} N(0, \sigma^2) & = \bar{h}^{(T)} - \bar{h} \text{ goes to 0 as } \left(\sqrt{T}\right)^{-1} \\ \bullet\ \sigma^{2(T)} \xrightarrow[a.s.]{T\to\infty} \sigma^2 \end{cases}$$

$$\Longrightarrow \text{error: } err^{(T)} = \bar{h}^{(T)} - \bar{h} \approx N\left(0, \frac{\sigma^2}{T}\right) \approx N\left(0, \frac{\sigma^{2(T)}}{T}\right)$$

we don't know σ^2, but once we simulate all rand. vars. $h(\theta^{(1)}), \dots, h(\theta^{(T)})$, we can compute the empirical mean and the empirical variance

and so:

$$\mathbb{P}\left(|err^{(T)}| > c\right) = \mathbb{P}\left(\frac{|\bar{h}^{(T)} - \bar{h}|}{\sqrt{\frac{\sigma^2(T)}{T}}} > \frac{c}{\sqrt{\frac{\sigma^2(T)}{T}}}\right) \overset{CLT}{\approx} 2\left(1 - \phi\left(\frac{c}{\sqrt{\frac{\sigma^2(T)}{T}}}\right)\right)$$

If T is large $\Rightarrow$ $\phi(\cdot\cdot) \approx 1$ (since $(\cdot\cdot) \to \infty$) $\Rightarrow$ $\mathbb{P}(|err^{(T)}| > c) \xrightarrow{T \to \infty} 0$

How do we apply this method?

Ex. Data on educational level and number of children of 155 women.
We divide women on educational level: with and without college degree.

conditionally $\perp\!\!\!\perp$
- $Y_{1,1}, \dots, Y_{n_1,1}$ — # children of women of group 1 (without college degree)
- $Y_{1,2}, \dots, Y_{n_2,2}$ — # children of women of group 2 (with a)

$n_1 = 111$, $\sum_i y_{i,1} = 217$, $\bar{y}_1 = 1.55$

$n_2 = 44$, $\sum_i y_{i,2} = 66$, $\bar{y}_2 = 1.5$

Assumptions:

$$Y_{1,1}, \dots, Y_{n_1,1} \mid \theta_1 \overset{iid}{\sim} Poi(\theta_1)$$ ← average number of children of women of group 1

$$Y_{1,2}, \dots, Y_{n_2,2} \mid \theta_2 \overset{iid}{\sim} Poi(\theta_2)$$ ← average number of children of women of group 2

$$\theta_1, \theta_2 \overset{iid}{\sim} gamma(\alpha, \beta)$$

$\alpha = 2$, $\beta = 1$

$\mathbb{E}[\theta_i] = 2$, $i = 1,2$

$Var(\theta_i) = 2$, $i = 1,2$

a priori they (θ_1, θ_2) seem to be equal

- We want to see $\mathbb{P}(\theta_1 > \theta_2 \mid y_1, y_2)$.

$$\theta_1, \theta_2 \mid y_1, y_2 = gamma(\alpha + \textstyle\sum_i y_{i,1}, \beta + n_1) \cdot gamma(\alpha + \sum_i y_{i,2}, \beta + n_2)$$

a posterior distribution

$\Rightarrow$ $\mathbb{P}(\theta_1 > \theta_2 \mid y_1, y_2) = 0.97$ (true value) computed with integration techniques, not with simulations

How can we compute it with Monte Carlo methods?

$$\mathbb{P}(\theta_1 > \theta_2 \mid y_1, y_2) = \int_{\{(\theta_1,\theta_2) \in (\mathbb{R}^+)^2 : \theta_1 > \theta_2\}} \pi(d\theta_1, d\theta_2 \mid y_1, y_2)$$

$$= \int_{\mathbb{R}^+ \times \mathbb{R}^+} \boxed{\mathbb{1}_{\{\theta_1 > \theta_2\}}} \pi(d\theta_1, d\theta_2 \mid y_1, y_2)$$ $\quad h(\theta) = \mathbb{1}_{\{\theta_1 > \theta_2\}}$

$\Rightarrow$ We simulate (from the posterior): $(\theta_1^{(1)}, \theta_2^{(1)}), (\theta_1^{(2)}, \theta_2^{(2)}), \dots, (\theta_1^{(T)}, \theta_2^{(T)})$

can we do it? Yes, since the posterior is the product of two gammas

$$\Rightarrow \bar{h}^{(T)} = \frac{\#\{\theta_1^{(j)} > \theta_2^{(j)}\}}{T}$$

We can also compute:

$$\mathbb{E}[\theta_2 \mid y_1, y_2] \approx \frac{1}{T} \sum_{t=1}^{T} \theta_2^{(t)} = \bar{\theta}_2, \quad Var(\theta_2 \mid y_1, y_2) = \frac{1}{T-1} \sum_{t=1}^{T} (\theta_2^{(t)} - \bar{\theta}_2)^2$$

[..]

Recall of the model:

$$\left.\begin{aligned} Y_{1,1}, \dots, Y_{n_1,1} \mid \theta_1 &\sim \text{Poisson}(\theta_1) \quad \text{iid} \\ Y_{1,2}, \dots, Y_{n_2,2} \mid \theta_2 &\sim \text{Poisson}(\theta_2) \quad \text{iid} \end{aligned}\right\} \perp\!\!\!\perp$$

$$\theta_1, \theta_2 \overset{\text{iid}}{\sim} \text{Gamma}(\alpha, \beta) \qquad \alpha = 2,\ \beta = 1$$

$$\left.\begin{aligned} \theta_1 \mid y_1 &\sim \text{Pois}(\bullet) \\ \theta_2 \mid y_2 &\sim \text{Pois}(\bullet) \end{aligned}\right\} \perp\!\!\!\perp$$

We want to check (via MC) that:

$$\mathbb{P}(\theta_1 > \theta_2 \mid y_1, y_2) = 0.97$$

METHOD OF COMPOSITION FOR SAMPLING

Useful when we're able to sample from the conditional distribution of the random element we're interested in, given the auxiliary variable and the marginal of the variable

Suppose that we're not able to sample from $\mathcal{L}(X_1)$ but we can sample from $\mathcal{L}(X_1 \mid X_2)$ and $\mathcal{L}(X_2)$

$$\Longrightarrow \quad \mathcal{L}(X_1) = \int \mathcal{L}(X_1, dX_2) = \int \underbrace{\mathcal{L}(X_1 \mid X_2)\, \mathcal{L}(dX_2)}_{\text{from here we can sample}}$$

$\mathcal{L}(X_1)$?
↓
$\mathcal{L}(X_1, X_2)$
↓
$\mathcal{L}(X_1 \mid X_2)\, \mathcal{L}(X_2)$

$\Longrightarrow$ To sample from $\mathcal{L}(X_1)$:
it's enough to sample from $\mathcal{L}(X_1, X_2)$:
$(X_1^{(j)}, X_2^{(j)})$ MC sample from $\mathcal{L}(X_1, X_2)$
but how can we simulate from this?

1. We simulate $X_2^{(j)} \sim \mathcal{L}(X_2)$
2. We simulate $X_1^{(j)} \sim \mathcal{L}(X_1 \mid X_2 = X_2^{(j)})$

$\Longrightarrow$ We obtain samples of the joint $\mathcal{L}(X_1, X_2)$.
How can we obtain samples of $\mathcal{L}(X_1)$?
We consider only the first component of each couple.

This is what we'll use when we'll need to sample from the posterior distribution.

How to: EVALUATE THE POSTERIOR PREDICTIVE DENSITY

(general settings)

$$\begin{aligned} X_1, \dots, X_n \mid \theta &\overset{\text{iid}}{\sim} f(\bullet \mid \theta) \\ \theta &\sim \pi \end{aligned}$$

The posterior predictive density is:

$$m_{X_{n+1} \mid X_1, \dots, X_n}(x \mid x_1, \dots, x_n) = \int_\Theta \underbrace{f(x \mid \theta)}_{h(\theta)}\, \pi(d\theta \mid x_1, \dots, x_n)$$

$$= \int_\Theta h(\theta)\, \pi(d\theta \mid \underline{x}) \quad \longleftrightarrow \quad \frac{1}{T} \sum_{j=1}^{T} h(\vartheta^{(j)})$$

it can be expressed as the Ergodic (/empirical) mean of the function $h(\cdot)$

Suppose we want to sample from the posterior predictive (we don't want to evaluate the posterior predictive density, we want to simulate sample from it):

$$\mathcal{L}(X_{n+1} \mid X_1, \dots, X_n) = \int_{\Theta} \mathcal{L}(X_{n+1}, d\theta \mid X_1, \dots, X_n)$$

$$= \int_{\Theta} \mathcal{L}(X_{n+1} \mid \theta, X_1, \dots, X_n)\ \mathcal{L}(d\theta \mid X_1, \dots, X_n)$$

$$= \int_{\Theta} \mathcal{L}(X_{n+1} \mid \theta)\ \pi(d\theta \mid \underline{X})$$

$\mathcal{L}(X_{n+1} \mid X_1, \dots, X_n)$: we want to sample from this law

$\mathcal{L}(X_{n+1} \mid \theta, X_1, \dots, X_n)$: we have assumed that $X_i \mid \theta$ are $\perp\!\!\!\perp$

$\mathcal{L}(d\theta \mid X_1, \dots, X_n)$: this is the posterior $\pi(d\theta \mid \underline{X})$

So, how can we sample from $\mathcal{L}(X_{n+1} \mid X_1, \dots, X_n)$?

1. We sample $\theta^{(j)} \sim \pi(d\theta \mid x_1, \dots, x_n)$
2. We sample $X_{n+1}^{(j)} \sim f(\cdot \mid \theta = \theta^{(j)})$

$\Longrightarrow$ we obtain $(X_{n+1}^{(j)}, \theta^{(j)})$ MC sample of the joint of X_{n+1} and θ (given $\underline{x}$)

$\Longrightarrow$ we consider only the first component : $X_{n+1}^{(j)}$

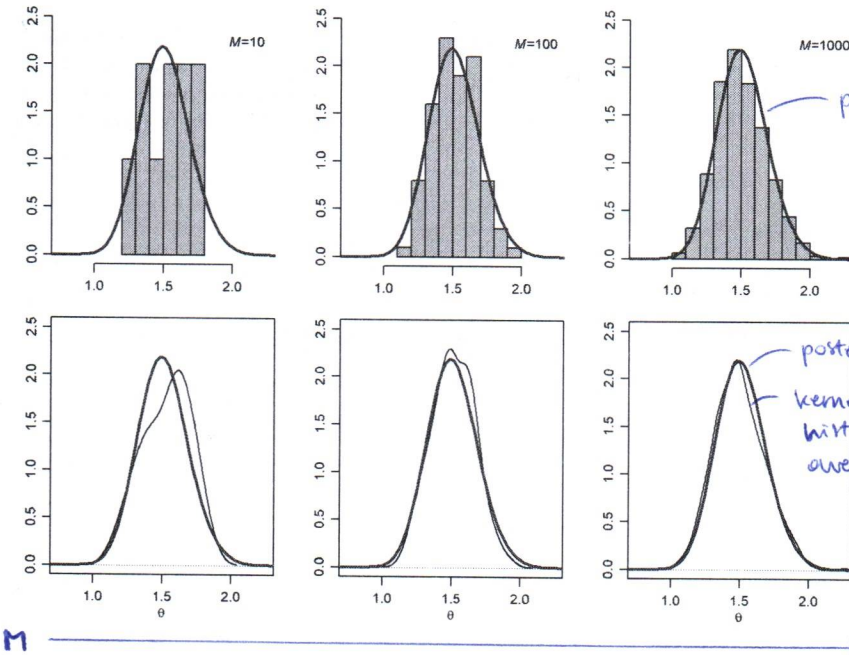
$\Longrightarrow$ we have a MC sample from the posterior predictive distribution

Monte Carlo Methods

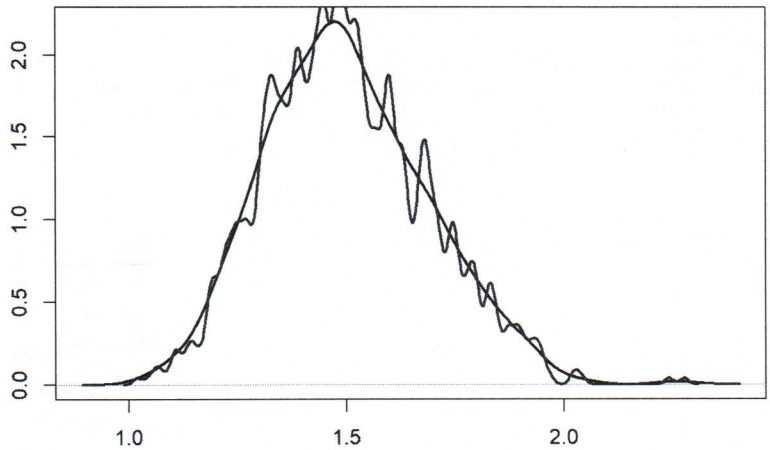
```
### ---------------------------------------------------------------------------
### ---------------------------------------------------------------------------
### From Chapter 4 in P. Hoff's book
### ---------------------------------------------------------------------------
### ---------------------------------------------------------------------------
# Y_1 is the vector of the numbers of children for the n_1 women WITHOUT college degrees
# Y_2 is the vector of the numbers of children for the n_2 women WITH college degrees
# The components of the two vectors are iid from Poisson with parameter theta_1 and theta_2
# PRIOR:     (theta_1,theta_2): theta_1,theta_2 iid  gamma(alpha,beta), alpha=2, beta=1, s.t. E(theta_i)=2, Var(theta_i)=2
# POSTERIOR: (theta_1,theta_2): gamma(a+\sum y_{1,i}, beta+n_1)\times gamma(a+\sum y_{2,i},beta+n_2)
#                             = gamma(219,112)\times gamma(68,45)
### ---------------------------------------------------------------------------
### Pat I: CONDITIONAL OF theta_2
### ---------------------------------------------------------------------------
# Random number draws from the posterior of  theta_2, given Y_{1,2},Y_{2,2},...Y_{n_2,2}
x11()
par(mar=c(3,3,.25,1),mgp=c(1.75,.75,0))
par(mfrow=c(2,3))
set.seed(1)
a <- 68  ;  b <- 45
set.seed(1)
theta.support <- seq(0,3,length=100)

# 3 different groups of iid draws from the posterior of theta_2
theta.sim10   <- rgamma(10,a,b)    # iid sample of size M=10 from the gamma(a,b)
theta.sim100  <- rgamma(100,a,b)   # size M = 100
theta.sim1000 <- rgamma(1000,a,b)  # size M = 1000

xlim = c(.75,2.25)
ylim = c(0,2.5)
lty  = 1

# Plot of the histogramm and kernel density estimate for the 3 groups of iid draws
hist( theta.sim10, prob=T,xlim=xlim,ylim=ylim,xlab="",main="",ylab="")
lines(theta.support,dgamma(theta.support,a,b),col="red",lwd=2,lty=lty)
text(2.1,2.25,expression(paste(italic(M),"=10",sep="")))

hist( theta.sim100, prob=T,xlim=xlim,ylim=ylim,xlab="",main="" ,ylab="")
lines(theta.support,dgamma(theta.support,a,b),col="red",lwd=2,lty=lty)
text(2.1,2.25,expression(paste(italic(M),"=100",sep="")))

hist( theta.sim1000, prob=T,xlim=xlim,ylim=ylim,xlab="",main="" ,ylab="")
lines(theta.support,dgamma(theta.support,a,b),col="red",lwd=2,lty=lty)
text(2.1,2.25,expression(paste(italic(M),"=1000",sep="")))

plot(density(theta.sim10),xlim=xlim,ylim=ylim,xlab=expression(theta),main="",ylab="")
lines(theta.support,dgamma(theta.support,a,b),col="red",lwd=2,lty=lty)

plot(density(theta.sim100),xlim=xlim,ylim=ylim,xlab=expression(theta),main="",ylab="")
lines(theta.support,dgamma(theta.support,a,b),col="red",lwd=2,lty=lty)

plot(density(theta.sim1000),xlim=xlim,ylim=ylim,xlab=expression(theta),main="",ylab="")
lines(theta.support,dgamma(theta.support,a,b),col="red",lwd=2,lty=lty)
```

Main goal of the example: we want to prove $\theta_1 > \theta_2$. We need to simulate iid from the posterior distribution of θ_1 and θ_2 and then compute the posterior prob. that $\theta_1 > \theta_2$.

↳ simulation of 1000 iid from a gamma(a,b)



```
# Something about: kernel density estimator
pippo = density(theta.sim1000)
plot(density(theta.sim1000),main="",ylab="",lwd=2)
lines(density(theta.sim1000,bw=0.01),main="",ylab="",col="blue",lwd=2)
```




```
### -----------------------------------------------------------------------------
### Monte Carlo computation of the posterior mean of theta_2
### -----------------------------------------------------------------------------
### A posteriori: theta_2 ~ gamma(a+sy=68, b+n=45)
set.seed(1)
a  <- 2
b  <- 1
sy <- 66
n  <- 44

theta.sim10   <- rgamma(10,a+sy,b+n)
theta.sim100  <- rgamma(100,a+sy,b+n)
theta.sim1000 <- rgamma(1000,a+sy,b+n)

# Since we know the a posteriori distribution: the exact value of the mean is
(a+sy)/(b+n)
```

we know that this is the true distribution, so the exact value for the mean will be the ratio of the two (posterior) parameters.

```
## [1] 1.511111
```

```
# Monte Carlo approximation
# MC estimate for 10 draws
mean(theta.sim10)
```

→ We compute the mean of the MC sample from the posterior distribution (10, 100, 1000 draws)

```
## [1] 1.532794
```

```
# MC standard error for 10 draws
sqrt(var(theta.sim10)/10)
```

→ $\sqrt{\frac{\text{empirical variance}(\hat{\theta})}{N}}$

why are we interested? This controls the probability of having an error estimating the true parameter through a MC estimator.

```
## [1] 0.05410572
```

```
# Probability that the absolut value of the error is Larger than c
c = 0.01
2*(1-pnorm(c/(sqrt(var(theta.sim10)/10))))
```

```
## [1] 0.8533676
```

→ very large, too much! We shouldn't use too little iterations

```
# MC estimate for 100 draws
mean(theta.sim100)
```

```
## [1] 1.513947
```

```
# MC standard error for 100 draws
sqrt(var(theta.sim100)/100)
```

```
## [1] 0.01556436
```

```
2*(1-pnorm(c/(sqrt(var(theta.sim100)/100))))
```

```
## [1] 0.5205529
```
→ much better, but still..

```
# MC estimate for 1000 draws
mean(theta.sim1000)
```

```
## [1] 1.501015
```

```
# MC standard error for 1000 draws
sqrt(var(theta.sim1000)/1000)
```

```
## [1] 0.005886511
```

```
2*(1-pnorm(c/(sqrt(var(theta.sim1000)/1000))))
```

```
## [1] 0.08935701
```
→ this is good (this is the one we should always check)

```
### -------------------------------------------------------------------------
### MC computation of the posterior distribution function at 1.75
### -------------------------------------------------------------------------
# Exact value
pgamma(1.75,a+sy,b+n)
```

We can evaluate the posterior distr. function in a specific point:

$$\mathbb{P}(\theta_2 < 1.75 \mid y_1, y_2) = \int_\Theta \underbrace{\mathbb{1}_{(-\infty, 1.75)}(\theta)}_{h(\theta)} \pi(d\theta \mid y_1, y_2)$$

```
## [1] 0.8998286
```

```
# The MC estimate is the relative frequency of the simulated draws less than 1.75
mean(theta.sim10   < 1.75)   # 10   MC draws
```

→ $\frac{1}{T}\sum_{j=1}^{T} \mathbb{1}_{(-\infty, 1.75)}(\theta^{(j)})$

```
## [1] 0.9
```

```
mean(theta.sim100  < 1.75)   # 100  MC draws
```

```
## [1] 0.94
```

```
mean(theta.sim1000 < 1.75)   # 1000 MC draws
```

```
## [1] 0.899
```

```
### -------------------------------------------------------------------------
### Posterior 95% credible interval, given by two symmetric quantiles
### -------------------------------------------------------------------------
qgamma(c(.025,.975),a+sy,b+n)                    #exact
```

We want now to compute posterior quantiles, because we want an interval estimate for θ_2

```
## [1] 1.173437 1.890836
```

```
quantile( theta.sim10,   c(.025,.975))          # empirical
```

```
##     2.5%    97.5%
## 1.260291 1.750068
```

```
quantile( theta.sim100,  c(.025,.975))          # empirical
```

```
##     2.5%    97.5%
## 1.231646 1.813752
```

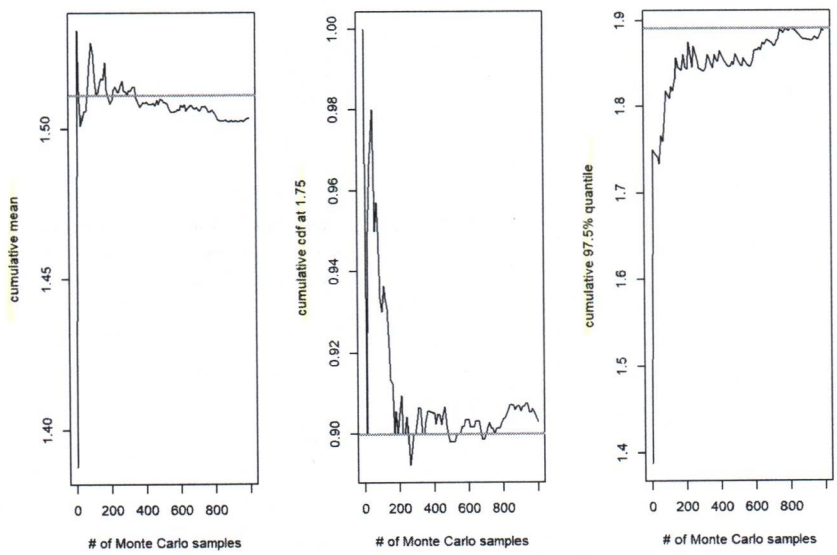
```
quantile( theta.sim1000, c(.025,.975))          # empirical
```

```
##     2.5%    97.5%
## 1.180194 1.892473
```

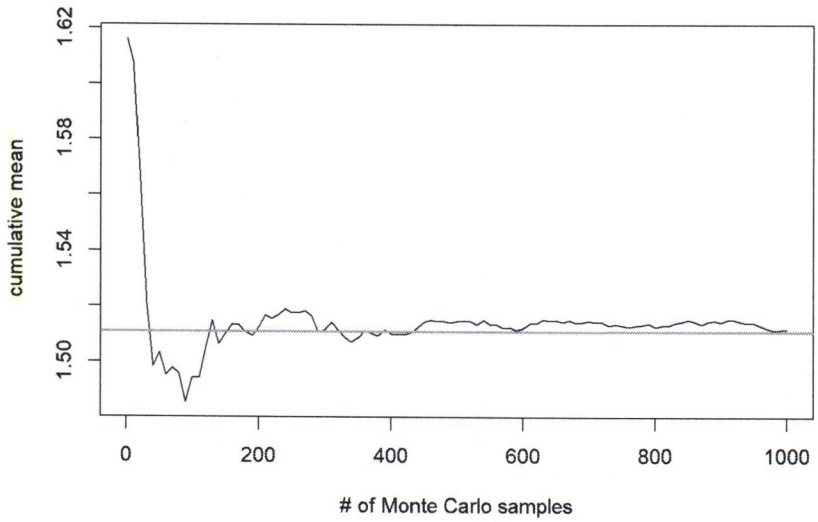
```
### ---------------------------------------------------------------------------
### Monte Carlo approximations as the sample size increases
### ---------------------------------------------------------------------------
x11()
par(mfrow=c(1,3), mar=c(2.75,2.75,.5,.5), mgp=c(1.70,.70,0))

set.seed(1)
a         <- 2
b         <- 1
sy        <- 66
n         <- 44
nsim      <- 1000
theta.sim <- rgamma(nsim,a+sy,b+n)
```

```
#cumulative mean
cmean <- cumsum(theta.sim)/(1:nsim)
cvar  <- cumsum(theta.sim^2)/(1:nsim) - cmean^2
ccdf  <- cumsum(theta.sim<1.75)/ (1:nsim)
cq    <-NULL
for(j in 1:nsim){ cq<-c(cq,quantile(theta.sim[1:j],probs=0.975)) }

sseq  <- c(1,(1:100)*(nsim/100))
cmean <- cmean[sseq]
cq    <- cq[sseq]
ccdf  <- ccdf[sseq]
```

```
# Plots ofthe MC estimates of the mean, the df at 1.75, the CI as the sample size M increases
# REM: there is no monotonicity!
par(mfrow=c(1,3))
plot(sseq,cmean,type="l",xlab="# of Monte Carlo samples",
     ylab="cumulative mean", col="black")
abline(h= (a+sy)/(b+n),col="gray",lwd=2)

plot(sseq,ccdf,type="l",xlab="# of Monte Carlo samples",
     ylab="cumulative cdf at 1.75", col="black")
abline(h= pgamma(1.75,a+sy,b+n),col="gray",lwd=2)

plot(sseq,cq,type="l",
     xlab="# of Monte Carlo samples",ylab="cumulative 97.5% quantile",col="black")
abline(h= qgamma(.975,a+sy,b+n),col="gray",lwd=2)
```



```
### ---------------------------------------------------------------------------
### Closer look at the MC computation of the posterior mean of theta2
### ---------------------------------------------------------------------------
a  <- 2
b  <- 1
sy <- 66
n  <- 44

nsim      <- 1000
theta.sim <- rgamma(nsim,a+sy,b+n)
sseq      <- c(1,(1:100)*(nsim/100))

cmean <- cumsum(theta.sim)/(1:nsim)
cmean <- cmean[sseq]

x11()
plot(sseq,cmean,type="l",xlab="# of Monte Carlo samples",ylab="cumulative mean",
     col="black")
abline(h= (a+sy)/(b+n),col="gray",lwd=2)
```




```
### ---------------------------------------------------------------------------
### Part II: MC COMPUTATION OF THE POSTERIOR PROBABILITY THAT theta1>theta2
### ---------------------------------------------------------------------------
set.seed(1)
a   <- 2
b   <- 1
sy1 <- 217 ;  n1 <- 111
sy2 <- 66  ;  n2 <- 44

a+sy1; b+n1
```

we're going to simulate (θ_1, θ_2) from the joint distribution. Here it's easy since the joint posterior factorizes.
⟹ we sample θ_1, θ_2 and then we consider the vector (θ_1, θ_2)

```
## [1] 219
```

```
## [1] 112
```

```
a+sy2; b+n2
```

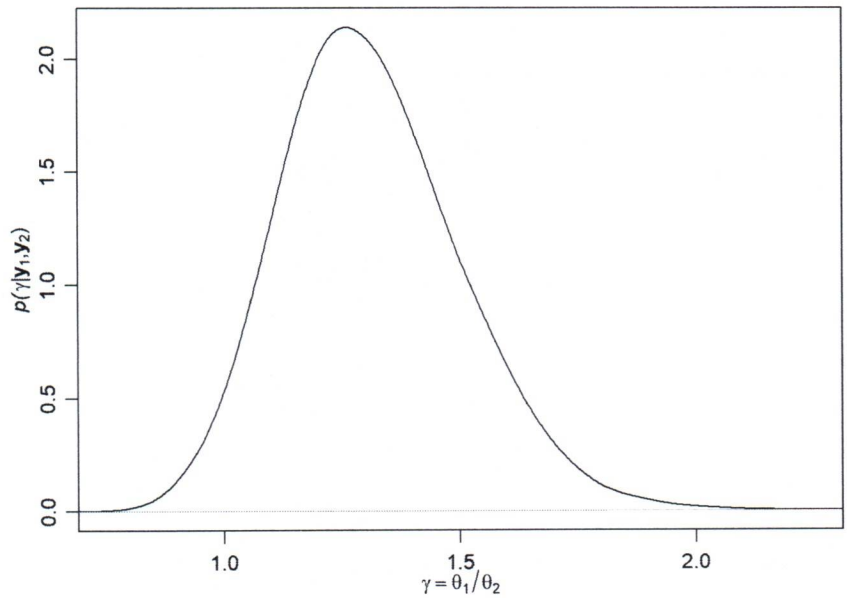
```
## [1] 68
```

```
## [1] 45
```

```
# MC sample of size M=10000 from the joint posterior
theta1.mc<-rgamma(10000,a+sy1, b+n1)
theta2.mc<-rgamma(10000,a+sy2, b+n2)

# The probability is estimated by the number of times in the sequence {(theta_1^(j),theta_2^(j)),j=1,2,...,M}
# the first component is > than the second, divided by M
mean(theta1.mc>theta2.mc)
```

```
## [1] 0.9708
```
(very high)

```
# questa probabilità è uguale alla prob che theta1 sia maggiore o UGUALE a theta2
# visto che (theta1,theta2) ha distribuzione assolutamente continua

# As an alternative, I can compute the posterior distribution of theta1/theta2;
# in yhis case, the analytic expression of its posterior is available
#pdf("fig4_4.pdf",family="Times",height=3.5,width=7)
x11()
par(mar=c(3,3,1,1),mgp=c(1.75,.75,0))
par(mfrow=c(1,1))
plot(density(theta1.mc/theta2.mc,adj=2),main="",xlim=c(.75,2.25),
xlab=expression(gamma==theta[1]/theta[2]),
ylab=expression(paste(italic("p("),gamma,"|",bold(y[1]),",",bold(y[2]),")",sep="")) )
```




```
### ------------------------------------------------------------------------
### Part III: COMPUTATION OF PREDICTIVE DISTRIBUTIONS
### ------------------------------------------------------------------------
# theta1.mc is a vector of size 10thousant from teh posterior distribution of theta1
y1.mc<-rpois(10000,theta1.mc)
y2.mc<-rpois(10000,theta2.mc)

mean(y1.mc>y2.mc)
```

(through composition method

predictive distr. that $Y_1 > Y_2$?

```
## [1] 0.4846
```

+

```
mean(y1.mc==y2.mc) # hence the posterior predictive prob. that Y_1 >= Y_2 is 0.69
```

we sum because the values are discrete (it changes having the > or ≥)

```
## [1] 0.2108
```

Note: also the joint predictive distribution factorizes in this case.

```
mean(y1.mc<y2.mc)
```

```
## [1] 0.3046
```

```
x11()
par(mar=c(3,3,1,1),mgp=c(1.75,.75,0))
par(mfrow=c(1,1))

diff.mc <- y1.mc-y2.mc
ds      <- -11:11
plot(ds,(table(c(diff.mc,ds))-1)/length(diff), type="h",lwd=3,
     xlab=expression(italic(D==tilde(Y)[1]-tilde(Y)[2])),
     ylab=expression(paste(italic("p(D)"),"|",bold(y[1]),",",bold(y[2]),")",sep="")))
```

children of a new woman of group 1

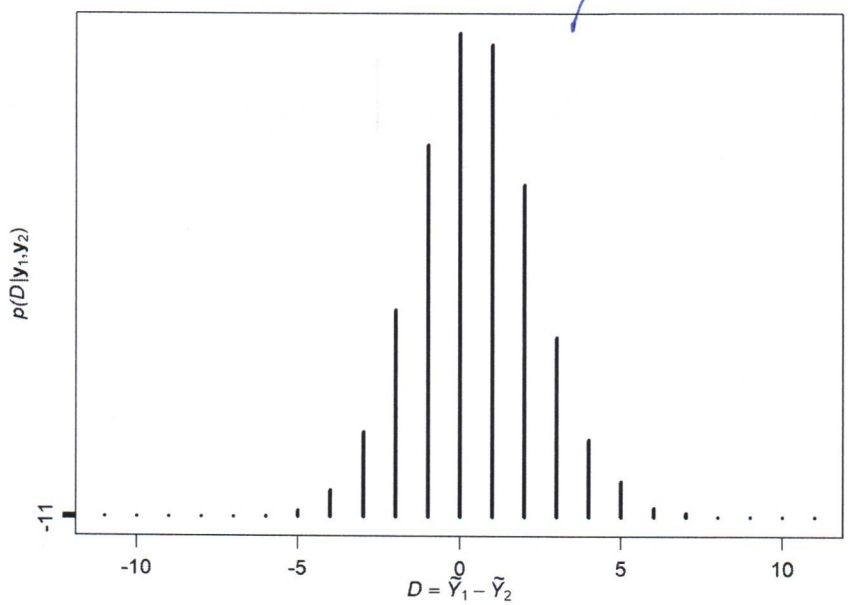
children .. group 2

$$\Upsilon\left(Y_{n_1+1,1}, Y_{n_2+1,2} \mid y_1, y_2\right) = \int_\Theta \Upsilon\left(Y_{n_1+1,1}, Y_{n_2+1,2} \mid \theta_1, \theta_2\right) \pi\left(\theta_1, \theta_2 \mid y_1, y_2\right) d\theta_1 d\theta_2$$

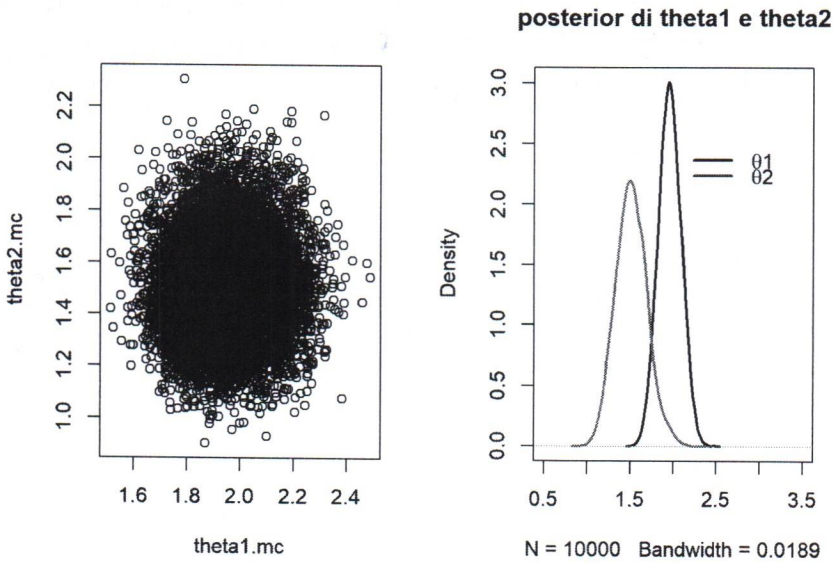
$$= \int_\Theta \Upsilon\left(Y_{n_1+1,1} \mid \theta_1\right) \Upsilon\left(Y_{n_2+1,2} \mid \theta_2\right) \pi\left(\theta_1 \mid y_1\right) \pi\left(\theta_2 \mid y_2\right) d\theta_1 d\theta_2$$

$$\Longrightarrow \Upsilon\left(Y_{n_1+1,1}, Y_{n_2+1,2} \mid y_1, y_2\right) = \Upsilon\left(Y_{n_1+1,1} \mid y_1\right) \cdot \Upsilon\left(Y_{n_2+1,2} \mid y_2\right)$$

predictive distribution of $Y_i|\theta_1 - Y_i|\theta_2$




```
### ---------------------------------------------------------------------------
### Marginali a posteriori di theta1 e di theta2
### ---------------------------------------------------------------------------
windows()
par(mfrow=c(1,2))
plot(theta1.mc,theta2.mc)
plot(density(theta1.mc),xlim=c(0.5,3.5),lwd=2,main='posterior di theta1 e theta2' )
lines(density(theta2.mc),main="",xlim=c(0.5,3.5),lwd=2, col=2)
legend(2,2.5,legend=c(expression(paste(theta,"1",sep="")),
                      expression(paste(theta,"2",sep=""))),
       lwd=c(2,2), col=c('black','red'),bty="n")
```




```
### ---------------------------------------------------------------------------
### Marginali predittive esatte - bionomiali negative
### ---------------------------------------------------------------------------
par(mfrow=c(1,2))
a+sy1; (b+n1)/(1+b+n1)  # parametri della predittiva marginale di Y_1
```

```
## [1] 219
```

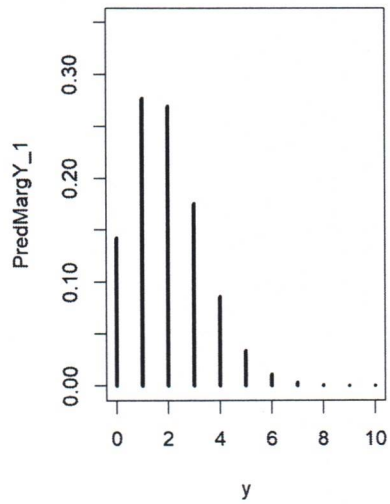
```
## [1] 0.9911504
```

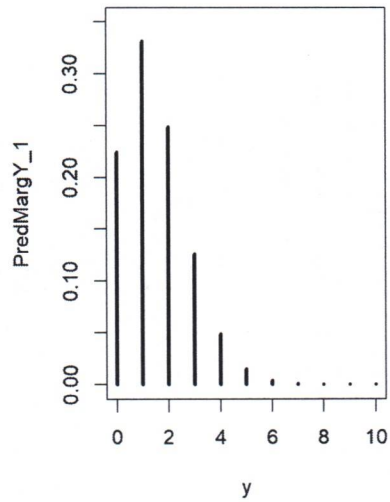
```
a+sy2; (b+n2)/(1+b+n2)  # parametri della predittiva marginale di Y_2
```

```
## [1] 68
```

```
## [1] 0.9782609
```

```
ds=seq(0,10)
plot(ds,dnbinom(ds,a+sy1, (b+n1)/(1+b+n1)), type="h",lwd=3,xlab='y', ylab='PredMargY_1',ylim=c(0,0.35))
plot(ds,dnbinom(ds,a+sy2, (b+n2)/(1+b+n2)), type="h",lwd=3,xlab='y', ylab='PredMargY_1',ylim=c(0,0.35))
```





What happens if we cannot simple iid from the posterior distribution?
We can build a Markov Chain whose limiting distribution is the posterior distribution and use a modification of the Ergodic theorem / SLLN (strong law of large Numbers) to approximate the integral w.r.t. the posterior distr. using MCMC method.

MC vs. MCMC?

the draw between two next values of the sequence are not ⫫ anymore as in MC (where we assume to sample iid)

Why do we simulate from the MCMC?
Because we can simulate the full posterior of our parameters through the MCMC, and use it to compute posterior estimates of the two components of the parameters simply considering the Ergodic mean. We could also be interested in the joint distribution.
Generally we know the posterior up to a normalizing constant, but, always generally, it's impossible to sample from the posterior (or compute mean, variance, ...) without knowing this constant.
With MCMC we're able to simulate and approximate the whole distr. of the parameters and hence we've also able to provide the predictive distribution.

Basic notions on General State Space Markov Chains

• discrete time
• discrete/continuous state space

Alessandra Guglielmi

Politecnico di Milano
Dipartimento di Matematica
Milano, Italia
e-mail: alessandra.guglielmi@polimi.it

16 October 2020

Homogeneous Markov chains on general spaces

E= state space, $E \subset \mathbb{R}^k$; $\mathcal{E} = \mathcal{B}(E)$

we suppose our parameter to be k-dimensional

Definition: $\{X_n, n \geq 0\}$ is a time-homogeneous Markov chain with values in E if $\{X_n, n \geq 0\}$ is a sequence of random elements with values in E such that

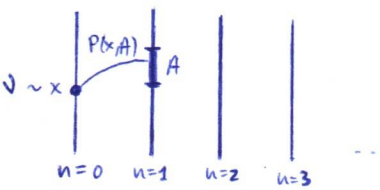
$$P(X_{n+1} \in A | X_0 = x_0, X_1 = x_1, \ldots, X_n = x_n) = P(X_{n+1} \in A | X_n = x_n)$$
$$=: P(x_n, A), \quad \forall A \in \mathcal{E}, \forall n, \forall x_0, x_1, \ldots, x_n \in E.$$

= conditioning to the present, the past and the future are ⊥⊥
$X_{n+1}|X_n,\ldots,X_1 \overset{d}{=} X_{n+1}|X_n$

$P(\cdot,\cdot)$ is the chain transition probability kernel (or transition probability), that is

(i) $x \mapsto P(x, A)$ is measurable $\forall A$

(ii) $A \mapsto P(x, A)$ is a probability on $(E, \mathcal{E})$ $\forall x \in E$



we fix x according to ν and then we jump at time 1 (generally at time n+1) according to P(x,A)

⇓

$\mathbb{P}(X_1 \in A) = \int_E P(x, A)\nu(dx)$

⇒ P(x,A) is a conditional distribution

Initial distribution of the chain

The chain is given when we fix:

- E state space
- P transition probability
- initial distribution ν of the chain (the law of X_0)

$$P(X_1 \in A) = \int_E P(X_1 \in A | X_0 = x)\nu(dx) = \int_E P(x, A)\nu(dx)$$

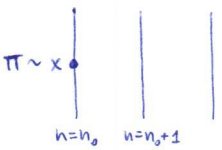
Invariant distribution

Very important: we will build MCMC chains which invariant distributions are the posteriors

Definition: π probability on E is called invariant or stationary for the MC $\{X_n, n \geq 0\}$ if

$$\int_E P(x, A)\pi(dx) = \pi(A) \quad \forall A$$

We write: $\pi P = \pi$



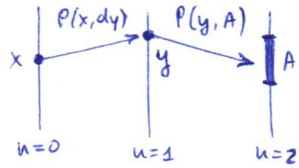
π is **invariant** if [we start at n_0 and we pick X according to π ⟹ the marginal at the next instant (n_0+1) is still π]

The n-step transition probability kernel

n-step transition probability kernel

$$P^n(x, A) = P(X_n \in A | X_0 = x)$$

$$n = 1 : P^1(x, A) = P(x, A)$$
$$n = 2 : P^2(x, A) = P(X_2 \in A | X_0 = x) = \int_E P(x, dy) P(y, A)$$
$$\dots\dots$$
$$n \geq 2 : P^n(x, A) = \int_E P(x, dy) P^{n-1}(y, A)$$



Notation:
$P_x(\cdot) := P(\cdot | X_0 = x)$

Key questions to our use of Markov chains

1. Over successive iterations, does the chain gravitate towards one state or another, or to a stable, equilibrium distribution over its state space?
2. If such distribution exists (also known as stationary, or invariant or limiting distribution), is it unique?
3. If a MC has a unique stationary distribution, how long does it take to get there after we start the MC at some arbitrary point in its state space?
4. How do we assess how close a MC has come to reaching its stationary distribution?
5. Can summaries of the trajectories of the MC be taken as summaries of the invariant distribution?

Irreducible Markov chains

If the state space is denumerable, a Markov chain is irreducible if every state i can be reached from any state j in a finite number of steps

Definition: ϕ probability on E

$\{X_n, n \geq 0\}$ is ϕ-irreducible if, for any A such that $\phi(A) > 0$, there exists $n = n(x, A) \geq 1$ with $P^n(x, A) > 0$ for any $x \in E$.

$\{X_n, n \geq 0\}$ is irreducible if there exists a probability ϕ w.r.t. the chain is ϕ-irreducible; ϕ is called irreducibility distribution.

The "standard" notion of irreducibility coincides when ϕ is the counting measure

If the chain is irreducible, it is NOT important where the chain starts from, since the chain will visit all remarkable states eventually (sooner or later).

Irreducible Markov chains

Criteria

« Return button: Check the assumptions

1. A MC with transition probability $P(x, A)$ is ϕ-irreducible if there exists $n \geq 1$ such that P^n has a (strictly) positive density f (w.r.t. ϕ) (Gibbs sampler)
2. If P has discrete and absolutely continuous components and there exists $n \geq 1$ such that the continuous component of P^n has a (strictly) positive density w.r.t. ϕ, then P is irreducible (Metropolis-Hastings)

REMARK: if $\{X_n, n \geq 0\}$ is irreducible, there exist different irreducibility distributions, but all are absolutely continuous w.r.t one of those, Ψ^*, which we call maximal irreducibility distribution

Recurrent Markov chains

sets whose measure is >0

If irreducibility means that all remarkable sets may be reached, recurrence means that these sets may be reached i.o. (infinitely often), at least from almost every initial point

$$X_n \in A \text{ i.o.} \Leftrightarrow \forall n\, \exists k \geq n : X_k \in A \Leftrightarrow \cap_{n=1}^{+\infty} \cup_{k \geq n} \{X_k \in A\}$$

Recurrent Markov chains

Definition: Let $\{X_n, n \geq 0\}$ be an irreducible MC with maximal irreducibility distribution Ψ^*. Then $\{X_n, n \geq 0\}$ is called recurrent if for all A such that $\Psi^*(A) > 0$:

- $P_x(X_n \in A \, i.o.) > 0$ for all x
- $P_x(X_n \in A \, i.o.) = 1$ Ψ^*-a.e. w.r.t. x.

If we start from almost any initial point x then $X_n \in A$ infinitely often with probability 1. The key point is "almost surely", that means that there are some subsets of X for which this condition does not hold, but in general for those points at least we have that this probability is >0.

Meaning: every *remarkable* set A is visited *infinitely often* with probability equal to 1, starting from *almost any* point x.

Definition: an irreducible and recurrent MC $\{X_n, n \geq 0\}$ is called positive recurrent if it admits an invariant probability; otherwise the chain is called null.

If the state space E is denumerable, an irreducible chain is recurrent if a state i (and therefore all the states) is recurrent, i.e. if $P_i(X_n = i \text{ i.o.}) = 1$.

Irreducibility + existence of a stationary distribution

Theorem

Let $\{X_n, n \geq 0\}$ be irreducible, and let π be a stationary distribution. Then

1. *$\{X_n, n \geq 0\}$ is π-irreducible and π is the maximal irreducibility distribution*
2. *π is the unique stationary distribution*
3. *$\{X_n, n \geq 0\}$ is positive recurrent.*

irreducible MC + $\exists$ at least one π (inv. distribution)

$\Rightarrow$
- $\{X_n\}_n$ is recurrent
- π is the unique distr.

Ergodic Theorem for MCs (Law of Large Numbers)

Theorem

Let $\{X_n, n \geq 0\}$ be an irreducible MC, with (unique) invariant distribution π. Let $f : E \to \mathbb{R}$ such that $E_\pi|f| := \int |f(x)|\pi(dx) < +\infty$. Then

$$P_x\left(\frac{1}{n+1}\sum_{i=0}^{n} f(X_i) \overset{n\to+\infty}{\longrightarrow} \int_E f d\pi\right) = 1 \quad \pi - a.e.\ w.r.t.\ x,$$

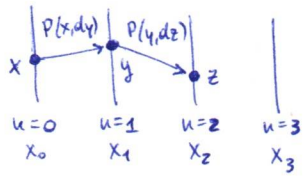
where $X_0 = x$ is the initial point of the MC.

If we start from almost all point x (initial point x) then the Ergodic mean of $f(X_i)$ will converge with probability 1 to the expectation of f ($:= \int_E f\, d\pi$).

REMARK: the limit (in the brackets) may NOT hold for initial points x belonging to a set C with $\pi(C) = 0$.

start from a "lucky point"

How can we simulate a Markov chain? If we start from x_0 and we're able to approximate the kernel we can!



we stop at X_m (m large enough).

$$\frac{1}{m+1}\sum_{i=0}^{m} f(X_i) \simeq \int f\,d\pi$$

with $f = \mathbb{1}_A$

$$\simeq \pi(A)$$

$$\Rightarrow \frac{\#\{X_1,\dots,X_m \in A\}}{m+1} \simeq \pi(A)$$

$\mathbb{P}(\theta \in A \mid \underline{x})$ posterior prob that $\theta \in A$

corresponding to the posterior probability that $\theta \in A$ (used for instance for credibility interval)

Application of the Ergodic Theorem (LLN)

AIM: compute $\pi(A)$ for a fixed A, where π is the target distribution. Here $f(x) = \mathbf{1}_A(x)$, and $\pi(A) = \int f d\pi$.

- Construct an irreducible MC $\{X_n, n \geq 0\}$ with invariant distribution π
- Consider a realization of this chain: $x_0, x_1, \dots, x_m$ for m LARGE and compute

$$\frac{\#\{i : x_i \in A\}}{m+1} = \text{relative frequency} =: \hat{\pi}(A)$$

This real number $\hat{\pi}(A)$ will be an approximation for the true value $\pi(A)$, if I've been lucky in picking up x_0!!!

Stronger convergence result

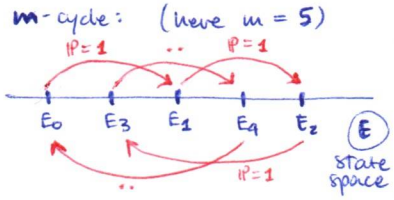
If I'd like to have a stronger convergence result (than convergence of the ergodic mean stated in the LLN) for $\mathcal{L}(X_n)$ to the limit distribution π, aperiodicity of the MC must be assumed.

Definition: Let $\{X_n, n \geq 0\}$ be irreducible; a m-cycle is a family of disjoint sets $\{E_0, E_1, \dots, E_{m-1}\}$ such that

$$P(x, E_j) = 1 \quad j = i+1 \mod m \quad \text{for all } x \in E_i.$$

d = period of the chain = the largest integer m for which an m-cycle exists

When $d = 1$, the chain is called aperiodic



once we're in a m-cycle we never get out.

Stronger result than Ergodic thm. →

Convergence in total variation a.s.-x

Theorem

Let $\{X_n, n \geq 0\}$ be irreducible and aperiodic, P its transition probability, π its invariant distribution. Then

$$\|P^n(x,\cdot) - \pi(\cdot)\| \to 0 \text{ as } n \to +\infty \quad \pi - \underline{a.s.}\ w.r.t.\ x.$$

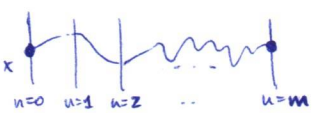
conditional law conditioned to the starting point x

Summing up, for almost every x, if the initial point of the MC is x,

$$\mathcal{L}(X_n | X_0 = x) \approx \pi \quad \text{for large } n$$

$$\|P_1 - P_2\| := \sup_{A \in \mathcal{E}} |P_1(A) - P_2(A)|$$

MCMC runs: aperiodicity is less important, since the interest is in computation of integrals. In these cases, it is enough to apply the LLN for MCs and use the ergodic mean $\frac{1}{n+1}\sum_{i=0}^{n} f(X_i)$ to approximate $\int f d\pi$, without assuming aperiodicity



At the m-th step, the conditional law (cond. w.r.t. the starting x) is:

$$P_x^m(x,\cdot) = \mathcal{L}(X_m | X_0 = x)$$

The theorem says:

$$\mathcal{L}(X_m | X_0 = x) \simeq \pi$$

This means that we can sample from π. If we simulate the chain a lot of times, every time starting from a different $X_0 = x$, we end up with many draws from $\mathcal{L}(X_m | X_0 = x)$. If these draws are independent we'll have approximately iid draws from π.

we want to get rid of the "for almost any"

Harris-recurrence

The Ergodic Theorem (LLN) **does NOT hold** when the initial point x of the chain is not the right one. However π-measure of the set of the initial points convergence does not hold for is null.

Rather, if I do not know π, how shall I be able to detect these wrong points, and avoid them?

Better to achieve stronger results than those we got so far, requiring they hold for all x. Of course, we need to assume a stronger property than recurrence.

Definition: Let $\{X_n, n \geq 0\}$ be irreducible, Ψ be its maximal irreducible distribution. Then $\{X_n, n \geq 0\}$ is Harris-recurrent if for all A such that $\Psi(A) > 0$:

$$P_x(X_n \in A \text{ i.o.}) = 1 \text{ for all } x \in E.$$

Ergodic Theorem for Harris-recurrent MCs

Theorem

Let $\{X_n, n \geq 0\}$ be an irreducible and Harris-recurrent MC, with invariant distribution π. Let $f : E \to \mathbb{R}$ be such that $E_\pi|f| < +\infty$. Then

$$P_x\left(\frac{1}{n+1}\sum_{i=0}^{n} f(X_i) \overset{n\to+\infty}{\longrightarrow} \int_E f d\pi\right) = 1 \text{ for all } x \in E.$$

Harris-ergodicity

Theorem

Let $\{X_n, n \geq 0\}$ be an aperiodic, irreducible, Harris-recurrent (=Harris-ergodic) MC, with invariant distribution π. Then

$$\|P^n(x,\cdot) - \pi(\cdot)\| \to 0 \text{ as } n \to +\infty \quad \forall x \in E.$$

BTW, the converse statement holds true too, so that Harris-ergodicity is equivalent to the limit above.

Basically, for all x

$$\mathcal{L}(X_n | X_0 = x) \approx \pi \quad \text{for large } n$$

we can approximate the posterior distribution with the conditional when n is large

How to use these theoretical results?

AIM: compute $E_\pi[g(\theta)|y_1,\ldots,y_n] = \int_\Theta g(\theta)\pi(\theta|y_1,\ldots,y_n)d\theta$

Suppose that θ has many components and we want to compute the expectation of any component of θ: $g(\theta_1,\ldots,\theta_K) = \theta_j$

Build (**simulate**) a MC $\{\theta_n, n \geq 0\}$ with state space Θ such that

- $\{\theta_n, n \geq 0\}$ is irreducible and Harris-recurrent;
- its (unique) invariant distribution is the posterior $\pi(\theta|y_1,\ldots,y_n)$.

the state space of the Markov chain is equal to the parametric space

Then estimate

$$E_\pi[g(\theta)|y_1,\ldots,y_n]$$

by the **ergodic mean** $\bar{g}_m = \frac{1}{m+1}\sum_{i=0}^{m} g(\theta_i)$

Markov Chain Monte Carlo methods (MCMC)

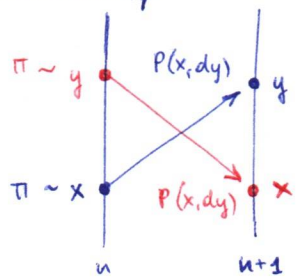
How shall we check these assumptions in the applications?

Irreducibility: criteria mentioned before (p. 13 ◂ Criteria) ; they apply to two core examples of MCMC algorithms, MH and Gibbs sampler

Harris-recurrence: these criteria not only guarantee irreducibility, but also Harris-recurrence for MH and Gibbs sampler

$\pi(\theta|y_1,\ldots,y_n)$ **is the target density, i.e. the invariant distribution of the MC**: we build a MC that is reversible wrt the posterior density

Reversibility:



the probability of picking $x \sim \pi$ and ending in y is the same as picking $y \sim \pi$ and ending in x

$\pi(dx)P(x,dy) = \pi(dy)P(y,dx)$

Reversibility

Definition: $\{X_n, n \geq 0\}$ MC with transition probability P. The chain is reversible wrt a probability π on E when

$$\pi(dx)P(x,dy) = \pi(dy)P(y,dx) \quad x, y \in E.$$

(kind of a symmetry condition)

Basically, the probability of the chain being in x and moving to y is equal to the probability of the chain being in y and coming back in x. Equivalently, if the chain $\{X_n\}$ is stationary, reversibility means that $\mathcal{L}(X_n, X_{n+1}) = \mathcal{L}(X_{n+1}, X_n)$

When π and P have densities, reversibility is equivalent to

$$\pi(x)p(x,y) = \pi(y)p(y,x) \quad \forall x, y \in E.$$

Reversibility wrt π implies that π is invariant

Theorem

If $\{X_n, n \geq 0\}$ is reversible wrt probability π, then π is invariant for the chain.

Proof.

$$\int_{x\in E} \pi(dx)P(x,dy) \overset{\text{reversibility}}{=} \int_{x\in E} \pi(dy)P(y,dx) = \pi(dy) \int_{x\in E} P(y,dx)$$
$$= \pi(dy)1 = \pi(dy),$$

that is π is stationary for the chain. □

$\pi(A) = \int_E \pi(dx)\, P(x,A) \quad \forall A \in \mathcal{B}(E)$

$\pi(dy) = \int_E \pi(dx)\, P(x,dy)$

Geometric and uniform ergodicity

If we aim at guaranteeing the CLT for MCs (in order to get a MC error bound), we need stronger assumptions for the MC.

$\mathcal{L}(X_n | X_0 = x) \approx \pi$

Definition: Let $\{X_n, n \geq 0\}$ be Harris-ergodic, π the invariant distribution. Then the chain is called geometrically ergodic if there exist $M : E \to \mathbb{R}^+$ con $E_\pi(M) < +\infty$ and $r \in (0,1)$ such that

$$\|P^n(x,\cdot) - \pi(\cdot)\| \leq M(x)r^n \quad \forall x \in E, \forall n.$$

When $M(x) = cost$, then $\{X_n, n \geq 0\}$ is called uniformly ergodic.

REM: if the MC is geometrically or uniformly ergodic, then it is Harris-ergodic, i.e.

$$\|P^n(x,\cdot) - \pi(\cdot)\| \to 0 \text{ as } n \to +\infty \quad \forall x \in E,$$

i.e. the MC is aperiodic, irreducible and Harris-recurrent, with invariant distribution π.

for the convergence of the chain

The Central Limit Theorem for MCs

Theorem

Let $\{X_n, n \geq 0\}$ be an Harris-ergodic MC with invariant distribution π and $f : E \to \mathbb{R}$ such that one of these conditions holds:

(a) $\{X_n, n \geq 0\}$ is geo-ergodic and $\mathbb{E}_\pi(|f|^{2+\epsilon}) < +\infty$ for some $\epsilon > 0$;

(b) $\{X_n, n \geq 0\}$ is unif-ergodic and $\mathbb{E}_\pi(f^2) < +\infty$.

Then:

$$\sqrt{n}\left(\bar{f}_n - \int f d\pi\right) \xrightarrow{d} \mathcal{N}(0, \sigma_f^2), \quad n \to +\infty,$$

where $\bar{f}_n = \frac{1}{n+1}\sum_{i=0}^{n} f(X_i)$ and

$$\sigma_f^2 = \mathrm{Var}_\pi(f(X_0)) + 2\sum_{k=1}^{+\infty} \mathrm{Cov}_\pi(f(X_0), f(X_k)), \; X_0, X_k \text{ marginal. } \sim \pi.$$

Suppose we want to approximate: $\int_\Theta g(\underline{\theta})\, \pi(d\underline{\theta}|\underline{x})$

$$\Rightarrow \boxed{\int_\Theta g(\underline{\theta})\, \pi(d\underline{\theta}|\underline{x}) \approx \frac{1}{n+1}\sum_{i=0}^{n} g(\underline{\theta}^{(i)})}$$

we know that the error of the approximation goes to zero, but how fast?

$$\left[\frac{1}{n+1}\sum_{i=0}^{n} g(\underline{\theta}^{(i)}) - \int_\Theta g(\underline{\theta})\, \pi(d\underline{\theta}|\underline{x})\right] \xrightarrow{d} N\left(0, \frac{\sigma_g^2}{n}\right)$$

ACCEPT / REJECT ALGORITHM

$\pi(\theta)$ density

We want to simulate from $\pi(\theta)$ but (suppose) we can't calculate the distribution function of this density.

Suppose that: $\pi(\theta) < c \cdot \underbrace{m(\theta)}_{\text{density from which we know how to simulate}}$ $\quad c>0, \ \forall \theta \in \Theta$

Algorithm:

1. $z \sim m \longrightarrow z \in \Theta \longrightarrow r := \dfrac{\pi(z)}{c \cdot m(z)} \in (0,1)$ (since $\pi(\theta) < c\,m(\theta)$ implies: $\frac{\pi(\theta)}{c\,m(\theta)} < 1$)
2. $U \sim \mathcal{U}([0,1]) \longrightarrow u \in (0,1)$
3. If $u \le r \Longrightarrow$ accept z and put $\tilde{\theta} = z$
   $u > r \Longrightarrow$ reject z and go to 1.

Notice: we can evaluate $\pi(z)$ because we can't sample from $\pi(\theta)$ but we know its form

$\Longrightarrow \hat{\theta} \sim \pi$

$\hat{\theta}$ so produced

proof. $(\tilde{\theta} \sim \pi)$

Recall: $\Theta \subseteq \mathbb{R}$, $U \sim \mathcal{U}([0,1])$, $z \sim m$

$$\mathbb{P}(\tilde{\theta} \le t) = \mathbb{P}\left(z \le t \mid U \le \frac{\pi(z)}{c \cdot m(z)}\right)$$

$$= \frac{\mathbb{P}\left(z \le t, \ U \le \frac{\pi(z)}{c \cdot m(z)}\right)}{\mathbb{P}\left(U \le \frac{\pi(z)}{c \cdot m(z)}\right)}$$

$$= \frac{\int_\Theta \mathbb{P}\left(z \le t, U \le \frac{\pi(z)}{c \cdot m(z)} \mid z = z\right) m(z)\, dz}{\int_\Theta \mathbb{P}\left(U \le \frac{\pi(z)}{m(z) \cdot c} \mid z = z\right) m(z)\, dz}$$

Once we condition to $z=z$ we have that U and z are $\perp\!\!\!\perp$

$$= \mathbb{P}\left(U \le \frac{\pi(z)}{c\,m(z)} \mid z = z\right) = \mathbb{P}\left(U \le \frac{\pi(z)}{c\,m(z)}\right)$$

$$= \mathbb{P}(z \le t \mid z = z) = \mathbb{P}(z \le t)$$

$$= \frac{\int_\Theta \mathbb{1}_{(-\infty, t)}(z) \frac{\pi(z)}{c \cdot m(z)} \cdot m(z)\, dz}{\int_\Theta \frac{\pi(z)}{c \cdot m(z)} \cdot m(z)\, dz}$$

$$= \int_{-\infty}^{t} \pi(z)\, dz = F_\pi(t) \Longrightarrow \tilde{\theta} \sim \pi$$

Suppose $\Theta = \mathbb{R}$ ■