I - Introduction

Consider the following simple problem:

A bag contains a ball of unknown colour, which may be either black or white. A white ball is added to the bag, then a ball is drawn at random from it. The drawn ball happens to be white. What is the colour of the ball that was initially in the bag?

A natural way to solve this problem is to proceed as follows:

- Let C_I be the colour of the ball that was initially in the bag and assume that $\mathbb{P}(C_I = w) = \mathbb{P}(C_I = b) = 1/2$.
- Denoting by C_D the colour of the drawn ball, we have

$$\mathbb{P}(C_I = w|C_D = w) = \frac{2}{3}.$$

- Then, we answer to the above question by saying that with probability 2/3 a white ball was initially in the bag.

Lessons from this problem

Most of you would have probably performed the same computations. However, these latter contain the two main ideas underlying Bayesian inference:

1. Prior uncertainty/prior beliefs ("Of which colour is the initial ball?") can be expressed in terms of probabilities:

$$\mathbb{P}(C_I = w) = \mathbb{P}(C_I = b) = 1/2.$$

2. Taking into account any new information that arises from the experiment ("the drawn ball is white") can be done by writing conditional probabilities:

$$\mathbb{P}(C_I = w|C_D = w) = 2/3.$$

The fact that we accept almost unnoticeably these concepts prove they are natural and convenient. Therefore we are all Bayesian!

Frequentist versus Bayesian interpretation of probabilities

From a frequentist perspective the probability of an event is the limit of its relative frequency as the number of trials goes to infinity.

From a Baysian perspective the probability of an event represents our reasonable expectation about it in a single trial.

To illustrate this difference of interpretation let X denotes the output of the roll of a fair dice (with 6 sides numbered from 1 to 6). Then, both schools agree that $\mathbb{P}(X=1)=1/6$ but

- For the frequentist school $\mathbb{P}(X=1)=1/6$ because, if we roll the dice n times and denote by x_i the outcome of the i-th roll, we have $\frac{1}{6}=\lim_{n\rightarrow+\infty}\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{\{1\}}(x_i)$.
- For the Bayesian school $\mathbb{P}(X=1)=1/6$ because all the 6 sides of the dice are equally likely to face upwards (since the dice is fair).

The frequentist view of the introductory example

Since C_I is deterministic (i.e. the color of the ball that was initially in the bag is maybe unknown but whether it is white or black is now part of the ‘state of the world’) the above computations do not make any sense from a frequentist point of view.

Bayesian formulae

Bayesian statistics is named after Reverend Thomas Bayes (1701-1761), who discovered the Bayes formula:

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A|B)\mathbb{P}(B)}{\mathbb{P}(A|B)\mathbb{P}(B) + \mathbb{P}(A|B^c)\mathbb{P}(B^c)},$$

where A and B are two events.

This formula allows for deducing the relation $A \to B$ (A gives B) from its opposite $B \to A$ (B gives A) and from the prior information $\mathbb{P}(B)$.

Actually, Thomas Bayes proved a continuous version of the above formula: Given random variables X and Y, with conditional distribution $f(x|y)$ and marginal distribution $g(y)$, the conditional distribution of y given x is

$$g(y|x) = \frac{f(x|y)g(y)}{\int f(x|y')g(y')\mathrm{d}y'}.$$

Historical remark: Pierre Simon de Laplace (1749-1827) rediscovered independently Bayes formula and can be considered as the second Bayesian scientist in History.

The Bayesian viewpoint in Statistical Science

Bayes and Laplace went one step further and consider that the uncertainty about the parameter θ of a parametric model $\{f(\cdot|\theta),\, \theta \in \Theta\}$ can be expressed through a probability distribution $\pi(\theta)$ on Θ.

Then, it is possible to apply the Bayes formula to the couple (θ, x):

$$\boxed{\pi(\theta|x) = \frac{\pi(\theta)f(x|\theta)}{m(x)}}$$

Bayesian terminology:

1. $\pi(\theta)$ is called prior distribution of θ,
2. $f(x|\theta)$ is called the likelihood of the model,
3. $\pi(\theta|x)$ is called the posterior distribution of θ,
4. $m(x) = \int_\Theta f(x|\theta)\pi(\theta)\mathrm{d}\theta$ is called the marginal density of x.

We can now provide a formal definition of Bayesian models:

Definition 1.1 *A Bayesian statistical model consists of a parametric statistical model $\{f(\cdot|\theta),\, \theta \in \Theta \subseteq \mathbb{R}^d\}$ and a prior distribution $\pi(\theta)$ on Θ.*

Example: Laplace measuring the mass of Saturn

Around 1812-1816 Laplace used Bayes formula to estimate the mass of Saturn given the available astronomical knowledge and data.

More precisely, he estimated the mass of Saturn applying the Bayes formula

$$\mathbb{P}(M|D,I) = \frac{\mathbb{P}(D|M,I)\mathbb{P}(M|I)}{\mathbb{P}(D|I)}$$

where

- M is the mass of Saturn
- D is the data from various measurements of planetary orbits
- I is the existing background information, e.g., Newtons laws of celestial mechanics.

In this example M is the parameter of interest, $\mathbb{P}(D|M,I)$ the likelihood and $\mathbb{P}(M|I)$ the prior distribution of M.

Remarkably, Laplace's estimated value of the mass of Saturn only differs from the modern value measured by the NASA by about 0.5%.

Remark: Probabilistic statements about M, the mass of Saturn, do not have any sense from a frequentist perspective.

General framework

In this course we will often replace distribution with probability density function (pdf), assuming that this latter is well defined with respect to a dominating measure (e.g. Lebesgue or counting measure).

In particular,

- The dominating measure for the model $\{f(\cdot|\theta),\, \theta \in \Theta \subseteq \mathbb{R}^d\}$ will be denoted by $\mathrm{d}x$;
- The dominated measure for the prior distribution will be denoted by $\mathrm{d}\theta$.

We denote by $\mathcal{X} \subseteq \mathbb{R}^k$ the observation space so that, for any $\theta \in \Theta$, $f(\cdot|\theta)\mathrm{d}x$ is a probability distribution on $\mathcal{X}$.

Following the standard convention, we will use capital letters for random variables and small letters for their realizations. For instance, X is a random variable and x is a realization of X.

However, in the Bayesian literature this standard convention is not used for the parameter θ (the notation Θ is already used for the parameter space!). Hence, throughout this course the notation θ is used for both the random variable and its realizations.

Back to the introductory example

Putting the introductory example into our general framework yields

- $\Theta = \{w, b\}$
- $\mathcal{X} = \{w, b\}$ and $X = C_D$
- $\{f(\cdot|\theta),\, \theta \in \Theta\}$ defined by

$$f(x|w) = \begin{cases} 1, & x = w \\ 0, & x = b \end{cases}, \quad f(x|b) = \begin{cases} \frac{1}{2}, & x = w \\ \frac{1}{2}, & x = b \end{cases}$$

- $\pi(\theta) = \frac{1}{2}\mathbf{1}_{\{w\}}(\theta) + \frac{1}{2}\mathbf{1}_{\{b\}}(\theta)$
- $\pi(\theta|x) = \mathbb{P}(C_I = \theta | C_D = x)$
- $\mathrm{d}x = \delta_{\{w\}}(\mathrm{d}x) + \delta_{\{b\}}(\mathrm{d}x)$
- $\mathrm{d}\theta = \delta_{\{w\}}(\mathrm{d}\theta) + \delta_{\{b\}}(\mathrm{d}\theta)$

Direct application of Bayes formula gives

$$\pi(w|x) = \frac{\pi(w)f(x|w)}{\int_\Theta f(x|\theta)\pi(\theta)\mathrm{d}\theta} = \frac{\frac{1}{2}\mathbf{1}_{\{w\}}(x)}{\frac{1}{2}f(x|b) + \frac{1}{2}f(x|w)} = \frac{\frac{1}{2}\mathbf{1}_{\{w\}}(x)}{\frac{1}{4} + \frac{1}{2}\mathbf{1}_{\{w\}}(x)}$$

and thus, as per above (!),

$$\mathbb{P}(C_I = w | C_D = w) = \pi(w|w) = \frac{\frac{1}{2}}{\frac{1}{4} + \frac{1}{2}} = \frac{2}{3}.$$

Some remarks about the general framework

The notation x for the observation and $f(x|\theta)$ for the likelihood that we will use throughout this course is quite general.

For instance, consider n observations $x_1, \ldots, x_n$. Then, we have $x := (x_1, \ldots, x_n)$ and

- If the observations $x_1, \ldots, x_n$ are modelled as n i.i.d. random variables taking values in $\mathcal{X}_1$ and if $\{\tilde{f}(\cdot|\theta),\, \theta \in \Theta \subseteq \mathbb{R}^d\}$ is the parametric model for observation x_1, then we have

$$\mathcal{X} = \mathcal{X}_1^n, \quad f(x|\theta) = \prod_{k=1}^{n} \tilde{f}(x_k|\theta).$$

- More generally, if the parametric model for x is defined by

$$\begin{aligned} &X_1|\theta \sim \tilde{f}_1(x_1|\theta)\mathrm{d}x_1 \\ &X_k|\big(X_1 = x_1, \ldots, X_{k-1} = x_{k-1}, \theta\big) \sim \tilde{f}_k(x_k|x_1, \ldots, x_{k-1}, \theta)\mathrm{d}x_k \end{aligned}$$

with $\theta \in \Theta$ and $x_k \in \mathcal{X}_k$, then we have

$$\mathcal{X} = \prod_{k=1}^{n} \mathcal{X}_k, \quad f(x|\theta) = \tilde{f}_1(x_1|\theta) \prod_{k=2}^{n} \tilde{f}_k(x_k|x_1, \ldots, x_{k-1}, \theta).$$

Example: Gaussian model with unknown mean

Let $\phi(\cdot)$ be the density of the $\mathcal{N}_1(0,1)$ distribution.

Proposition 1.1 *Let $(\sigma, \sigma_0, \mu_0) \in \mathbb{R}_{>0} \times \mathbb{R}_{>0} \times \mathbb{R}$ and consider the Bayesian statistical model defined by*

$$\pi(\theta) = \frac{1}{\sigma_0}\phi\Big(\frac{\theta - \mu_0}{\sigma_0}\Big), \quad f(x|\theta) = \prod_{k=1}^{n} \frac{1}{\sigma}\phi\Big(\frac{x_k - \theta}{\sigma}\Big),$$

where $n \in \mathbb{N}$ and $x = (x_1, \dots, x_n) \in \mathbb{R}^n$. Then,

$$\pi(\theta|x) = \frac{1}{\sigma_n}\phi\Big(\frac{\theta - \mu_n}{\sigma_n}\Big)$$

with

$$\mu_n = \lambda_n \mu_0 + (1 - \lambda_n)\bar{x}_n, \quad \sigma_n^2 = \frac{\sigma^2}{n_0 + n}$$

where

$$\bar{x}_n = \frac{1}{n}\sum_{k=1}^{n} x_k, \quad \lambda_n = \frac{n_0}{n_0 + n} \in (0,1), \quad n_0 = \frac{\sigma^2}{\sigma_0^2}.$$

Proof: Done in class

Comments:

- The posterior mean μ_n is a weighted average between the empirical mean $\bar{x}_n$ and the prior mean μ_0. The posterior distribution is therefore a compromise between the prior information and the information carried by the observations.
- When n is large, $\mu_n \approx \bar{x}_n$ and $\sigma_n^2 \approx \sigma^2/n$ so that the impact of the prior distribution becomes less important as n increases.

Gaussian model with unknown mean

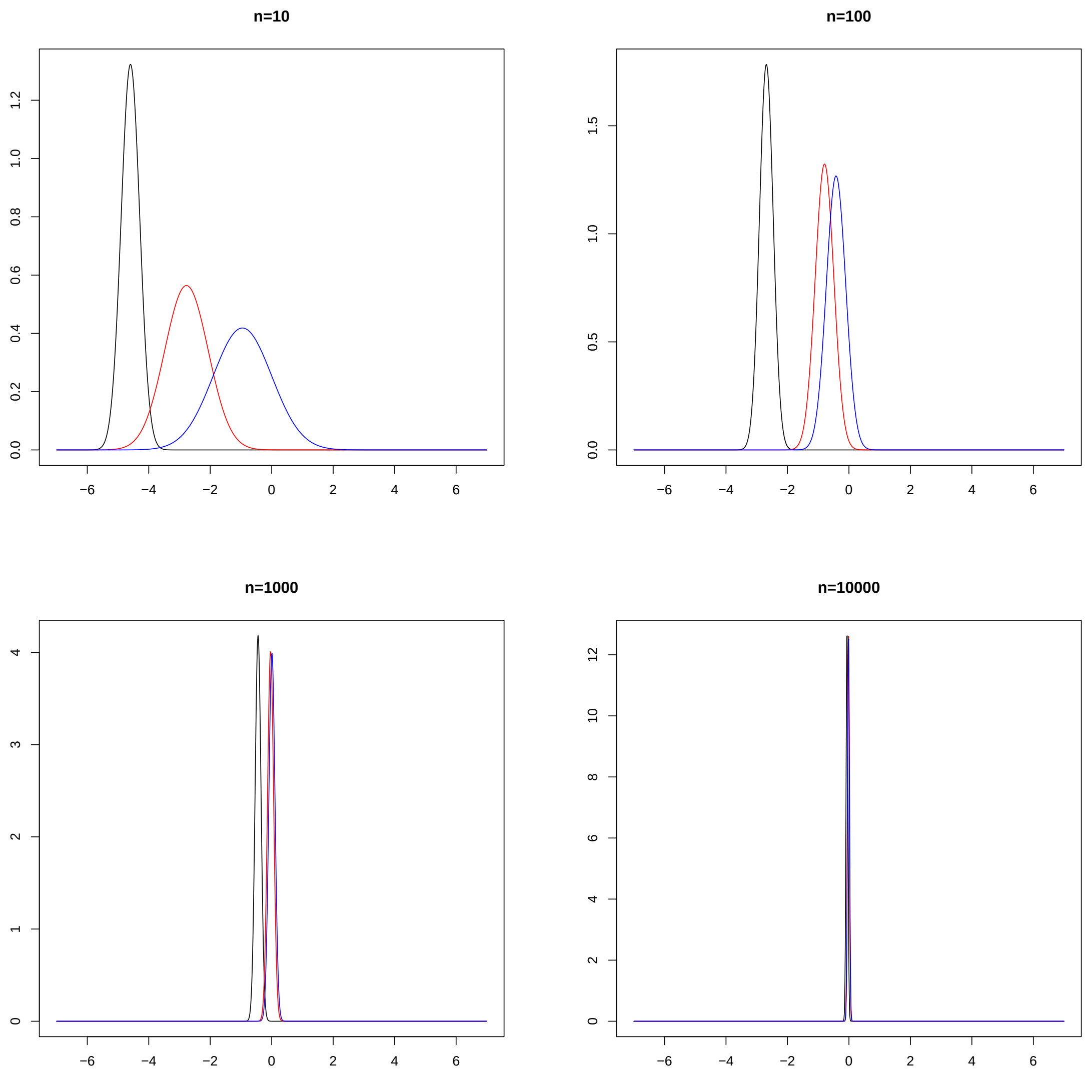


Figure 1: Posterior distribution for $n \in \{10, 100, 1\,000, 10\,000\}$ and for $\sigma_0^2 = 0.1$ (black), $\sigma_0^2 = 1$ (red) and $\sigma_0^2 = 10$. The observations are realizations of i.i.d. $\mathcal{N}_1(0, 10)$ random variables and $\mu_0 = 5$.

General comments about Bayesian inference

1. In the Bayesian framework, all the inference about θ is carried out
   (a) Conditionally on the observation x;
   (b) Uniquely through the posterior distribution $\pi(\theta|x)$.

   In particular,

   - Point estimators of θ are derived from $\pi(\theta|x)$ and are justified only for the current observation x (**Chapter 2**).
   - Hypotheses testing (**Chapter 4**) and model choice (**Chapter 5**) are fully based on $\pi(\theta|x)$ (and not on some asymptotic arguments).

2. As illustrated with the Gaussian example, for a fixed sample size n the choice of the prior distribution can have an important impact on the posterior distribution. The choice of a "good" prior distribution will be the object of **Chapter 3**.

3. As illustrated with the Gaussian example, as the number of observations n increases the impact of a fixed prior distribution becomes less and less important. In particular, as $n \to +\infty$, the posterior distribution concentrates around the true parameter value. Results on Bayesian asymptotics will be presented and derived in **Chapter 6**.

Computational aspects

By definition,

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{m(x)}, \quad m(x) = \int_\Theta f(x|\theta)\pi(\theta)\mathrm{d}\theta$$

where often (but not always!) $f(x|\theta)$ and $\pi(\theta)$ are known.

Assuming this is the case, three scenarios are possibles:

- The quantity $m(x)$ can be explicitly computed. This happens only when the model $\{f(\cdot|\theta),\, \theta \in \Theta\}$ admits a conjugate prior.
- The quantity $m(x)$ cannot be explicitly computed but can be approximated by a quadrature rule. In this case,

  $$m(x) \approx \sum_{i_1}^{N} w_1^{i_1} \sum_{i_2}^{N} w_2^{i_2} \cdots \sum_{i_d}^{N} w_d^{i_d} f(x|\theta_1^{i_1}, \dots, \theta_d^{i_d})\pi(\theta_1^{i_1}, \dots, \theta_d^{i_d})$$

  where the $w_j^{i_j}$'s, the $\theta_j^{i_j}$'s and N depend on the quadrature rule and on the desired level of approximation. This method is only applicable when d is small (say $d \leq 2$) since the cost to reach a fixed approximation error is exponential in d.
- The quantity $m(x)$ cannot be explicitly computed and $d > 2$ (say). In this case, we can use Monte Carlo methods to approximate $\pi(\theta|x)$; that is, we generate a random weighted sample $(W_N^1, \theta_N^1), \dots, (W_N^N, \theta_N^N)$ such that

  $$\lim_{N\to+\infty} \sum_{i=1}^{N} W_N^i \delta_{\theta_N^i}(\mathrm{d}\theta) = \pi(\theta|x)\mathrm{d}\theta, \quad \text{(almost surely)}.$$

  In **Chapter 7** we will introduce the Metropolis-Hastings algorithm which is arguably the most popular Monte Carlo algorithm to approximate the posterior distribution in this scenario (which is from far the most frequent in practice).

Why do we need Bayesian statistics?

- As Bayesian statistics is based on non-asymptotic arguments it is particularly relevant when the sample size is small (e.g. because collecting observations is costly).
- In some cases (e.g. when the sample size is small) we really want to incorporate some prior information in the analysis.
- Often in science it is easy to predict the outcome given the cause while deducting the cause of a given outcome is much harder. Bayesian theory provides a unified framework to address this problem.
- Frequentist methods are justified as the number of observations goes to infinity and/or according to their average performance over an infinite number of samples. In some cases there exist only a limited number of observations (e.g. there exists only one planet Saturn) so that this kind of justifications does not make any sense.
- Computational reasons: As we will see in Chapters 2 and 6 Bayesian methods have very good frequentist properties while, in some cases, computing Bayesian estimators is easier than computing frequentist estimators.
- In some cases taking a Bayesian approach facilitates the modelling process because specifying the joint distribution $f(x, \theta)$ is simpler than specifying the conditional distribution $f(x|\theta)$. The typical example where this is true is when the observations are similar but not identical (see **Chapters 8 and 9**).
- Philosophical reasons...

A second historical example (Laplace, 1786)

In Paris, $n_m = 251\,527$ male births and $n_f = 241\,945$ female births was recorded in 1786. Using these data, Laplace wanted to test if the probability $\theta \in \Theta := [0,1]$ of a male birth is above $1/2$.

- Let $x = n_m$ be the observation and $n = n_m + n_f$. Then, the underlying statistical model can be stated as follows:
$$X \sim \mathrm{Bin}(n, \theta).$$
- Laplace had no substantial prior information on the value of θ. Therefore he decided to assign to θ an uniform prior distribution on Θ,
$$\pi(\theta) = \mathbf{1}_{[0,1]}(\theta),$$
in order to represent the idea that ‘a priori all values of θ are equally likely’.
- The corresponding posterior distribution of θ given $X = x$ is the $\mathrm{Beta}(x+1, n-x+1)$ distribution; that is
$$\pi(\theta|x) = \frac{\Gamma(n+2)}{\Gamma(x+1)\Gamma(n-x+1)}\theta^x(1-\theta)^{n-x}, \quad \theta \in [0,1].$$
- Using some numerical integration routine, we get
$$\pi(\{\theta : \theta > 1/2\}|x) = 1 - \int_0^{0.5} \pi(\theta|x)\mathrm{d}x \approx 1 - 1.15 \times 10^{-42}.$$
- From this result Laplace deduced that θ is more likely to be above $1/2$.

II - Decision theory and Bayesian inference

We saw in the previous chapter that Bayesian inference is based on the following two principles:

- We can express our ignorance/information about the unknown parameter $\theta \in \Theta$ by a probability distribution $\pi(\theta)$ on Θ.
- We can use the Bayes rule and the likelihood $f(x|\theta)$ of an observation $x \in \mathcal{X}$ to update our prior knowledge about θ.

The first output of Bayesian inference is therefore the posterior distribution $\pi(\theta|x) \propto f(x|\theta)\pi(\theta)$.

However, we often want to derive an estimator of the unknown parameter θ; that is, a point $\widehat{\theta}$ in the parameter space that approximates θ in some sense.

The "most natural" estimators we can derive from $\pi(\theta|x)$ are:

- the posterior mean;
- the posterior median;
- the posterior mode (also called MAP for maximum a posteriori).

The goal of this chapter is to justify (or not!) these estimators from a decision theoretic perceptive.

Decision theory: General framework

Let $\mathcal{D}$ be the set of all possible decisions.

Definition 2.1 *A loss function is any function* $L : \Theta \times \mathcal{D} \to [0, +\infty)$.

Definition 2.2 *A decision rule is any mapping* $\delta : \mathcal{X} \to \mathcal{D}$.

For $\theta \in \Theta$ and $x \in \mathcal{X}$, the quantity $L(\theta, \delta(x))$ therefore gives the cost induced by the decision rule δ when we observe x.

The question of interest is then the following:

Given a loss function L, what is the "optimal" decision rule δ?

In this chapter we mainly focus on the scenario $\mathcal{D} = \Theta$. In this case, δ is an estimator of the unknown parameter θ and the Bayesian answer to the above question for different choices of loss L leads to the derivation of different Bayesian estimators of θ.

Remark: Often we address the reverse question; that is, for which (if any) loss function L is the decision rule δ optimal? This is helpful to understand in what sense δ is a good decision rule.

The frequentist approach

The frequentist approach considers the frequentist risk

$$R(\theta, \delta) = \int_{\mathcal{X}} L(\theta, \delta(x)) f(x|\theta) \mathrm{d}x.$$

Because the frequentist risk is a function of θ there exists in general no decision rule δ such that, for every decision rule $\delta' \neq \delta$, we have

$$R(\theta, \delta) \leq R(\theta, \delta'), \quad \forall \theta \in \Theta.$$

Consequently, the frequentist risk alone is not sufficient to select a particular decision rule and other criteria are needed to choose δ.

For instance, in the frequentist approach we can

1. Select the decision rule δ which minimizes the frequentist risk on a given restricted set of decision rules (e.g. the set of unbiased and linear decision rules).

2. Select a decision rule δ which is minimax; that is, such that for every decision rule $\delta' \neq \delta$ we have

$$\max_{\theta \in \Theta} R(\theta, \delta) \leq \max_{\theta \in \Theta} R(\theta, \delta').$$

3. Select a decision rule δ which is admissible.

Definition 2.3 *A decision rule δ is admissible if there exists no decision rule δ' such that*

$$R(\theta, \delta') \leq R(\theta, \delta), \quad \forall \theta \in \Theta$$

with the above inequality being strict for at least one $\theta \in \Theta$.

Admissibility and Stein's result

Admissibility seems to be a weak requirement for an estimator since e.g. the estimator δ_{θ^*} such that $\delta_{\theta^*}(x) = \theta^*$ for all $x \in \mathcal{X}$ and for a $\theta^* \in \Theta$ is in general admissible. (This is for instance the case when $L(\theta, \theta') = 0$ if and only $\theta = \theta'$ while $f(x|\theta) > 0$ for all $(x, \theta) \in \mathcal{X} \times \Theta$.)

However, Stein (1956) shows the following surprising result.

Theorem 2.1 *Let $\Theta = \mathbb{R}^d$, $f(\cdot\,|\theta)$ be the probability density function of the $\mathcal{N}_d(\theta, I_d)$ distribution and $L : \Theta \times \Theta \to [0, +\infty)$ be the quadratic loss function. Then, when $d \geq 3$, the maximum likelihood estimator (MLE) defined by $\delta_0(x) = x$, $x \in \mathbb{R}^d$, is not admissible.*

Proof: See Appendix 1.

Remark: For $d \in \{1, 2\}$, the estimator δ_0 is admissible and therefore a (surprising) corollary of Theorem 2.1 is that the aggregation of several admissible estimators of unrelated quantities is not necessarily admissible.

Remark: Theorem 2.1 has been extended to alternative loss functions and to non-Gaussian models.

The main message of this theorem is that there are no general guarantees that the MLE is admissible.

However, it can be shown that, in the set-up of Theorem 2.1, δ_0 is minimax for any $d \geq 1$ and is asymptotically efficient, and therefore inadmissible estimators are not necessarily bad estimators.

To sum-up: Admissible estimators are not necessarily good estimators and inadmissible estimators are not necessarily bad estimators.

The Bayesian approach

The Bayesian approach considers the posterior expected loss

$$\rho(\pi, d|x) = \int_\Theta L(\theta, d)\pi(\theta|x)\mathrm{d}\theta, \quad d \in \mathcal{D}.$$

Hence, while the frequentist approach integrates on $\mathcal{X}$, the Bayesian approach integrates on Θ. Say differently, the Bayesian approach uses the posterior distribution to integrate out the unknown quantity θ while the frequentist approach uses the likelihood to integrate out the known quantity x.

Using the posterior expected loss, we define $\delta^\pi : \mathcal{X} \to \mathcal{D}$ an estimator such that

$$\delta^\pi(x) \in \operatorname*{argmin}_{d\in\mathcal{D}} \rho(\pi, d|x), \quad \forall x \in \mathcal{X}. \tag{1}$$

Lastly, the integrated risk of δ is given by

$$r(\pi, \delta) = \int_\Theta R(\theta, \delta)\pi(\theta)\mathrm{d}\theta.$$

Definition 2.4 *A Bayes estimator associated with a prior distribution π and a loss function L is any estimator δ^π (defined in* (1)*) such that $r(\pi, \delta^\pi) < +\infty$. The value of $r(\pi) := r(\pi, \delta^\pi)$ is called the Bayes risk.*

Two important properties of Bayes estimators

The following result shows that if δ^π is a Bayes estimator then δ^π is a minimizer of the integrated risk; that is

$$r(\pi) \leq r(\pi, \delta), \quad \forall \delta,$$

and therefore the Bayesian risk is the minimum possible integrated risk.

Theorem 2.2 *An estimator minimising the integrated risk* $r(\pi, \delta)$ *can be obtained by selecting, for every* $x \in \mathcal{X}$*, a value* $\delta(x)$ *belonging to* $\operatorname{argmin}_{d \in \mathcal{D}} \rho(\pi, d|x)$*.*

Proof: This is a direct consequence of the fact that

$$r(\pi, \delta) = \int_\Theta R(\theta, \delta)\pi(\theta)\mathrm{d}\theta = \int_{\mathcal{X}} \rho(\pi, \delta(x)|x)m(x)\mathrm{d}x.$$

Bayes estimators are attractive beyond Bayesian statisticians because they have good frequentist properties. For instance, the next result shows that, under mild conditions, the Bayes estimator is admissible

Theorem 2.3 *If the Bayes estimator is the unique minimizer of the integrated risk then it is admissible.*

Proof: Done in class.

Other good frequentist properties of Bayes estimators are related to their asymptotic behaviours (as the number of observations goes to infinity); see Chapter 6.

The quadratic loss function

Assuming $\mathcal{D} = \Theta$, the quadratic loss function is defined by

$$L(\theta, d) = \|\theta - d\|^2, \quad (\theta, d) \in \Theta^2$$

where $\| \cdot \|$ stands for the Euclidean norm on $\mathbb{R}^d$.

Theorem 2.4 *Assume that $\mathbb{E}_\pi[\theta^T \theta | x] < +\infty$ for all $x \in \mathcal{X}$ and that Θ is a convex set. Then, the estimator δ^π associated with the quadratic loss function is unique and is the posterior expectation,*

$$\delta^\pi(x) = \mathbb{E}_\pi[\theta|x] = \frac{\int_\Theta \theta \pi(\theta) f(x|\theta) \, \mathrm{d}\theta}{\int_\Theta \pi(\theta) f(x|\theta) \, \mathrm{d}\theta}, \quad x \in \mathcal{X}.$$

Proof: Done in class.

Remark: The condition that Θ is a convex set ensures that $\mathbb{E}_\pi[\theta|x] \in \Theta$ for any $x \in \mathcal{X}$.

Proposition 2.1 *Consider the set-up of Theorem 2.4. If $r(\pi) < +\infty$ then the estimator $\delta^\pi(x) = \mathbb{E}_\pi[\theta|x]$ is admissible.*

Proof: If $r(\pi) < +\infty$ the estimator δ^π is the unique Bayes estimator associated with the quadratic loss function and the result follows from Theorem 2.3.

Exercise: Show that the posterior expectation is also the Bayes estimator associated with the more general loss function

$$L(\theta, d) = (\theta - d)^T Q(\theta - d), \quad (\theta, d) \in \Theta^2$$

where Q is an arbitrary symmetric positive definite matrix.

The absolute error loss function

Assuming $\mathcal{D} = \Theta \subseteq \mathbb{R}$, the absolute error loss function is defined by

$$L(\theta, d) = |\theta - d|, \quad (\theta, d) \in \Theta^2.$$

Theorem 2.5 *Assume that $\Theta \subseteq \mathbb{R}$ is a convex set and that $\mathbb{E}_\pi[|\theta||x] < +\infty$ for all $x \in \mathcal{X}$. Then, an estimator δ^π associated with the absolute error loss function is such that, for all $x \in \mathcal{X}$, $\delta^\pi(x)$ is a median of $\pi(\theta|x)$.*

Proof: See Appendix 2.

Recall that $m \in \mathbb{R}$ is a median of $\pi(\theta|x)$ if

$$\pi\big(\{\theta : \theta \leq m\}\big) \geq \frac{1}{2}, \quad \pi\big(\{\theta : \theta \geq m\}\big) \geq \frac{1}{2}.$$

Remarks:

1. The result of Theorem 2.5 still holds if Θ is a countable set.
2. In comparison with the quadratic loss, the absolute error loss penalizes less large errors.
3. The posterior median may not be unique but always exists.
4. Even when the posterior median is not unique $\delta^\pi(x)$ is in general unique.

Exercise: Show that, for $k_1, k_2 > 0$, an estimator δ^π associated with the loss

$$L(\theta, d) = \begin{cases} k_2(\theta - d) \text{ if } \theta > d, \\ k_1(d - \theta) \text{ otherwise} \end{cases}$$

is a $k_2/(k_1 + k_2)$ quantile of the posterior distribution.

The 0–1 loss function

Assuming $\mathcal{D} = \Theta$, the 0–1 loss function is defined by

$$L(\theta, d) = 1 - \mathbb{I}_{\theta}(d), \quad (\theta, d) \in \Theta^2.$$

When the support of $\pi(\theta|x)$ is a countable set we have the following result:

Theorem 2.6 *Assume that the support of $\pi(\theta|x)$ is a countable set. Then, an estimator δ^{π} associated with the above 0–1 loss function is such that, for any $x \in \mathcal{X}$, $\delta^{\pi}(x)$ is a mode of $\pi(\theta|x)$.*

Proof: Done in class.

Remarks:

1. The posterior mode may not be unique.
2. When the posterior mode is not unique $\delta^{\pi}(x)$ is an arbitrary mode of $\pi(\theta|x)$ (and thus δ^{π} is not unique).
3. If $\mathrm{d}\theta$ is a continuous measure (i.e. $\pi(\theta|x)\mathrm{d}\theta$ is a continuous probability distribution) then the posterior expected loss associated with the above 0–1 loss function is one for any decision rule δ since

$$\int_{\Theta} (1 - \mathbb{I}_{\theta}(d)\pi(\theta|x)\mathrm{d}\theta = 1, \quad \forall (d, x) \in \Theta \times \mathcal{X}.$$

The MAP estimator for continuous parameter spaces

Assuming $\mathcal{D} = \Theta \subseteq \mathbb{R}^d$ we consider, for $\epsilon > 0$, the 0–1 loss function defined by

$$L_\epsilon(\theta, d) = \mathbb{I}_{\|\theta - d\| > \epsilon}, \quad (\theta, d) \in \Theta^2.$$

For $\epsilon > 0$ let δ_ϵ^π be an estimator such that

$$\delta_\epsilon^\pi(x) \in \operatorname*{argmin}_{d \in \Theta} \pi\big(\{\theta : \|\theta - d\| > \epsilon\}|x\big), \quad \forall x \in \mathcal{X}$$

and, for $x \in \mathcal{X}$, let $\delta_{\text{MAP}}^\pi(x)$ be a posterior mode of $\pi(\theta|x)$; that is

$$\delta_{\text{MAP}}^\pi(x) \in \operatorname*{argmax}_{\theta \in \Theta} \pi(\theta|x).$$

Then, we have the following result for the MAP estimator.

Theorem 2.7 *Let $x \in \mathcal{X}$, assume that $\pi(\theta|x)$ is continuous. Then, under some technical conditions, $\delta_{MAP}^\pi(x) = \lim_{\epsilon \to 0} \delta_\epsilon^\pi(x)$.*

Proof: See Bassett, R. and Deride J (2016). “Maximum a posteriori estimators as a limit of Bayes estimators”. *Mathematical Programming*, p. 1-16.

Important remarks:

1. Because the MAP estimator is obtained as a limit of Bayes estimators (and not by minimizing a posterior expected loss for a given loss function) it is not a Bayes estimator when $\pi(\theta|x)$ is continuous.

2. Marginal MAP estimates are usually not coherent with the joint MAP estimate.

Example: The Binomial model

Recall that, for parameters $\alpha, \beta > 0$, the density of the Beta(α, β) distribution is defined

$$f_{\alpha,\beta}(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \quad x \in (0,1)$$

where Γ stands for the Gamma function, i.e.

$$\Gamma(t) = \int_0^{+\infty} x^{t-1} e^{-x}\, \mathrm{d}x, \quad t > 0.$$

Proposition 2.2 *Let $(n, \alpha_0, \beta_0) \in \mathbb{N} \times \mathbb{R}_{>0} \times \mathbb{R}_{>0}$ and consider the Bayesian statistical model defined by*

$$\pi(\theta) = f_{\alpha_0,\beta_0}(\theta), \quad f(x_n|\theta) = \binom{n}{x_n} \theta^{x_n}(1-\theta)^{n-x_n},$$

where $x_n \in \{0, \dots, n\}$. Then, $\pi(\theta|x_n) = f_{\alpha_n,\beta_n}(\theta)$ with $\alpha_n = \alpha_0 + x_n$ and $\beta_n = \beta_0 + n - x_n$. Consequently, $\mathbb{E}_\pi[\theta|x_n] = \alpha_n/(\alpha_n + \beta_n)$ and, assuming $\alpha_0, \beta_0 > 1$,

$$\frac{\alpha_n - 1}{\alpha_n + \beta_n - 2} = \operatorname*{argmax}_{\theta \in (0,1)} \pi(\theta|x_n)$$

and

$$\pi\Big(\Big[0, \Big(\alpha_n - \frac{1}{3}\Big)/\Big(\alpha_n + \beta_n - \frac{2}{3}\Big)\Big]\Big|x_n\Big) \approx \frac{1}{2}.$$

Proof: Done in class (but the formula for the posterior median is admitted).

The Binomial model

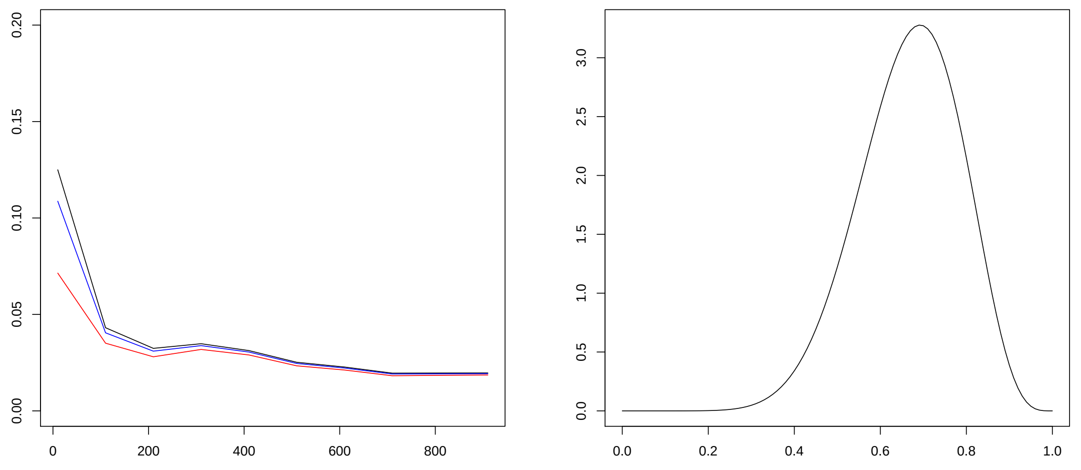


Figure 1: The left plot gives the posterior mean (black), posterior mode (red) and posterior median as a function of n while the right plot shows the prior distribution. The parameters of this latter are $(\alpha_0, \beta_0) = (1, 5)$ while $X_n \sim \text{Binomial}(n, 0.02)$.

Some lessons from the Binomial example:

1. We observe some gaps between the different Bayes estimates when n is small ($n < 200$, say).
2. However, these gaps disappear as n increases and, in fact, the different Bayes estimators converge toward the true parameter value as $n \to +\infty$.
3. One reason for the phenomenon described in 2. is that the posterior distribution is approximatively Gaussian when n is large (see Chapter 6).

Appendix 1: Proof of Theorem 2.1

To prove Theorem 2.1 we will need the following result known as "Stein's lemma".

Lemma 2.1 *Let* $Z \sim \mathcal{N}_1(0,1)$ *and* $h : \mathbb{R} \to \mathbb{R}$ *be a differentiable function such that* $\mathbb{E}[h'(Z)] < +\infty$. *Then,*

$$\mathbb{E}[Zh(Z)] = \mathbb{E}[h'(Z)].$$

Proof: We have

$$\begin{aligned}\mathbb{E}[Zh(Z)] &= \int_{-\infty}^{+\infty} zh(z)\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}\mathrm{d}z \\ &= \int_{-\infty}^{+\infty} zh(z)\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}\mathrm{d}z - h(0)\mathbb{E}[Z] \\ &= \int_{-\infty}^{+\infty} z(h(z)-h(0))\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}\mathrm{d}z \\ &= \int_{-\infty}^{+\infty} z\left(\int_0^z h'(u)\mathrm{d}u\right)\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}\mathrm{d}z\end{aligned} \tag{2}$$

while

$$\begin{aligned}\mathbb{E}[h'(Z)] &= \int_{-\infty}^{0} h'(z)\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}\mathrm{d}z + \int_{0}^{+\infty} h'(z)\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}\mathrm{d}z \\ &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{0} h'(z)\left(\int_{-\infty}^{z} -ue^{-\frac{u^2}{2}}\mathrm{d}u\right)\mathrm{d}z \\ &+ \frac{1}{\sqrt{2\pi}}\int_{0}^{+\infty} h'(z)\left(\int_{z}^{+\infty} ue^{-\frac{u^2}{2}}\mathrm{d}u\right)\mathrm{d}z.\end{aligned} \tag{3}$$

We now study the two integrals that appear on the right-hand side of the second equality sign.

Appendix 1: Proof of Theorem 2.1 (continued)

Using Fubini's theorem,

$$\begin{aligned}
&\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{0} h'(z)\Big(\int_{-\infty}^{z} -ue^{-\frac{u^2}{2}}\mathrm{d}u\Big)\mathrm{d}z \\
&\quad= \int_{-\infty}^{0}\left(\int_{-\infty}^{0}\frac{1}{\sqrt{2\pi}}h'(z)\mathbf{1}_{(-\infty,z]}(u)\mathrm{d}z\right)(-u)e^{-\frac{u^2}{2}}\mathrm{d}z\mathrm{d}u \\
&\quad= \int_{-\infty}^{0}\left(\int_{u}^{0}\frac{1}{\sqrt{2\pi}}h'(z)\mathrm{d}z\right)(-u)e^{-\frac{u^2}{2}}\mathrm{d}u \\
&\quad= \int_{-\infty}^{0}\left(\int_{0}^{u}\frac{1}{\sqrt{2\pi}}h'(z)\mathrm{d}z\right)ue^{-\frac{u^2}{2}}\mathrm{d}u
\end{aligned}$$

and, similarly, one can easily check that

$$\frac{1}{\sqrt{2\pi}}\int_{0}^{+\infty} h'(z)\Big(\int_{z}^{+\infty} ue^{-\frac{u^2}{2}}\mathrm{d}u\Big)\mathrm{d}z = \int_{0}^{+\infty}\left(\int_{0}^{u}\frac{1}{\sqrt{2\pi}}h'(z)\mathrm{d}z\right)ue^{-\frac{u^2}{2}}\mathrm{d}u.$$

Together with (3), this shows that

$$\mathbb{E}[h'(Z)] = \int_{-\infty}^{+\infty} u\left[\int_{0}^{u}\frac{1}{\sqrt{2\pi}}h'(z)\mathrm{d}z\right]e^{-\frac{u^2}{2}}\mathrm{d}u = \mathbb{E}[Zf(Z)]$$

where the second equality is due to (2). This concludes the proof of Lemma 2.1.

Appendix 1: Proof of Theorem 2.1 (continued)

We now prove Theorem 2.1. To this end remark first that

$$R(\theta,\delta_0)=\int_{\mathbb{R}^d}\sum_{i=1}^{d}(\theta_i-x_i)^2 f(x|\theta)\mathrm{d}x=\sum_{i=1}^{d}\mathbb{E}_\theta[(X_i-\theta_i)^2]=d,\quad \forall\theta\in\Theta.$$

Next, let $\delta^{JS}:\mathbb{R}^d\to\mathbb{R}^d$ be defined by

$$\delta^{JS}(x)=x-\frac{d-2}{\|x\|^2}x,\quad x\in\mathbb{R}^d$$

and we now compute $R(\theta,\delta^{JS})$ for all $\theta\in\mathbb{R}^d$ and $d\geq 2$.

Let $d\geq 2$ and $\theta\in\mathbb{R}^d$. Then,

$$\begin{aligned}R(\theta,\delta^{JS})&=\int_{\mathbb{R}^d}\left\|\theta-x+\frac{d-2}{\|x\|^2}x\right\|^2 f(x|\theta)\mathrm{d}x\\&=\int_{\mathbb{R}^d}\|\theta-x\|^2 f(x|\theta)\mathrm{d}x+(d-2)^2\int_{\mathbb{R}^d}\frac{\|x\|^2}{\|x\|^4}f(x|\theta)\mathrm{d}x\\&+2(d-2)\int_{\mathbb{R}^d}(\theta-x)^T\frac{x}{\|x\|^2}f(x|\theta)\mathrm{d}x\\&=R(\theta,\delta_0)+(d-2)^2\int_{\mathbb{R}^d}\frac{1}{\|x\|^2}f(x|\theta)\mathrm{d}x\\&+2(d-2)\sum_{i=1}^{d}\frac{(\theta_i-x_i)x_i}{\|x\|^2}f(x|\theta)\mathrm{d}x\\&=R(\theta,\delta_0)+(d-2)^2\mathbb{E}_\theta\left[\frac{1}{\|X\|^2}\right]\\&+2(d-2)\sum_{i=1}^{d}\mathbb{E}_\theta\left[\frac{(\theta_i-X_i)X_i}{\|X\|^2}\right].\end{aligned}\tag{4}$$

Appendix 1: Proof of Theorem 2.1 (end)

To proceed further let $i \in \{1,\dots,d\}$, $x_{-i}\in\mathbb{R}^{d-1}$ and $h_{i,x_{-i}}:\mathbb{R}\rightarrow\mathbb{R}$ be defined by

$$h_{i,x_{-i}}(z)=\frac{z+\theta_i}{\|(z+\theta_i,x_{-i})\|^2},\quad z\in\mathbb{R}.$$

We have

$$h'_{i,x_{-i}}(z)=\frac{\|(z+\theta_i,x_{-i})\|^2-2(z+\theta_i)^2}{\|(z+\theta_i,x_{-i})\|^4},\quad z\in\mathbb{R}$$

and thus $\mathbb{E}[h'_{i,x_{-i}}(Z)]<+\infty$ when $Z\sim\mathcal{N}_1(0,1)$. Therefore, using Lemma 2.1 we have (with '$X_{-i}=X$ without component i')

$$\begin{aligned}\mathbb{E}_\theta\left[\frac{(\theta_i-X_i)X_i}{\|X\|^2}\Big|X_{-i}=x_{-i}\right]&=-\mathbb{E}_\theta\big[(X_i-\theta_i)h_{i,x_{-i}}(X_i-\theta_i)\big]\\&=-\mathbb{E}_\theta\left[\frac{\|(X_i,x_{-i})\|^2-2X_i^2}{\|(X_i,x_{-i})\|^4}\right]\\&=\mathbb{E}_\theta\left[\frac{2\,X_i^2}{\|(X_i,x_{-i})\|^4}-\frac{1}{\|(X_i,x_{-i})\|^2}\right].\end{aligned}$$

Then, because this equality holds for any $x_{-i}\in\mathbb{R}^{d-1}$,

$$\mathbb{E}_\theta\left[\frac{(\theta_i-X_i)X_i}{\|X\|^2}\right]=2\,\mathbb{E}_\theta\left[\frac{X_i^2}{\|X\|^4}\right]-\mathbb{E}_\theta\left[\frac{1}{\|X\|^2}\right],\quad\forall i\in\{1,\dots,d\}$$

and thus

$$\sum_{i=1}^d\mathbb{E}_\theta\left[\frac{(\theta_i-X_i)X_i}{\|X\|^2}\right]=-(d-2)\mathbb{E}_\theta\left[\frac{1}{\|X\|^2}\right].$$

Then, using (4), it follows that for any $d\geq 2$ we have

$$R(\theta,\delta^{JS})=R(\theta,\delta_0)-(d-2)^2\mathbb{E}_\theta\left[\frac{1}{\|X\|^2}\right],\quad\forall\theta\in\mathbb{R}^d$$

and the proof is complete.

Remark: The above computations are not valid when $d=1$ since in this case $R(\theta,\delta^{JS})=+\infty$ for any $\theta\in\Theta$.

Appendix 2: Proof of Theorem 2.5

To prove Theorem 2.5 let $d \in \Theta$, $x \in \mathcal{X}$ and assume that $\Theta = \mathbb{R}$ (the extension to an arbitrary convex set being trivial). Below we use the shorthand $\pi(\mathrm{d}\theta|x) = \pi(\theta|x)\mathrm{d}\theta$ and $\int_{\mathbb{R}} f(y)\mathrm{d}y$ denotes the (improper) Riemman integral of f on $\mathbb{R}$.

Then, because $\mathbb{E}_\pi[|\theta|\, x] < +\infty$, we have

$$\begin{aligned}\rho(d) := \rho(\pi, d|x) = &\int_\Theta \mathbf{1}_{(-\infty,d]}(\theta)(d-\theta)\pi(\mathrm{d}\theta|x)\\ &+ \int_\Theta \mathbf{1}_{(d,+\infty)}(\theta)(\theta-d)\pi(\mathrm{d}\theta|x)\end{aligned}$$

where (using Fubini's theorem for the third equality)

$$\begin{aligned}\int_{-\infty}^{d} \pi(\{\theta:\ \theta \leq y\}|x)\mathrm{d}y &= \int_{-\infty}^{d}\Big(\int_\Theta \mathbf{1}_{(-\infty,y]}(\theta)\pi(\mathrm{d}\theta|x)\Big)\mathrm{d}y\\ &= \int_{-\infty}^{d}\Big(\int_\Theta \mathbf{1}_{(-\infty,y]}(\theta)\mathbf{1}_{(-\infty,d]}(\theta)\pi(\mathrm{d}\theta|x)\Big)\mathrm{d}y\\ &= \int_\Theta \mathbf{1}_{(-\infty,d]}(\theta)\Big(\int_{-\infty}^{d}\mathbf{1}_{(-\infty,y]}(\theta)\mathrm{d}y\Big)\pi(\mathrm{d}\theta|x)\\ &= \int_\Theta \mathbf{1}_{(-\infty,d]}(\theta)\Big(\int_{\theta}^{d}\mathrm{d}y\Big)\pi(\mathrm{d}\theta|x)\\ &= \int_\Theta \mathbf{1}_{(-\infty,d]}(\theta)(d-\theta)\pi(\mathrm{d}\theta|x).\end{aligned}$$

Remark: Funini's theorem can be used because $\mathbb{E}_\pi[|\theta|\, x] < +\infty$.

Appendix 2: Proof of Theorem 2.5 (continued)

Similarly (using again Fubini's theorem for the third equality),

$$\begin{aligned}\int_d^{+\infty} \pi(\{\theta:\,\theta>y\}|x)\mathrm{d}y &= \int_d^{+\infty}\Big(\int_\Theta \mathbf{1}_{(y,+\infty)}(\theta)\pi(\mathrm{d}\theta|x)\Big)\mathrm{d}y\\ &= \int_d^{+\infty}\Big(\int_\Theta \mathbf{1}_{(y,+\infty)}(\theta)\mathbf{1}_{(d,+\infty)}(\theta)\pi(\mathrm{d}\theta|x)\Big)\mathrm{d}y\\ &= \int_\Theta \mathbf{1}_{(d,+\infty)}(\theta)\Big(\int_d^{+\infty}\mathbf{1}_{(y,+\infty)}(\theta)\mathrm{d}y\Big)\pi(\mathrm{d}\theta|x)\\ &= \int_\Theta \mathbf{1}_{(d,+\infty)}(\theta)\Big(\int_d^{\theta}\mathrm{d}y\Big)\pi(\mathrm{d}\theta|x)\\ &= \int_\Theta \mathbf{1}_{(d,+\infty)}(\theta)(\theta-d)\pi(\mathrm{d}\theta|x)\end{aligned}$$

so that

$$\rho(d)=\int_{-\infty}^d \pi(\{\theta:\,\theta\leq y\}|x)\mathrm{d}y+\int_d^{+\infty}\pi(\{\theta:\,\theta>y\}|x)\mathrm{d}y.$$

Then, using Leibniz integral rule,

$$\begin{aligned}\rho'(d)&=\pi(\{\theta:\,\theta\leq d\}|x)-\pi(\{\theta:\,\theta>d\}|x)\\ &=2\pi(\{\theta:\,\theta\leq d\}|x)-1.\end{aligned}$$

Let $d^*\in\mathbb{R}$ be such that $\pi(\{\theta:\,\theta\leq d^*\}|x)\geq 1/2$ and remark that, since the mapping $y\mapsto|y|$ is convex on $\mathbb{R}$, the mapping $d\mapsto\rho(d)$ is convex on $\mathbb{R}$ (see Problem Sheet 1, Problem 5). Hence,

$$\rho(d)\geq\rho(d^*)+\rho'(d^*)(d-d^*)\geq\rho(d^*),\quad\forall d>d^*. \tag{5}$$

Appendix 2: Proof of Theorem 2.5 (continued)

Next, because (5) holds for any d^* such that $\rho'(d^*) \geq 0$, this inequality holds in particular for

$$d^* = \min\{d \in \mathbb{R} : \pi(\{\theta :\ \theta \leq d\}|x) \geq 1/2\}. \tag{6}$$

(Note that d^* is well defined since the mapping $d \mapsto \pi(\{\theta :\ \theta \leq d\}|x)$ is right continuous.)

Let $d < d^*$. Then, there exists an $\epsilon_d > 0$ such that, for all $\epsilon \in (0, \epsilon_d)$, we have $d \leq d^* - \epsilon < d^*$ and thus

$$\rho(d) \geq \rho(d^* - \epsilon) + \rho'(d^* - \epsilon)(d - (d^* - \epsilon)) \geq \rho(d^* - \epsilon), \quad \forall \epsilon \in (0, \epsilon_d).$$

Then, because the mapping $d \mapsto \rho(d)$ is continuous on $\mathbb{R}$ (because it is convex on this set), together with (5) this shows that d^* defined in (6) is such that

$$\rho(d) \geq \rho(d^*), \quad \forall d \neq d^*.$$

Hence, $d^* \in \operatorname{argmin}_{d \in \Theta} \rho(d)$.

To conclude the proof it remains to show that d^* is a median of $\pi(\theta|x)\mathrm{d}\theta$; that is,

$$\pi(\{\theta :\ \theta \leq d^*\}|x) \geq 1/2, \quad \pi(\{\theta :\ \theta \geq d^*\}|x) \geq 1/2 \tag{7}$$

where the first inequality holds by the definition of d^*.

Appendix 2: Proof of Theorem 2.5 (end)

We show the second inequality in (7) by contradiction and assume that $\pi(\{\theta : \theta \geq d^*\}|x) < 1/2$.

In this case

$$\begin{aligned} \pi(\{\theta : \theta \leq d^*\}|x) &= \pi(\{\theta : \theta < d^*\}|x) + \pi(\{\theta : \theta = d^*\}|x) \\ &> \frac{1}{2} + \pi(\{\theta : \theta = d^*\}|x) \end{aligned} \tag{8}$$

and we consider below the two possible cases.

1. $\pi(\{\theta : \theta = d^*\}|x) = 0$. In this case, (8) imples that

   $$\pi(\{\theta : \theta \leq d^*\}|x) > 1/2$$

   and thus there exists a $d' < d^*$ such that $\pi(\{\theta : \theta \leq d'\}|x) \geq 1/2$, contradicting the definition of d^*. (Such a d' exists because the mapping $d \mapsto \pi(\{\theta : \theta \leq d\}|x)$ is continuous at d^* when $\pi(\{\theta : \theta = d^*\}|x) = 0$.)

2. $\pi(\{\theta : \theta = d^*\}|x) > 0$. In this case,

   $$\pi(\{\theta : \theta \leq d^*\}|x) > \pi(\{\theta : \theta < d^*\}|x) > \frac{1}{2}$$

   and again (since the first equality is strict) there exists a $d' < d^*$ such that $\pi(\{\theta : \theta \leq d'\}|x) \geq 1/2$, contradicting the definition of d^*.

Therefore, (8) holds and the proof is complete.

III - From prior information to prior distribution

The choice of the prior distribution is probably the most delicate part of Bayesian analysis because

1. The choice of $\pi(\theta)$ can have a large impact on the posterior distribution (and thus on all the inference derived from it).
2. There is no (decision) theory to choose $\pi(\theta)$.

Three approaches can be distinguished:

- Subjective priors. Prior information is available (expert knowledge, previous experiments, etc.). Typically, $\pi(\theta)$ is obtained by using this information to (i) select a parametric family of densities and (ii) determine the corresponding parameters.
- Conjugate priors. Limited prior information is available. The choice of the parametric form of $\pi(\theta)$ is made for ease of computations while the corresponding parameters are determined from some prior information.
- Objective priors. Prior information is not available, or too sparse to be taken into account. One must find a way to express this lack of information.

Conjugate prior distributions

Definition 3.1 *A parametric family $\mathcal{F}$ of distributions is said to be conjugate for the model $\{f(\cdot|\theta), \theta \in \Theta\}$ if every prior distribution in $\mathcal{F}$ yields a posterior distribution for this model that is still in $\mathcal{F}$; that is,*

$$\pi \in \mathcal{F} \Rightarrow \pi(\cdot|x) \in \mathcal{F}, \quad \forall x \in \mathcal{X}.$$

The main advantage of using a conjugate prior distribution is therefore that the posterior distribution has a known expression and, usually, classical Bayesian estimators (notably the posterior mean) have a closed form expression.

However, the existence of a conjugate prior distribution is (mostly) limited to the case where $\{f(\cdot|\theta), \theta \in \Theta\}$ is an exponential family of distributions.

Remark: When the model of interest $\{f(\cdot|\theta), \theta \in \Theta\}$ does not admit a conjugate prior, conjugate prior distributions may still be useful to obtain a posterior distribution $\pi(\theta|x)$ which is “easy” to approximate using Markov Chain Monte Carlo methods. This is in particular the case for hierarchical models (see Chapter 9).

Example: The univariate Gaussian model

Let $\theta = (\mu, \sigma^2)$ and $\Theta = \mathbb{R} \times \mathbb{R}_{>0}$.

Proposition 3.1 *Let* $(\alpha_0, \beta_0, \kappa_0) \in \mathbb{R}^3_{>0}$, $\mu_0 \in \mathbb{R}$, $n \in \mathbb{N}_{>0}$ *and consider the Bayesian statistical model where the prior distribution* $\pi(\theta)$ *is such that*

$$\mu|\sigma^2 \sim \mathcal{N}_1(\mu_0, \kappa_0^{-1}\sigma^2), \quad \sigma^2 \sim \textit{Inv-Gamma}(\alpha_0, \beta_0)$$

and where, with $x = (x_1, \dots, x_n) \in \mathbb{R}^n$,

$$f(x|\theta) = \prod_{k=1}^{n} \frac{1}{\sigma}\phi\left(\frac{x_k - \mu}{\sigma}\right), \quad \theta \in \Theta.$$

Then, the posterior distribution $\pi(\theta|x)$ *is such that*

$$\mu|(\sigma^2, x) \sim \mathcal{N}_1(\mu_n, \kappa_n^{-1}\sigma^2), \quad \sigma^2|x \sim \textit{Inv-Gamma}(\alpha_n, \beta_n)$$

with

$$\alpha_n = \alpha_0 + \frac{n}{2}, \quad \beta_n = \beta_0 + \sum_{k=1}^{n} \frac{(x_k - \bar{x}_n)^2}{2} + \frac{n\kappa_0}{n + \kappa_0}\frac{(\bar{x}_n - \mu_0)^2}{2}$$

$$\kappa_n = \kappa_0 + n, \quad \mu_n = \frac{\kappa_0\mu_0 + n\bar{x}_n}{\kappa_0 + n}.$$

Proof: See Problem Sheet 1, Problem 1.

The non-informative approach

When no prior information is available it may still be worth employing Bayesian approaches because they enjoy some classical optimality criteria (e.g. admissibility of Bayes estimators) and because sometimes computing Bayesian estimators is simpler than computing frequentist estimators.

Below we will see two types of non-informative prior distributions:

1. Laplace's prior
2. Jeffreys prior.

Definition 3.2 *A prior density $\pi(\theta)$ is said to be improper if it does not integrate to a finite value; that is, if*

$$\int_\Theta \pi(\theta)\,\mathrm{d}\theta = +\infty.$$

Often, non-informative prior densities are improper but the resulting Bayesian model is considered as valid provided that the corresponding posterior density (as defined by Bayes formula) is proper; i.e.

$$\int_\Theta \pi(\theta) f(x|\theta)\,\mathrm{d}\theta < +\infty.$$

In this case the estimator δ^π is called generalized Bayes estimator because typically the Bayes risk is infinite when an improper prior is used (i.e. $r(\pi) = +\infty$) and thus δ^π is not a Bayes estimator as defined in Chapter 2.

Laplace's prior

Laplace, who was the first to use non-informative techniques (recall the last example of Chapter 1), defines a non-informative prior density as a prior density $\pi(\theta)$ that puts equal value on every $\theta \in \Theta$; that is

$$\pi(\theta) \propto 1, \quad \forall \theta \in \Theta.$$

Note that if Θ is unbounded, the prior density $\pi(\theta)$ is improper.

Example: *Consider the multivariate Gaussian model*

$$X \sim \mathcal{N}_d(\theta, I_d), \quad \theta \in \mathbb{R}^d.$$

Then, the Laplace's prior yields a proper posterior density since in this case

$$\theta|(X = x) \sim \mathcal{N}_d(x, I_d), \quad \forall x \in \mathbb{R}^d.$$

Under a quadratic loss function, δ^π is unique and defined by

$$\delta^\pi(x) = \mathbb{E}_\pi[\theta|x] = x, \quad x \in \mathbb{R}^d.$$

However, $r(\pi) = +\infty$ and thus δ^π is not a Bayes estimator. In addition, δ^π is not admissible when $d > 2$ (see Theorem 2.7).

Remark: In the above example, the inference derived under the Laplace's prior corresponds to the "limiting inference" where the prior variance goes to infinity in a conjugate analysis. This observation often holds true when a conjugate prior distribution is available.

Laplace's prior: re-parametrization issues

The most important problem with the Laplace's prior is the problem of invariance under re-parametrization.

Namely, if $\pi(\theta) = 1$ for all $\theta \in \Theta$ and $g : \Theta \to \Theta$ is a one-to-one and continuously differentiable mapping, then the parameter $\eta = g^{-1}(\theta)$ has prior density

$$\pi^*(\eta) = \left| \det \left[\frac{\partial}{\partial \eta} g(\eta) \right] \right| \pi(g(\eta)) = \left| \det \left[\frac{\partial}{\partial \eta} g(\eta) \right] \right|$$

by the Jacobian formula.

Therefore, the prior density $\pi^*(\eta)$ is, in general, not constant (and hence not non-informative in the sense of Laplace) although the information on η is the same as on θ!

The Jeffreys prior

The Jeffreys prior is based on the idea that a non-informative prior density is a prior density derived entirely from the model $\{f(\cdot|\theta), \theta \in \Theta\}$ since this is the only available information.

Definition 3.3 *For a given parametric model $\{f(\cdot|\theta), \theta \in \Theta\}$, the Jeffreys prior is defined by the density*

$$\pi_J(\theta) \propto \{\det[I(\theta)]\}^{1/2}, \quad \theta \in \Theta$$

with $I(\theta)$ the Fisher Information matrix; that is,

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

where the second inequality holds only under some conditions and where $\mathbb{E}_\theta$ denotes the expectations under $f(x|\theta)\mathrm{d}x$.

Compared to the Laplace's prior, the Jeffreys prior has the advantage to be invariant by re-parametrisation, as shown in the next result.

Proposition 3.2 *Assume that θ has prior density $\pi_J(\theta)$ and let $g : \Theta \to \Theta$ be a one-to-one and continuously differentiable mapping. Then, $\eta = g^{-1}(\theta)$ has prior density*

$$\pi^*(\eta) \propto \left| \det \left[\frac{\partial}{\partial \eta} g(\eta) \right] \right| \pi_J(g(\eta)) = \left\{ \det[\tilde{I}(\eta)] \right\}^{1/2}$$

where

$$\tilde{I}(\eta) = \mathbb{E}_{g(\eta)} \left[\frac{\partial \log f(X|g(\eta))}{\partial \eta} \frac{\partial \log f(X|g(\eta))}{\partial \eta^T} \right].$$

Proof: Use the fact that $\det(AB) = \det(A)\det(B)$.

Example: The Jeffreys prior for the Binomial model is the Beta$(1/2, 1/2)$ distribution.

Some remarks concerning the Jeffreys prior

- The Fisher information matrix $I(\theta)$ is related to the curvature of the Kullback-Leiber divergence and hence provides a measure of the ability of the model to discriminate between θ and $\theta + \mathrm{d}\theta$ (see the Appendix for more details on this).

  Therefore, the Jeffreys prior puts

  - a small prior probability on the values of θ for which the observations provide little information.
  - a large prior probability on the values of θ for which the observations provide a "significant" amount of information.

- The Jeffreys prior is often improper.

- The Jeffreys prior is not always computable.

Jeffreys prior as a bias reduction technique

Let $X^{(n)} = (X_1, \dots, X_n)$, where the X_i's are i.i.d. with common distribution $\tilde{f}(\cdot|\theta_0)$ for some $\theta_0 \in \Theta \subset \mathbb{R}^d$ and where $\{\tilde{f}(\cdot|\theta), \theta \in \Theta\}$ is an exponential family model with canonical parameter θ.

Let $f(X^{(n)}|\theta) = \prod_{i=1}^n \tilde{f}(X_i|\theta)$, $\pi(\theta)$ be a prior distribution for θ such that $\pi(\theta) > 0$ for all $\theta \in \Theta$ and $\hat{\theta}_n$ be the maximum likelihood estimator (MLE) of θ_0.

Then, it is well-known that

- $\|\mathbb{E}_{\theta_0}[\hat{\theta}_n] - \theta_0\| = \mathcal{O}(n^{-1})$
- $\|\mathbb{E}_{\theta_0}[\delta^{\pi}_{\text{MAP}}(X^{(n)})] - \theta_0\| = \mathcal{O}(n^{-1})$ (with $\delta^{\pi}_{\text{MAP}}$ as in Chapter 2).

However, it can be shown that[a]

$$\|\mathbb{E}_{\theta_0}[\delta^{\pi_J}_{\text{MAP}}(X^{(n)})] - \theta_0\| = o(n^{-1}) \tag{1}$$

so that, under the Jeffreys prior, the MAP outperforms the MLE in term of bias!

[a] see Firth, D. (1993). “Bias reduction of maximum likelihood estimates”. *Biometrika*, 80(1), 27-38.

General remarks on the objective approach

- A more basic way of obtaining a non-informative prior density is to take a vague prior; that is, a prior distribution with a ‘large’ variance.
- Automated use of non-informative prior distributions does not make sense in any settings. In particular, this is ‘sub-optimal’ if some prior information is available.
- Careless use of improper prior densities is dangerous. In particular, it may be that the posterior density does not define a proper probability distribution.
- Generalized Bayes estimators do not share the same optimality properties than Bayes estimators. In particular, the formers may not be admissible and may have very poor properties (see e.g. the next example).
- Improper prior densities should be avoided for hypothesis testing (see Chapter 4).

Generalized Bayes estimators and admissibility

The goal of this example is to show that, although Bayes estimators are (under weak conditions) admissible, generalized Bayes estimators can have very poor properties.

Let $f(\cdot|\theta)$ be the p.d.f. of the $\mathcal{N}_d(\theta, I_d)$ distribution, $\theta \in \mathbb{R}^d$ and $\pi(\theta) \propto 1$ so that

$$\theta|(X = x) \sim \mathcal{N}_d(x, I_d), \quad \forall x \in \mathbb{R}^d.$$

Assuming a quadratic loss function, the corresponding generalized Bayes estimator $\delta^\pi(x) := \mathbb{E}_\pi[\theta|x] = x$ is reasonable as mentioned in Chapter 2 (i.e. it is minimax and has good asymptotic properties).

Assume now that the parameter of interest is $\eta = ||\theta||^2$. Then, under the above prior distribution, the generalized Bayes estimator of η is

$$\mathbb{E}_\pi[\eta|x] = ||x||^2 + d.$$

However, as shown in the next result, this estimator has very poor properties:

Proposition 3.3 *For $c \in \mathbb{R}$, let $\delta_c : \mathbb{R}^d \to \mathbb{R}_+$ be the estimator of η defined by*

$$\delta_c(\tilde{x}) = ||\tilde{x}||^2 + c, \quad \tilde{x} \in \mathbb{R}^d.$$

Then, under a quadratic loss function,

$$R(\eta, \delta_{-d}) < R(\eta, \delta_d), \quad \forall \eta \in \mathbb{R}_+$$

and $\sup_{\eta \in \mathbb{R}_+} R(\eta, \delta_{-d}) = \sup_{\eta \in \mathbb{R}_+} R(\eta, \delta_d) = +\infty$.

Proof: See Problem Sheet 2, Problem 3.

Alternative approaches

As mentioned above, the use of improper prior densities as non-informative prior may be problematic.

Below we describe two approaches that aim at reducing the impact of the prior distribution on the inference.

Let $\{\tilde{\pi}(\theta|\lambda),\, \lambda \in \Lambda\}$ be a family of probability density functions on Θ.

- **Hierarchical modelling:** Instead of choosing a fixed $\lambda \in \Lambda$, we can reduce the influence of the prior distribution by assigning a prior distribution $\pi_2(\lambda)$ on Λ and then by considering the prior distribution $\pi(\theta)$ defined by

$$\pi(\theta) = \int_\Lambda \tilde{\pi}(\theta|\lambda)\pi_2(\lambda)\mathrm{d}\lambda, \quad \theta \in \Theta.$$

- **Empirical Bayes:** The idea is to use the observation $x \in \mathcal{X}$ to select a value $\lambda_x \in \Lambda$ and then to use the prior distribution $\pi(\theta)$ defined by

$$\pi(\theta) = \tilde{\pi}(\theta|\lambda_x), \quad x \in \mathcal{X}.$$

  Typically, λ_x is estimated from the marginal distribution

$$m(x|\lambda) = \int_\Theta f(x|\theta)\tilde{\pi}(\theta|\lambda)\mathrm{d}\theta$$

  by taking, for instance, $\lambda_x \in \operatorname{argmax}_{\lambda\in\Lambda} m(x|\lambda)$.

  Remark: This approach is not fully Bayesian since the observation x is used twice (i.e. in the prior and in the likelihood).

Appendix: KL divergence and Fisher information matrix

We first recall the definition of the Kullback-Leiber (KL) divergence.

Definition 3.4 *The KL divergence $KL(\cdot|\cdot) : \Theta \times \Theta \to \mathbb{R}$ is defined by*

$$KL(\theta'|\theta) = \mathbb{E}_{\theta'}\left[\log \frac{f(X|\theta')}{f(X|\theta)}\right] = \int_{\mathcal{X}} \log \frac{f(x|\theta')}{f(x|\theta)} f(x|\theta')\mathrm{d}x, \quad (\theta', \theta) \in \Theta^2.$$

The next result provides two basic properties of the KL divergence which imply that the quantity $KL(\theta'|\theta)$ can be used to measure the difference between the probability distributions $f(x|\theta')\mathrm{d}x$ and $f(x|\theta)\mathrm{d}x$.

Proposition 3.4 *For any $(\theta, \theta') \in \Theta^2$, $KL(\theta'|\theta) \geq 0$ and, if the model $\{f(\cdot|\theta),\, \theta \in \Theta\}$ is well identified, $KL(\theta'|\theta) = 0$ if and only of $\theta' = \theta$.*

Proof: Let $(\theta', \theta) \in \Theta^2$. Then,

$$\mathbb{E}_{\theta'}\left[\log \frac{f(X|\theta')}{f(X|\theta)}\ \right] = \mathbb{E}_{\theta'}\left[-\log \frac{f(X|\theta)}{f(X|\theta')}\right] \geq -\log \mathbb{E}_{\theta'}\left[\frac{f(X|\theta)}{f(X|\theta')}\right] = \log(1)$$

where the inequality uses Jensen's inequality. The second part of the proposition is trivial.

Remark: If $KL(\theta'|\theta)$ is a measure of the difference between the probability distributions $f(x|\theta')\mathrm{d}x$ and $f(x|\theta)\mathrm{d}x$, the KL divergence is not a true metric since it does not satisfy the triangular inequality and is, in general, not symmetric.

Proposition 3.5 below shows that the Fisher information matrix is related to the curvature of the KL divergence and hence provides a measure of the ability of the model to discriminate between θ and $\theta + \mathrm{d}\theta$.

Appendix: KL divergence and Fisher information matrix (end)

Proposition 3.5 *Under some regularity conditions (that notably ensure that we can swap integration and differentiation operations)*

$$KL(\theta'|\theta) = \frac{1}{2}(\theta'-\theta)^T I(\theta)(\theta'-\theta) + o(\|\theta'-\theta\|^2) \quad \textit{as } \|\theta'-\theta\|^2 \to 0.$$

Proof: Using a Taylor expansion of order 2, we have (as $\|\theta'-\theta\|^2 \to 0$)

$$KL(\theta'|\theta) = (\theta'-\theta)^T \frac{\partial KL(\tilde{\theta}|\theta)}{\partial \tilde{\theta}}\Big|_{\tilde{\theta}=\theta} + \frac{1}{2}(\theta'-\theta)^T \frac{\partial^2 KL(\tilde{\theta}|\theta)}{\partial \tilde{\theta}\partial \tilde{\theta}^T}\Big|_{\tilde{\theta}=\theta}(\theta'-\theta) + o(\|\theta'-\theta\|^2)$$

where

$$\frac{\partial KL(\tilde{\theta}|\theta)}{\partial \tilde{\theta}} = \frac{\partial}{\partial \tilde{\theta}} \int_{\mathcal{X}} f(x|\theta) \log \frac{f(x|\theta)}{f(x|\tilde{\theta})} \mathrm{d}x = \int_{\mathcal{X}} \frac{\partial}{\partial \tilde{\theta}} \Big(\log \frac{f(x|\theta)}{f(x|\tilde{\theta})} \Big) f(x|\theta) \mathrm{d}x = -\int_{\mathcal{X}} \frac{\partial \log f(x|\tilde{\theta})}{\partial \tilde{\theta}} f(x|\theta)\mathrm{d}x$$

so that

$$\frac{\partial KL(\tilde{\theta}|\theta)}{\partial \tilde{\theta}}\Big|_{\tilde{\theta}=\theta} = -\int_{\mathcal{X}} \frac{\partial f(x|\theta)}{\partial \theta} \mathrm{d}x = -\frac{\partial}{\partial \theta} \int_{\mathcal{X}} f(x|\theta)\mathrm{d}x = 0.$$

Moreover,

$$\frac{\partial^2 KL(\tilde{\theta}|\theta)}{\partial \tilde{\theta}\partial \tilde{\theta}^T} = -\int_{\mathcal{X}} \frac{\partial^2 \log f(x|\tilde{\theta})}{\partial \tilde{\theta}\partial \tilde{\theta}^T} f(x|\theta)\mathrm{d}x$$

so that

$$\frac{\partial^2 KL(\tilde{\theta}|\theta)}{\partial \tilde{\theta}\partial \tilde{\theta}^T}\Big|_{\tilde{\theta}=\theta} = -\int_{\mathcal{X}} \frac{\partial^2 \log f(x|\theta)}{\partial \tilde{\theta}\partial \tilde{\theta}^T}\Big|_{\tilde{\theta}=\theta} f(x|\theta)\mathrm{d}x = I(\theta).$$

The proof is complete.

IV - Hypothesis testing and credible sets

For a given statistical model $\{f(\cdot|\theta), \theta \in \Theta\}$ the problem of hypothesis testing consists in answering the following question:

> Is the hypothesis that θ belongs to the subset $\Theta_0 \subset \Theta$ of the parameter space acceptable?

In short, we want to test the null hypothesis $H_0 : \theta \in \Theta_0$ against the alternative hypothesis $H_1 : \theta \in \Theta_1$, with $\Theta_1 = \Theta \setminus \Theta_0$.

Example: *Consider the following logistic regression model:*

$$\mathbb{P}(Y = 1|Z = z) = \exp(\beta_0 + \beta_1 z)/\{1 + \exp(\beta_0 + \beta_1 z)\}.$$

- *How gender affects a given behaviour?*

$$H_0 : \beta_1 > 0, \quad H_1 : \beta_1 \leq 0.$$

- *Does the presence of a nuclear plant increases the risk of leukemia in its vicinity?*

$$H_0 : \beta_1 = 0, \quad H_1 : \beta_1 > 0.$$

- *Does weather affects the sales of a given product?*

$$H_0 : \beta_1 = 0, \quad H_1 : \beta_1 \neq 0.$$

Hypothesis testing as a decision theoretic problem

Hypothesis testing is clearly a decision theoretic problem (see Chapter 2) where the set of possible decisions contains only two elements

$$\mathcal{D} = \{0, 1\}.$$

By convention, 1 stands for the acceptance of H_0 and 0 for its rejection.

The most natural loss function in that particular context is the a_0–a_1 loss defined, for $(\theta, d) \in \Theta \times \mathcal{D}$ and $a_0, a_1 \geq 0$, by

$$L(\theta, d) = \begin{cases} 0 & \text{if } d = \mathbf{1}_{\Theta_0}(\theta), \\ a_0 & \text{if } \theta \in \Theta_0 \text{ and } d = 0, \\ a_1 & \text{if } \theta \in \Theta_1 \text{ and } d = 1. \end{cases}$$

Example: *Spam filtering consists in detecting and discarding automatically any unsolicited e-mail that gets into your mailbox. The most elaborate spam filters (e.g. Mozilla/Netscape) rely on a Bayesian modelling of the occurrence of particular terms. This is a testing problem (spam/not spam) in which the loss function is clearly not symmetrical (the 'cost' of discarding a non-spam e-mail being more important than of keeping a spam mail).*

In some contexts it is not necessarily clear what are the respective costs of each decision, so one may set arbitrarily $a_0 = a_1 = 1$ (and recover the 0–1 loss function introduced in Chapter 2).

Remark: Hypothesis testing can also be considered as the problem of estimating the function of θ defined by $g(\theta) := \mathbf{1}_{\Theta_0}(\theta)$.

Bayesian estimator and a_0–a_1 loss function

Proposition 4.1 *The decision rule $\delta^\pi : \mathcal{X} \to \{0,1\}$ associated with the a_0–a_1 loss function is defined, for $x \in \mathcal{X}$, by*

$$\delta^\pi(x) = \begin{cases} 1 & \textit{if } \pi(\Theta_0|x) > a_1/(a_0+a_1), \\ 0 & \textit{otherwise.} \end{cases}$$

Proof: Done in class.

Remarks:

- If $a_0 = a_1$ this procedure simply consists in selecting the hypothesis with the highest posterior probability.
- This procedure only depends on a_0/a_1. The larger this ratio is, the more important a wrong answer under H_0 is relative to H_1.
- The quantity $a_1/(a_0+a_1)$ is referred to as the acceptance level.
- Since the a_0–a_1 loss is bounded, as soon as $\pi(\theta)$ is a proper prior density the Bayes risk is finite (i.e. $r(\pi) < +\infty$) and thus δ^π is the Bayes estimator of $g(\theta) := \mathbf{1}_{\Theta_0}(\theta)$.

Notation: We recall that for a measurable set $A \subseteq \Theta$ we use the notation

$$\pi(A|x) = \int_A \pi(\theta|x)\mathrm{d}\theta, \quad \pi(A) = \int_A \pi(\theta)\mathrm{d}\theta.$$

The Bayes factor

Bayesian hypothesis testing (and Bayesian model choice, see Chapter 5) are often performed using the Bayes factor.

Definition 4.1 *The Bayes factor* $B_{01}^{\pi}:\mathcal{X}\to[0,+\infty)$ *for the test* $H_0:\theta\in\Theta_0$ *versus* $H_1:\theta\in\Theta_1$ *is defined by*

$$B_{01}^{\pi}(x)=\frac{\pi(\Theta_0|x)}{\pi(\Theta_1|x)}\frac{\pi(\Theta_1)}{\pi(\Theta_0)}\quad x\in\mathcal{X}.$$

The Bayes factor therefore measures the modification of the odds of H_0 against H_1 due to the observation of x. It allows for (partly) reducing the impact of the prior distribution on the decision and as such is considered as an ‘objective' quantity.

To understand this last point let $\pi_i(\theta)$ be a prior density under H_i $(i=0,1)$, $\rho_0\in(0,1)$ and

$$\pi(\theta)=\rho_0\pi_0(\theta)+(1-\rho_0)\pi_1(\theta)\implies\pi(\Theta_0)=\rho_0.$$

Then,

$$B_{01}^{\pi}(x)=\frac{\int_{\Theta}f(x|\theta)\pi_0(\theta)\mathrm{d}\theta}{\int_{\Theta}f(x|\theta)\pi_1(\theta)\mathrm{d}\theta}=\frac{m_0(x)}{m_1(x)}\tag{1}$$

where

$$m_i(x)=\int_{\Theta}f(x|\theta)\pi_i(\theta)\mathrm{d}\theta,\quad i=0,1.\tag{2}$$

Consequently, $B_{01}^{\pi}(x)$ does not depend on $\pi(\Theta_0)$, the prior probability that H_0 is true.

Remark: If $\Theta_0=\{\theta_0\}$ and $\Theta_1=\{\theta_1\}$, then

$$B_{01}^{\pi}(x)=\frac{f(x|\theta_0)}{f(x|\theta_1)}$$

so that the Bayes factor reduces to the likelihood ratio.

Duality between loss and prior distribution

Proposition 4.2 *The decision rule $\delta^\pi : \mathcal{X} \to \{0, 1\}$ associated with the a_0–a_1 loss function can be alternatively defined, for $x \in \mathcal{X}$, by*

$$\delta^\pi(x) = \begin{cases} 1 & \textit{if } B_{01}^\pi(x) > \frac{a_1 \pi(\Theta_1)}{a_0 \pi(\Theta_0)}, \\ 0 & \textit{otherwise.} \end{cases}$$

Proof: Done in class.

Remark: In the expression of δ^π given in Proposition 4.1 the impact of $\pi(\Theta_0)$ on the testing procedure is hidden in the quantity $\pi(\Theta_0|x)$ while in the one given in Proposition 4.2 it is explicit.

Remark: $a_i \pi(\Theta_i)$ is the prior expected loss of not choosing H_i.

This alternative condition for accepting H_0 illustrates the fundamental duality between the prior distribution and the loss function.

Indeed, it is equivalent to:

- To modify the penalties as follows

  $$a_0 \to a_0\, \pi(\Theta_0), \quad a_1 \to a_1\, \pi(\Theta_1)$$

  and then set $\pi(\Theta_0) = \pi(\Theta_1) = 1/2$;

- To modify the prior probabilities as follows:

  $$\pi(\Theta_0) \to \frac{a_0\, \pi(\Theta_0)}{a_0 \pi(\Theta_0) + a_1 \pi(\Theta_1)}, \quad \pi(\Theta_1) \to \frac{a_1\, \pi(\Theta_1)}{a_0 \pi(\Theta_0) + a_1 \pi(\Theta_1)}$$

  and then set $a_0 = a_1$.

Jeffreys' scale

The strength of the evidence in favour of H_0 is sometimes measured through a scale proposed by Jeffreys (1961) which goes a follows:

- The evidence is poor if $\log_{10} B_{01}^{\pi}(x)$ is between 0 and 0.5;
- The evidence is substantial if $\log_{10} B_{01}^{\pi}(x)$ is between 0.5 and 1;
- The evidence is strong if $\log_{10} B_{01}^{\pi}(x)$ is between 1 and 2;
- The evidence is decisive if $\log_{10} B_{01}^{\pi}(x)$ is larger than 2;

Remark: The bounds separating one strength from another are mostly a matter of convention (and not derived using decision theory).

Point-null hypotheses

We now consider tests where $\Theta_0 = \{\theta^*\}$ for some $\theta^* \in \Theta$.

Some statisticians argue that point-null hypotheses are intrinsically absurd when Θ is a continuous space. Indeed, how can we determine that a parameter attains an *exact* value (in other words, that it is exactly known) from a finite amount of data?

In some cases however, point-null hypotheses make more sense because they are related to a *qualitative* aspect of the model.

Example: *In the Logistic regression model*

$$\mathbb{P}(Y = 1|Z = z) = \exp(\beta_0 + \beta_1 z)/\{1 + \exp(\beta_0 + \beta_1 z)\}$$

the hypothesis $H_0 : \beta_1 = 0$ *is equivalent to assuming that the factor* z *has no impact on the event* y*.*

In any cases point-null hypotheses are very popular in practice and thus there is a need for the Bayesian point-null hypothesis testing procedure that we introduce below.

Bayes factors for point-null hypotheses

If a given null hypothesis is reasonable for the problem at hand, then its prior probability must be positive (and thus $\pi(\theta)$ cannot be a continuous prior distribution).

Let $\rho_0 \in (0,1)$ be the prior probability that $\theta = \theta^*$ and $g_1(\theta)$ be a prior density on $\Theta_1 = \Theta \setminus \{\theta^*\}$.

Then, we consider the prior distribution $\pi(\theta)$ defined by

$$\pi(\theta) = \rho_0 \mathbf{1}_{\{\theta^*\}}(\theta) + (1-\rho_0) g_1(\theta) \mathbf{1}_{\{\theta \neq \theta^*\}}(\theta), \quad \theta \in \Theta. \tag{3}$$

Remark: When we write $\pi(\theta)$ as in (3) we implicitly assume that $\mathrm{d}\theta$ is such that $\int_{\{\theta^*\}} \pi(\theta)\mathrm{d}\theta = \rho_0$.

In this case,

$$\pi(\{\theta^*\}|x) \propto \rho_0 f(x|\theta^*)$$

while

$$\pi(\Theta_1|x) \propto (1-\rho_0) \int_{\Theta_1} f(x|\theta) g_1(\theta)\, \mathrm{d}\theta.$$

Using the notation introduced above,

$$\pi_0(\theta) = \mathbf{1}_{\{\theta^*\}}(\theta), \quad \pi_1(\theta) = g_1(\theta)\mathbf{1}_{\{\theta \neq \theta^*\}}(\theta)$$

and thus, by (1),

$$B_{01}^{\pi}(x) = \frac{f(x|\theta^*)}{m_1(x)}$$

where $m_1(x)$ is defined in (2); that is,

$$m_1(x) = \int_{\Theta} f(x|\theta)\pi_1(\theta)\mathrm{d}\theta = \int_{\Theta_1} f(x|\theta) g_1(\theta)\mathrm{d}\theta.$$

Remark: As per above the Bayes factor does not depend on ρ_0.

Hypothesis testing with noninformative prior distributions

Remark first that the testing setting is not coherent with an absolute lack of information since it implies to partition the parameter space into two sets.

In addition, for point-null hypotheses, the set Θ_0 has in general measure zero under the Laplace and the Jeffreys prior densities.

Moreover, Laplace's prior (and more generally improper prior densities) poses problems for hypothesis testing as illustrated in the following example.

Example 4.1 *Let $f(\cdot|\,\theta)$ be the p.d.f. of the $\mathcal{N}_1(\theta,1)$ distribution and consider the test $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$. Then, if we use the prior $\pi(\theta)$ defined by*

$$\pi(\theta) = \frac{1}{2}\mathbf{1}_{\{0\}}(\theta) + \frac{1}{2}\mathbf{1}_{\{\theta\neq 0\}}(\theta), \quad \theta \in \Theta := \mathbb{R}$$

we have

$$\pi(\{0\}|x) = \frac{1}{1+\sqrt{2\pi}e^{x^2/2}} \leq \frac{1}{1+\sqrt{2\pi}} \approx 0.285, \quad \forall x \in \mathbb{R}^d.$$

The posterior probability $\pi(\{0\}|x)$ is therefore bounded above so that the posterior distribution is biased toward H_1. Consequently, if the loss function does not take this bias into account the null hypothesis will be often rejected.

More one testing with improper prior distributions

Example 4.1 (continued) *Assume now that $\pi(\theta)$ is defined by*

$$\pi(\theta) = \rho_0 \mathbf{1}_{\{0\}}(\theta) + (1-\rho_0)\, c\, \mathbf{1}_{\{\theta \neq 0\}}(\theta), \quad \theta \in \Theta$$

for some constant $c > 0$ and where $\rho_0 \in (0,1)$.

Then, we have

$$B_{01}^{\pi}(x) = \frac{1}{c\sqrt{2\pi}\, e^{x^2/2}}$$

and therefore, by increasing (resp. decreasing) the parameter c, we can make the Bayes factor arbitrarily small (resp. large).

Conclusion: Improper prior densities should be avoided in the context of hypothesis testing.

Remark: Another reason why improper prior densities should be avoided in the context of hypothesis testing is that $\pi(\Theta_i) = +\infty$ for at least on $i \in \{0,1\}$, so that the testing procedure given in Proposition 4.2 is meaningless.

Testing with a vague prior distribution

Example 4.1 (end) *Assume now that $\pi(\theta)$ is defined by*

$$\pi(\theta) = \rho_0 \mathbf{1}_{\{0\}}(\theta) + (1-\rho_0) g_1(\theta) \mathbf{1}_{\{\theta \neq 0\}}(\theta), \quad \theta \in \Theta := \mathbb{R}$$

where $g_1(\theta)$ is the density of the $\mathcal{N}_1(0, \sigma_0^2)$ distribution. Then, we have

$$\pi(\{0\}|x) = \left[1 + \frac{1-\rho_0}{\rho_0} \sqrt{\frac{1}{1+\sigma_0^2}} \exp\left(\frac{\sigma_0^2 x^2}{2(1+\sigma_0^2)}\right)\right]^{-1}, \quad x \in \mathbb{R}^d$$

which converges to 1 as $\sigma_0^2 \to +\infty$ for every observation x!

If instead we use for $g_1(\theta)$ the Laplace's prior $g_1(\theta) \equiv 1$ we obtain, as shown above,

$$\pi(\{0\}|x) = \left[1 + \sqrt{2\pi} e^{x^2/2}\right]^{-1}, \quad x \in \mathbb{R}^d.$$

Consequently, and contrary to what we saw for point estimation (see Chapter 3), limiting arguments (i.e. letting the prior variance going to infinity) do not allow to derive uninformative answers in the context of hypothesis testing.

Credible intervals and confidence intervals

A topic closely related to hypothesis testing is the derivation of regions of the parameter space that contain the ‘most likely values’ for the parameter, called credible sets in the Bayesian approach.

The frequentist counterpart of credible sets are the *confidence interval* which are usually derived from the asymptotic distribution of the estimator.

For instance, for an univariate parameter θ, an $(1-\alpha)$-confidence interval is typically

$$\left[\widehat{\theta}_n - q_{1-\alpha/2}\frac{\widehat{\sigma}_n}{\sqrt{n}}, \widehat{\theta}_n + q_{1-\alpha/2}\frac{\widehat{\sigma}_n}{\sqrt{n}}\right]$$

where n is the sample size, $\widehat{\theta}_n$ is the frequentist estimator of θ (e.g. the MLE), $\widehat{\sigma}^2_n$ is an estimator of the asymptotic variance of $\sqrt{n}(\widehat{\theta}_n - \theta_0)$, with θ_0 the “true” parameter value, and $q_{1-\alpha/2}$ is the $(1-\alpha/2)$-quantile of the $\mathcal{N}_1(0,1)$ distribution, e.g. $q_{1-\alpha/2} = 1.96$ for $\alpha = 0.05$.

Reminder: In the frequentist perspective, the parameter is fixed and the confidence interval is random, having a probability of $(1-\alpha)$ to actually contain θ_0 (when we repeat the same experiment a great number of times). It is therefore not possible to interpret $(1-\alpha)$ as the probability that the parameter lies in the confidence interval *for the considered experiment.*

Credible sets

Definition 4.2 *Let $\alpha \in (0,1)$. A subset $C \subseteq \Theta$ is an $(1-\alpha)$-credible set for θ if*

$$\pi(C|x) \geq 1-\alpha.$$

The notion of credible sets is not very useful by itself as there are obviously an infinite number of credible sets for a given α.

In an univariate setting, one may restrict its attention to the credible interval centred at a given Bayesian estimator. This is arbitrary however, as it depends on a particular choice of Bayesian estimator (posterior mean, median, etc.) and such intervals may not exist when Θ is bounded, as illustrated in the next example.

Example: Assume that the posterior distribution is the Beta(1,30) distribution. Then, the posterior mean is $1/31$, and the posterior probability of being above $1/31$ is approximatively 0.37. Therefore credible intervals centred at $1/31$ exists only if $1-\alpha$ is smaller than 0.74.

Highest posterior density regions

A more satisfactory approach is to restrict our attention to the credible set that contains the 'most likely values'.

Definition 4.3 *The subset $C_\alpha(x)$ of the parameter space is a highest posterior density (HPD) region at level $(1-\alpha)$ if it is of the form*

$$\{\theta \in \Theta : \pi(\theta|x) > \gamma_\alpha\} \subset C_\alpha(x) \subset \{\theta \in \Theta : \pi(\theta|x) \geq \gamma_\alpha\}$$

where γ_α is the largest bound such that

$$\pi(C_\alpha(x)|x) \geq 1 - \alpha.$$

Remarks:

- HPD regions minimize the volume among $(1-\alpha)$-credible regions.
- If $\pi(\theta|x)$ is a continuous density the HPD region is simply

$$C_\alpha(x) = \{\theta \in \Theta : \pi(\theta|x) \geq \gamma_\alpha\}.$$

- If the choice of an HPD region among $(1-\alpha)$-credible sets is natural, it is also justified (to some extend) from a decision theoretic perspective.

V - Model Choice

The goal of this short chapter is to present the Bayesian answer to the important problem of selecting one model among a set of $M \in \bar{\mathbb{N}}$ competing models:

$$\mathcal{M}_i = \{f_i(\cdot|\theta_i),\, \theta_i \in \Theta_i \subset \mathbb{R}^{d_i}\}, \quad i = 1, ..., M.$$

Model choice can be seen as an extension of hypothesis testing since testing $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_1$ is equivalent to choosing between the model

$$\mathcal{M}_0 = \{f(\cdot|\theta),\, \theta \in \Theta_0\}$$

and the model

$$\mathcal{M}_1 = \{f(\cdot|\theta),\, \theta \in \Theta_1\}.$$

However, the inference in this chapter is on much ‘bigger’ objects than in Chapter 4 since we are now dealing with models rather than parameters. As a consequence of this, and as briefly explained below, the Bayesian solution of model choice is usually hard to justify from a purely Bayesian perspective.

Model choice as an estimation problem

The standard Bayesian solution to model choice consists to extend the prior modelling from parameters to models by considering the index of the model $\mu \in \{1, \dots, M\}$ as an additional parameter to estimate.

More precisely, let

$$\Theta = \cup_{i=1}^{M} \{i\} \times \Theta_i$$

be the parameter space, $\pi_i(\theta_i)$ be a prior distribution on Θ_i and $(p_1, \dots, p_M)$ be the prior distribution of μ.

Then, using Bayes theorem, the posterior distribution of μ given the observation x is given by

$$\begin{aligned} \pi(i|x) &= \frac{p_i \int_{\Theta_i} f_i(x|\theta_i)\pi_i(\theta_i)\mathrm{d}\theta_i}{\sum_{j=1}^{M} p_j \int_{\Theta_j} f_j(x|\theta_j)\pi_j(\theta_j)\mathrm{d}\theta_j} \\ &= \frac{p_i\, m_i(x)}{\sum_{j=1}^{M} \, p_j m_j(x)}, \quad i = 1, \dots, M \end{aligned}$$

and we can use the estimator δ^π derived in Chapter 2 to estimate μ.

Typically, the MAP (i.e. the posterior mode) is used so that, in this case, the estimator $\delta^\pi : \mathcal{X} \to \{1, \dots, M\}$ is defined by

$$\delta^\pi(x) \in \operatorname*{argmax}_{i \in \{1, \dots, M\}} \{p_i\, m_i(x)\}, \quad x \in \mathcal{X}.$$

Remark: The posterior distribution $\pi(i|x)$ is usually hard to compute (even with advanced Monte Carlo techniques).

Model choice as a testing problem

While the previous approach treats the problem of model choice as an estimation problem, the approach described below treats this problem as a testing problem.

As in Chapter 4, models $\mathcal{M}_1$ and $\mathcal{M}_2$ can be compared using the Bayes factor:

$$B_{12}^{\pi} = \frac{\pi(\{1\}|x)}{\pi(\{2\}|x)} \frac{p_2}{p_1} = \frac{\int_{\Theta_1} f_1(x|\theta_1)\pi_1(\theta_1)\mathrm{d}\theta_1}{\int_{\Theta_2} f_2(x|\theta_2)\pi_2(\theta_2)\mathrm{d}\theta_2} = \frac{m_1(x)}{m_2(x)}.$$

Remarks:

- Assume that for any i, j model $\mathcal{M}_i$ is preferred to model $\mathcal{M}_j$ when $B_{ij}^{\pi} > 1$. Then, since

  $$B_{ij}^{\pi} = B_{ik}^{\pi} B_{kj}^{\pi}$$

  the resulting model ordering is transitive.
- The quantity $m_i(x)$ is called the evidence of model i.
- This approach is useful only when M is small because it requires to compute $m_i(x)$ for all $i = 1, \dots, M$.
- The difficulties with this approach are the same as for hypothesis testing (Chapter 4), namely that improper and vague prior densities should me avoided.

Some comments on Bayesian model choice

The Bayesian solution of model choice is hard to justify from a purely Bayesian perspective.

Indeed,

- In the estimation approach of model choice, there should be some coherence in the choice of $(p_1, \ldots, p_M)$. For instance, if $\mathcal{M}_1 = \mathcal{M}_2 \cup \mathcal{M}_3$ we should have $\max(p_2, p_3) \leq p_1 \leq p_2 + p_3$. The construction of such a prior distribution is therefore complicated when M is large.
- In the testing approach of model choice, the Bayes factor does not depend on the prior probabilities $(p_1, \ldots, p_M)$ but one need to specify the thresholds $\tilde{a}_{ij}$ such that model i is preferred to model j when $B^{\pi}_{ij} > \tilde{a}_{ij}$. As for the prior probabilities $(p_1, \ldots, p_M)$, there should be some coherence in the choice of the $\tilde{a}_{ij}$'s and therefore the construction of these bounds is complicated when M is large.

For these reasons,

1. In practice we usually choose the model $\mu^* \in \operatorname{argmax}_{i \in 1:M} m_i(x)$ (which amounts to set $p_i = \frac{1}{M}$ for all i in the estimation approach and $\tilde{a}_{ij} = 1$ for all i, j in the testing approach).
2. Bayesian model choice is often justified using asymptotic arguments, as explained in the rest of this chapter.

Asymptotic expansion of the evidence

Let $X_1, \ldots, X_n$ be i.i.d. random variables with common density function $\tilde{f}(\cdot|\theta)$,

$$l_n(\theta) = \sum_{i=1}^{N} \log \tilde{f}(X_i|\theta), \quad \theta \in \Theta$$

be the log-likelihood function, $X^{(n)} = (X_1, \ldots, X_n)$ and $\hat{\theta}_n$ be the MLE of θ.

Then, under some regularity conditions (see Chapter 6),

$$\log m(X^{(n)}) = l_n(\hat{\theta}_n) - \frac{d}{2} \log n + \mathcal{O}_{\mathbb{P}}(1)$$

so that, as in the frequentist approach, the criterion used to carry out Bayesian model choice penalizes the number of parameters d.

Recall that the Bayesian information criterion (BIC) is defined by

$$BIC_n = -2l_n(\hat{\theta}_n) + d \log n$$

and therefore

$$\log m(X^{(n)}) = -\frac{BIC_n}{2} + \mathcal{O}_{\mathbb{P}}(1).$$

Consequently, selecting the model $i \in \{1, \ldots, M\}$ having the largest evidence $m_i(x)$ is asymptotically equivalent to choosing the model that minimizes the BIC criterion.

Important remark: Since it is known that (under suitable assumptions), the BIC criterion chooses the ‘true’ model with probability one as the number observations n tends to infinity, the above expansion of $\log m(X^{(n)})$ shows that selecting the model having the highest evidence is an asymptotically ‘correct’ procedure.

An example

Let $X_1,\dots,X_n$ be i.i.d. $\mathcal{N}_1(\theta_0,1)$ random variables for some $\theta_0\in\mathbb{R}$ and $f(\cdot|\theta)$ be the p.d.f. of $\mathcal{N}_n(\theta,I_n)$ distribution, with $\theta\in\mathbb{R}$.

We consider the two following two models for $X^{(n)}=(X_1,\dots,X_n)$

$$\mathcal{M}_1=\{f(\cdot|0)\},\quad \mathcal{M}_2=\{f(\cdot|\theta),\,\theta\in\mathbb{R}\setminus\{0\}\}$$

and we assume that $\pi_1(\theta)=\mathbf{1}_{\{0\}}(\theta)$ while $\pi_2(\theta)$ is the density of the $\mathcal{N}_1(\mu_0,\sigma_0^2)$ distribution.

Then, with $\bar{X}_n=\frac{1}{n}\sum_{i=1}^n X_i$,

$$\log B_{12}^{\pi}(X^{(n)})=-\frac{n\bar{X}_n^2}{2}+\frac{n(\bar{X}_n-\mu_0)^2}{2\sigma_0^2(n+1/\sigma_0^2)}+\frac{1}{2}\log\big(n\sigma_0^2+1\big),$$

so that, if $\theta_0=0$ and a $n\to+\infty$

$$\log B_{12}^{\pi}(X^{(n)})\to+\infty\quad\text{(in probability)}$$

at speed $\log n$ while, if $\theta_0\neq 0$,

$$\lim_{n\to+\infty}\log B_{12}^{\pi}(X^{(n)})=-\infty\quad\text{(almost surely)}$$

at speed n.

Proof of these results: See Problem Sheet 3, Problem 4.

VI - Bayesian asymptotics

In the frequentist perspective, most statistical procedures (estimators, tests, etc.) are evaluated through their *asymptotic properties*. This is because it is generally difficult to evaluate the frequentist properties of the considered procedure for a given sample size n.

In addition, as mentioned in Chapter 2, the frequentist risk does not allow to derive an optimal estimator from a decision theoretic point of view and thus asymptotic results are often the main justification for using a particular estimator.

Example: *Under mild conditions, the asymptotic properties of the maximum likelihood estimator (MLE) $\hat{\theta}_n$ are the following:*

1. *The MLE is convergent*

$$\hat{\theta}_n \rightarrow \theta_0 \textit{ in probability}$$

where θ_0 is the 'true' parameter.

2. *The MLE is asymptotically Normal,*

$$n^{1/2}(\hat{\theta}_n - \theta_0) \overset{\text{dist.}}{\Rightarrow} \mathcal{N}_d(0, I_1(\theta_0)^{-1})$$

as the number n of observations goes towards infinity.

The first result indicates that $\hat{\theta}_n$ gets closer and closer to θ_0 as the sample size increases. The second result gives a rough evaluation of the estimation error as $\sqrt{n} I_1(\theta_0)^{-1/2}$.

Asymptotics for Bayesian methods

In the Bayesian framework asymptotic properties are less an issue:

- From a pure Bayesian perspective, all the inference is done conditionally to the observations $x_1, \ldots, x_n$ and therefore asymptotic results are not relevant to choose an estimator.
- Bayesian estimators are justified from a decision theoretic point of view (see Chapter 2).
- The uncertainty on the value of θ is already expressed through the spread of the posterior distribution. In particular, the posterior variance (for instance) is the Bayesian estimator of the quadratic loss incurred by the use of the posterior mean (see Problem Sheet 2).
- In connection with the previous point, it is always possible to derive credible regions that are exactly of level $1 - \alpha$ in a Bayesian framework without resorting to asymptotic approximations (in contrast with frequentist confidence regions).

Nonetheless,

- A Bayes estimator which, for instance, would not be convergent would have little appeal.
- The asymptotic properties of Bayesian estimators explain why Bayesian methods are also appealing from a frequentist point of view.

The goal of this chapter is to present the most classical results on Bayesian asymptotics.

Set-up and notation

- We assume in this chapter that $X_1,\dots,X_n$ are i.i.d. from the distribution on $\mathcal{X}_1$ having density $\tilde{f}(\cdot|\theta_0)$ for some $\theta_0\in\Theta\subseteq\mathbb{R}^d$. In other words, we assume that the statistical model $\{\tilde{f}(\cdot|\theta),\theta\in\Theta\}$ is well-specified.
- In this chapter we use the shorthand $X^{(n)}=(X_1,\dots,X_n)$ so that the posterior distribution can be written as

$$\pi(\theta|X^{(n)})=\frac{f(X^{(n)}|\theta)\pi(\theta)}{\int_\Theta f(X^{(n)}|\theta)\pi(\theta)\mathrm{d}\theta},\quad f(X^{(n)}|\theta)=\prod_{i=1}^n\tilde{f}(X_i|\theta).$$

- As in Chapter 5, we denote by $l_n(\theta)$ the log-likelihood function; that is,

$$l_n(\theta)=\sum_{i=1}^n\log\tilde{f}(X_i|\theta),\quad \theta\in\Theta.$$

- As in Chapter 3, we denote by $\hat{\theta}_n$ the MLE of θ_0; that is

$$\hat{\theta}_n\in\operatorname*{argmax}_{\theta\in\Theta} l_n(\theta),$$

by $I_1(\theta)$ the Fisher information matrix for a single observation; that is

$$I_1(\theta)=\mathbb{E}_\theta\left[\frac{\partial\log\tilde{f}(X_1|\theta)}{\partial\theta}\frac{\partial\log\tilde{f}(X_1|\theta)}{\partial\theta^T}\right],\quad \theta\in\Theta,$$

and by $KL(\theta'|\theta)$ Kullback-Leibler (KL) divergence between $\tilde{f}(x_1|\theta')\mathrm{d}x_1$ and $\tilde{f}(x_1|\theta)\mathrm{d}x_1$; that is

$$KL(\theta'|\theta)=\mathbb{E}_{\theta'}\left[\log\frac{\tilde{f}(X_1|\theta')}{\tilde{f}(X_1|\theta)}\right],\quad (\theta,\theta')\in\Theta^2.$$

Consistency of posterior distributions

Definition 6.1 *We say that the sequence of posterior distributions* $\pi(\theta|X^{(n)})$ *is consistent if for every* $\epsilon > 0$ *we have*

$$\pi(\{\theta : \|\theta - \theta_0\| \geq \epsilon\}|X^{(n)}) \to 0, \quad \mathbb{P}_{\theta_0}\text{-}\textit{almost surely.}$$

Informally speaking, the posterior distribution is consistent if, as the sample size n increases, it puts more and more mass around θ_0.

The following lemma provides an alternative but equivalent definition of consistent posterior distributions.

Lemma 6.1 *A sequence of posterior distributions* $\pi(\theta|X^{(n)})$ *is consistent if and only if*

$$\pi(\theta|X^{(n)})\mathrm{d}\theta \overset{\text{dist.}}{\Rightarrow} \delta_{\theta_0}, \quad \mathbb{P}_{\theta_0}\text{-}\textit{almost surely.}$$

Proof: admitted.

In words, this lemma says that posterior consistency is equivalent to the convergence in distribution of the posterior distribution to a Dirac mass at θ_0.

Remark: As a corollary of Lemma 6.1, $\pi(\theta|X^{(n)})$ is consistent if and only if for any continuous and bounded function $g : \Theta \to \mathbb{R}$ we have

$$\lim_{n\to+\infty} \mathbb{E}_\pi[g(\theta)|X^{(n)}] = g(\theta_0), \quad \mathbb{P}_{\theta_0}\text{-almost surely.}$$

Schwartz's consistency theorem

The following well-known result provides a sufficient condition for posterior consistency.

Theorem 6.1 (Schwartz's theorem) *Assume the following:*

(A1) For every $\eta > 0$, $\pi\big(\{\theta : KL(\theta_0|\theta) \leq \eta\}\big) > 0$;

(A2) For every $\epsilon > 0$ *there exists a sequence of tests* $(\phi_n)_{n\geq 1}$ *(i.e.* $\phi_n : \mathcal{X}_1^n \to \{0,1\}$*) such that, for some constants* $D_1, D_2 \in \mathbb{R}_{>0}$,

$$\mathbb{E}_{\theta_0}\big[\phi_n(X^{(n)})\big] \leq e^{-nD_1}, \qquad \sup_{\{\theta:\, \|\theta-\theta_0\|\geq\epsilon\}} \mathbb{E}_\theta\big[1-\phi_n(X^{(n)})\big] \leq e^{-nD_2}.$$

Then, for every $\epsilon > 0$ *we have*

$$\pi(\{\theta : \|\theta - \theta_0\| \geq \epsilon\}|X^{(n)}) \to 0, \quad \mathbb{P}_{\theta_0}\text{-almost surely.}$$

Proof: See Appendix 1.

Assumption (A1) ensures that the prior distribution π puts some mass on a neighbourhood of θ_0 (otherwise there would be no hope to get posterior consistency!).

Assumption (A2) is about the identifiability of θ_0. It assumes the existence of tests $(\phi_n)_{n\geq 1}$ that separate the singleton $\{\theta_0\}$ from the alternative $\{\theta : \|\theta - \theta_0\| \geq \epsilon\}$ in an uniform fashion.

Remark: Surprisingly, (A2) is equivalent to

(A2') For every $\epsilon > 0$ there exists a sequence of tests $(\phi_n)_{n\geq 1}$ such that

$$\lim_{n\to+\infty} \mathbb{E}_{\theta_0}\big[\phi_n(X^{(n)})\big] = \lim_{n\to+\infty} \sup_{\{\theta:\, \|\theta-\theta_0\|\geq\epsilon\}} \mathbb{E}_\theta\big[1-\phi_n(X^{(n)})\big] = 0.$$

Convergence rate of the posterior distributions

Definition 6.2 *We say that the sequence of posterior distributions* $\pi(\theta|X^{(n)})$ *converges to* θ_0 *at rate* $\epsilon_n \to 0$ *if for all sequences* $M_n \to +\infty$ *we have*

$$\pi\big(\{\theta : \|\theta - \theta_0\| \geq M_n\epsilon_n\}|X^{(n)}\big) \to 0, \quad \textit{in } \mathbb{P}_{\theta_0}\textit{-probability.}$$

The following result illustrates how the convergence rate of the posterior distribution is related to the convergence rate of point-estimates derived from it (such as the posterior mean or the posterior median; see Chapter 2).

Lemma 6.2 *Assume that the sequence of posterior distributions* $\pi(\theta|X^{(n)})$ *converges to* θ_0 *at rate* $\epsilon_n \to 0$ *and let* $\tilde{\theta}_n$ *be the centre of the smallest ball that contains posterior mass of at least* $1/2$*. Then, for every sequence* $M_n \to +\infty$ *we have*

$$\mathbb{P}_{\theta_0}\big(\|\tilde{\theta}_n - \theta_0\| \leq 2M_n\epsilon_n\big) \to 1.$$

Proof: Done in class.

In other words, Lemma 6.2 shows that the convergence rate of the estimator $\tilde{\theta}_n$ is at least ϵ_n.

A classical result for the convergence rate of posterior distributions

Theorem 6.2 *Assume that for every sequence $M_n \to +\infty$ there exists a sequence of tests $(\phi_n)_{n\geq 1}$ such that $\mathbb{E}_{\theta_0}[\phi_n(X^{(n)})] \to 0$ and such that, for some constants $\epsilon > 0$ and $D > 0$, and for n large enough,*

$$\mathbb{E}_\theta[1 - \phi_n(X^{(n)})] \leq e^{-D(\|\theta-\theta_0\|^2 \wedge \epsilon)}$$

for all θ such that $\|\theta - \theta_0\| \geq M_n/\sqrt{n}$.

Then, under a set of technical conditions[a] *on $\{\tilde{f}(\cdot|\theta), \theta \in \Theta\}$ and on $\pi(\theta)$, we have*

$$\pi\big(\{\theta : \|\theta - \theta_0\| \geq M_n/\sqrt{n}\}|X^{(n)}\big) \to 0, \quad \textit{in } \mathbb{P}_{\theta_0}\textit{-probability}$$

for every sequence $M_n \to +\infty$.

Proof: See Problem Sheet 4 (for a particular model $\{\tilde{f}(\cdot|\theta), \theta \in \Theta\}$).

Remark: The assumption on the identifiability of θ_0 is stronger than in Schwartz's theorem (Theorem 6.1): Theorem 6.2 assumes the existence of tests ϕ_n that separate the singleton $\{\theta_0\}$ from the alternative $\{\|\theta - \theta_0\| > M_n/\sqrt{n}\}$, which is the complement of a shrinking ball around θ_0.

Remark: This result shows that the convergence rate of the estimator $\tilde{\theta}_n$ defined in Lemma 6.2 is at least $n^{-1/2}$.

[a] See e.g. Kleijn, B. J. K., and A. W. Van der Vaart. "The Bernstein-von-Mises theorem under misspecification." Electronic Journal of Statistics (2012): 354-381.

The Bernstein-von Mises theorem

Informally speaking, the Bernstein-von Mises theorem states that, as n increases, the posterior distribution behaves more and more like the $\mathcal{N}_d(\hat{\theta}_n, n^{-1}I_1(\theta_0)^{-1})$ distribution.

Theorem 6.3 (Bernstein von Mises theorem) *Under some technical conditions we have*

$$\sqrt{n}\big(\theta - \hat{\theta}_n)|X^{(n)} \overset{\text{dist.}}{\Rightarrow} \mathcal{N}_d(0, I_1(\theta_0)^{-1}), \quad \textit{in } \mathbb{P}_{\theta_0}\textit{-probability.}$$

Proof: See Appendix 2.

A first implication of this result is that a highest posterior density (HPD) region at level $(1-\alpha)$ (see Chapter 4) is asymptotically equivalent to a Wald $(1-\alpha)$-confidence interval based on the MLE; that is, a HPD at level $(1-\alpha)$ is a valid $(1-\alpha)$ confidence interval when the model is well-specified.

The following corollary gives the expansion of the log evidence we saw in Chapter 5.

Corollary 6.1 *Under some technical conditions,*

$$\log m(X^{(n)}) = l_n(\hat{\theta}_n) - \frac{d}{2}\log n + \mathcal{O}_{\mathbb{P}_{\theta_0}}(1).$$

Proof: See Appendix 3.

Posterior mean and maximum likelihood estimator

Theorem 6.4 *Under some technical conditions*[a] *we have*

$$\sqrt{n}\,(\mathbb{E}_\pi[\theta|X^{(n)}] - \hat{\theta}_n) \to 0, \quad \textit{in } \mathbb{P}_{\theta_0}\textit{-probability}$$

and therefore, if $\sqrt{n}\,(\hat{\theta}_n - \theta_0) \overset{\text{dist.}}{\Rightarrow} \mathcal{N}_d(0, I_1(\theta_0)^{-1})$, *we have*

$$\sqrt{n}\,(\mathbb{E}_\pi[\theta|X^{(n)}] - \theta_0) \overset{\text{dist.}}{\Rightarrow} \mathcal{N}_d(0, I_1(\theta_0)^{-1})$$

Proof: Admitted.

Remarks:

1. The first result shows that the difference between the MLE $\hat{\theta}_n$ and $\mathbb{E}_\pi[\theta|X^{(n)}]$ decreases quickly with n (i.e. faster than $n^{-1/2}$).
2. The second result shows that the estimator $\mathbb{E}_\pi[\theta|X^{(n)}]$ has the same asymptotic distribution as the MLE.

Altogether, these two results show that the estimator $\mathbb{E}_\pi[\theta|X^{(n)}]$ is asymptotically equivalent to the MLE, justifying the use of Bayesian methods from a frequentist perspective.

[a] See e.g. Chapter 1 of Ghosh, J. K. and Ramamoorthi, R.V. *Bayesian Nonparametrics.* Springer-Verlag New York (2013).

Bayesian asymptotics under misspecified models

In this chapter we assumed that the model was well-specified; that is, that there exists a $\theta_0 \in \Theta$ such that $X_1 \sim \tilde{f}(x_1|\theta_0)\mathrm{d}x_1$.

In practice, this assumption is never verified meaning that a model is always misspecified and it is therefore important to understand the asymptotic behaviour of Bayesian quantities in this context.

Let $\theta_0 \in \mathrm{argmax}_{\theta\in\Theta}\, \mathbb{E}_0[\log \tilde{f}(X_1|\theta)]$ and

$$V_{\theta_0} = -\frac{\partial^2 \mathbb{E}_0\big[\log \tilde{f}(X_1|\theta_0)\big]}{\partial\theta\,\partial\theta^T}, \quad \dot{l}_{\theta_0}(X_1) = \frac{\partial \log \tilde{f}(X_1|\theta_0)}{\partial\theta},$$

where the notation $\mathbb{E}_0$ is used to denote expectation under the true distribution of X_1.

Remark: If the model is well specified then θ_0 is as before the true parameter and $V_{\theta_0} = I_1(\theta_0)$ (under some additional conditions).

Then, under some technical conditions,

- the MLE is such that $\hat{\theta}_n \to \theta_0$ in $\mathbb{P}_0$-probability while

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \overset{\text{dist.}}{\Rightarrow} \mathcal{N}_d\big(0, V_{\theta_0}^{-1}\mathbb{E}_0[\dot{l}_{\theta_0}(X_1)\dot{l}_{\theta_0}(X_1)^T]V_{\theta_0}^{-1}\big)$$

- the Bernstein von Misses theorem (Theorem 6.3) becomes

$$\sqrt{n}\big(\theta - \hat{\theta}_n\big)|X^{(n)} \overset{\text{dist.}}{\Rightarrow} \mathcal{N}_d(0, V_{\theta_0}^{-1}), \quad \text{in } \mathbb{P}_{\theta_0}\text{-probability}$$

  while the results in Theorem 6.1 and in Theorem 6.2 still hold[a].

Consequently,

1. The two ‘procedures’ converge to θ_0.
2. Since in general $V_{\theta_0}^{-1} \neq V_{\theta_0}^{-1}\mathbb{E}_0[\dot{l}_{\theta_0}(X_1)\dot{l}_{\theta_0}(X_1)^T]V_{\theta_0}^{-1}$, a HPD at level $(1-\alpha)$ is not a valid $(1-\alpha)$ confidence interval when the model misspecified.

[a] See Kleijn, B. J. K., and A. W. Van der Vaart. “The Bernstein-von-Mises theorem under misspecification.” Electronic Journal of Statistics (2012): 354-381.

Appendix 1: Proof of Theorem 6.1

We start with two preliminary results.

Lemma 6.3 *Let* $(Y_n)_{n\geq 1}$ *be a sequence of random variables such that*

$$\sum_{n=1}^{\infty} \mathbb{P}(|Y_n| \geq \epsilon) < +\infty, \quad \forall \epsilon > 0.$$

Then, $\lim_{n\to+\infty} Y_n = 0$, $\mathbb{P}$*-almost surely.*

Proof: This is a direct consequence of the Borel-Cantelli lemma.

Lemma 6.4 *Let* $(g_n)_{n\geq 1}$ *be a sequence of non-negative (measurable) functions* $g_n : \Theta \to [0, +\infty)$. *Then,*

$$\liminf_{n\to+\infty} \int_{\Theta} g_n(\theta)\pi(\theta)\mathrm{d}\theta \geq \int_{\Theta} \liminf_{n\to+\infty} g_n(\theta)\pi(\theta)\mathrm{d}\theta.$$

Proof: This is a direct consequence of Fatou's lemma.

To prove Theorem 6.1, remark first that, for every $\epsilon > 0$,

$$\sum_{n=1}^{\infty} \mathbb{P}_{\theta_0}(|\phi_n(X^{(n)})| \geq \epsilon) \leq \sum_{n=1}^{\infty} \frac{\mathbb{E}_{\theta_0}[\phi_n(X^{(n)})]}{\epsilon} \leq \sum_{n=1}^{\infty} \frac{e^{-nD_1}}{\epsilon} < +\infty$$

where the first inequality uses Markov's inequality (see Problem Sheet 4, Problem 2) while the second one holds under (A1).

Then, by Lemma 6.3,

$$\lim_{n\to+\infty} \phi_n(X^{(n)}) = 0, \quad \mathbb{P}_{\theta_0}\text{-almost surely.} \tag{1}$$

Appendix 1: Proof of Theorem 6.1 (continued)

Next, let $\epsilon > 0$ and $V = \big\{\theta : \ \|\theta - \theta_0\| \geq \epsilon\big\}$. We show below that

$$\limsup_{n\to+\infty} \pi(V|X^{(n)}) = 0 \quad \mathbb{P}_{\theta_0}\text{-almost surely.}$$

To this end note first that ($\mathbb{P}_{\theta_0}$-a.s.),

$$\begin{aligned}
&\limsup_{n\to+\infty} \pi(V|X^{(n)}) \\
&\leq \limsup_{n\to+\infty} \phi_n(X^{(n)})\pi(V|X^{(n)}) + \limsup_{n\to+\infty} \pi(V|X^{(n)})\big(1-\phi_n(X^{(n)})\big) \\
&\leq \limsup_{n\to+\infty} \phi_n(X^{(n)}) + \limsup_{n\to+\infty} \pi(V|X^{(n)})\big(1-\phi_n(X^{(n)})\big) \\
&= \limsup_{n\to+\infty} \pi(V|X^{(n)})\big(1-\phi_n(X^{(n)})\big)
\end{aligned} \tag{2}$$

where the equality follows from (1).

Next, write

$$\pi(V|X^{(n)})(1-\phi_n) = \frac{(1-\phi_n)\int_V \pi(\theta)\prod_{k=1}^n \frac{\tilde{f}(X_k|\theta)}{\tilde{f}(X_k|\theta_0)}\mathrm{d}\theta}{\int_\Theta \pi(\theta)\prod_{k=1}^n \frac{\tilde{f}(X_k|\theta)}{\tilde{f}(X_k|\theta_0)}\mathrm{d}\theta} \tag{3}$$

and we now study the denominator of the term appearing on the r.h.s. of the equality sign.

Let $\eta > 0$, $K_\eta = \big\{\theta : \ KL(\theta_0|\theta) \leq \eta\big\}$ and $\theta \in K_\eta$. (Remark that under (A1) $\pi(K_\delta) > 0$ and thus $K_\delta \neq \emptyset$). Then, by the law of large numbers,

$$\lim_{n\to+\infty}\left|\frac{1}{n}\sum_{k=1}^n \log\frac{\tilde{f}(X_k|\theta)}{\tilde{f}(X_k|\theta_0)} - \mathbb{E}_{\theta_0}\left[\log\frac{\tilde{f}(X_1|\theta)}{\tilde{f}(X_1|\theta_0)}\right]\right| = 0, \quad \mathbb{P}_{\theta_0} - a.s.$$

so that, for any $\delta > 0$, there exists ($\mathbb{P}_{\theta_0}$-a.s) an $n_\delta \geq 1$ such that

$$\frac{1}{n}\sum_{k=1}^n \log\frac{\tilde{f}(X_k|\theta)}{\tilde{f}(X_k|\theta_0)} \geq -KL(\theta_0|\theta) - \delta, \quad \forall n \geq n_\delta.$$

Appendix 1: Proof of Theorem 6.1 (continued)

Since $\theta \in K_\eta$, we have $KL(\theta_0|\theta) \leq \eta$ and thus ($\mathbb{P}_{\theta_0}$-a.s)

$$\frac{1}{n}\sum_{k=1}^{n} \log \frac{\tilde{f}(X_k|\theta)}{\tilde{f}(X_k|\theta_0)} \geq -(\delta+\eta), \quad \forall n \geq n_\delta.$$

Applying this result with $\delta = \eta$ implies that, for any $\theta \in K_\eta$, there exists ($\mathbb{P}_{\theta_0}$-a.s) a $n_\delta \geq 1$ such that

$$\frac{1}{n}\sum_{k=1}^{n} \log \frac{\tilde{f}(X_k|\theta)}{\tilde{f}(X_k|\theta_0)} \geq -2\eta \Leftrightarrow \prod_{k=1}^{n} \frac{\tilde{f}(X_k|\theta)}{\tilde{f}(X_k|\theta_0)} \geq e^{-n2\eta}, \quad \forall n \geq n_\delta.$$

Therefore ($\mathbb{P}_{\theta_0}$-a.s),

$$\begin{aligned}
\liminf_{n\to+\infty} e^{2n\eta} \int_\Theta \pi(\theta) \prod_{k=1}^{n} \frac{\tilde{f}(X_k|\theta)}{\tilde{f}(X_k|\theta_0)} \mathrm{d}\theta &\geq \liminf_{n\to+\infty} e^{2n\eta} \int_{K_\eta} \pi(\theta) \prod_{k=1}^{n} \frac{\tilde{f}(X_k|\theta)}{\tilde{f}(X_k|\theta_0)} \mathrm{d}\theta \\
&\geq \int_{K_\eta} \pi(\theta) \liminf_{n\to+\infty} e^{2n\eta} \prod_{k=1}^{n} \frac{\tilde{f}(X_k|\theta)}{\tilde{f}(X_k|\theta_0)} \mathrm{d}\theta \\
&= \pi(K_\eta)
\end{aligned}$$

where the second inequality uses Lemma 6.4.

Then, using (3) we have ($\mathbb{P}_{\theta_0}$-a.s),

$$\begin{aligned}
&\limsup_{n\to+\infty} \pi(V|X^{(n)})(1-\phi_n(X^{(n)})) \\
&\leq \frac{\limsup_{n\to+\infty} e^{2n\eta}\big(1-\phi_n(X^{(n)})\big) \int_V \pi(\theta) \prod_{k=1}^{n} \frac{\tilde{f}(X_k|\theta)}{\tilde{f}(X_k|\theta_0)} \mathrm{d}\theta}{\liminf_{n\to+\infty} e^{2n\eta} \int_\Theta \pi(\theta) \prod_{k=1}^{n} \frac{\tilde{f}(X_k|\theta)}{\tilde{f}(X_k|\theta_0)} \mathrm{d}\theta} \qquad (4) \\
&\leq \frac{\limsup_{n\to+\infty} e^{2n\eta}\big(1-\phi_n(X^{(n)})\big) \int_V \pi(\theta) \prod_{k=1}^{n} \frac{\tilde{f}(X_k|\theta)}{\tilde{f}(X_k|\theta_0)} \mathrm{d}\theta}{\pi(K_\eta)}.
\end{aligned}$$

where $\pi(K_\eta) > 0$ under (A1).

Appendix 1: Proof of Theorem 6.1 (end)

To proceed further let

$$Z_n = \big(1-\phi_n(X^{(n)})\big)\int_V \pi(\theta)\prod_{k=1}^n \frac{\tilde{f}(X_k|\theta)}{\tilde{f}(X_k|\theta_0)}\mathrm{d}\theta$$

and remark that

$$\begin{aligned}
\mathbb{E}_{\theta_0}[Z_n] &= \int_{\mathcal{X}_1^n}\left(\big(1-\phi_n(x^{(n)})\big)\int_V \pi(\theta)\prod_{k=1}^n \frac{\tilde{f}(x_k|\theta)}{\tilde{f}(x_k|\theta_0)}\mathrm{d}\theta\right)\prod_{i=1}^n \tilde{f}(x_i|\theta_0)\mathrm{d}x_i\\
&= \int_{\mathcal{X}_1^n}\left(\big(1-\phi_n(x^{(n)})\big)\int_V \pi(\theta)\prod_{k=1}^n \tilde{f}(x_k|\theta)\mathrm{d}\theta\right)\prod_{i=1}^n \mathrm{d}x_i\\
&= \int_V\left(\int_{\mathcal{X}_1^n}\big(1-\phi_n(x^{(n)})\big)\prod_{i=1}^n \tilde{f}(x_i|\theta)\mathrm{d}x_i\right)\pi(\theta)\mathrm{d}\theta\\
&= \int_V \mathbb{E}_\theta[1-\phi_n(X^{(n)})]\pi(\theta)\mathrm{d}\theta\\
&\leq \sup_{\theta\in V}\mathbb{E}_\theta[1-\phi_n(X^{(n)})]\\
&\leq e^{-nD_2}
\end{aligned}$$

where the third equality holds by Tonelli's theorem and the last inequality holds under (A2).

Then, taking $\eta>0$ sufficiently small so that $\beta := D_2-2\eta>0$, we have (using Markov's inequality for the first inequality)

$$\sum_{n=1}^\infty \mathbb{P}_{\theta_0}(|e^{2n\eta}Z_n|\geq\epsilon)\leq\sum_{n=1}^\infty\frac{e^{2n\eta}\mathbb{E}_{\theta_0}[Z_n]}{\epsilon}\leq\sum_{n=1}^\infty\frac{e^{-n\beta}}{\epsilon}<+\infty,\quad\forall\epsilon>0.$$

Consequently, by Lemma 6.3,

$$\lim_{n\to+\infty} e^{2n\eta}Z_n=0,\quad \mathbb{P}_{\theta_0}\text{-almost surely}$$

which, together with (2) and (4), completes the proof.

Appendix 2[a]: A proof of Theorem 6.3

We assume in this Appendix that $\Theta = \mathbb{R}$ (and thus that $d = 1$). This assumption is made to simplify the presentation but what follows can be easily generalized to any $d \geq 1$.

Below we shall consider the following assumptions.

(B1) $\tilde{f}(x|\theta) > 0$ for all $(x, \theta) \in \mathcal{X}_1 \times \Theta$.

(B2) $l(\theta, x) := \log \tilde{f}(x|\theta)$ is thrice differentiable with respect to θ in a neighbourhood $(\delta_0 - \theta_0, \theta_0 + \delta_0)$ of θ_0. If $\dot{l}$, $\ddot{l}$ and $\dddot{l}$ stand for the first, second and third derivatives, then $\mathbb{E}_{\theta_0}[\dot{l}(X_1, \theta_0)]$ and $\mathbb{E}_{\theta_0}[\ddot{l}(X_1, \theta_0)]$ are both finite and

$$\sup_{\theta \in (\delta_0 - \theta_0, \theta_0 + \delta_0)} |\dddot{l}(X_1, \theta_0)| \leq M(x), \quad \text{and } \mathbb{E}_{\theta_0}[M(X_1)] < +\infty.$$

(B3) Interchange of the order of expectation with respect to θ_0 and differentiation at θ_0 are justified, so that

$$\mathbb{E}_{\theta_0}[\dot{l}(X_1, \theta_0)] = 0, \quad \mathbb{E}_{\theta_0}[\ddot{l}(X_1, \theta_0)] = -\mathbb{E}_{\theta_0}[\dot{l}(X_1, \theta_0)]^2.$$

(B4) $I_1(\theta_0) = \mathbb{E}_{\theta_0}[\dot{l}(X_1, \theta_0)]^2 > 0$.

(B5) $\hat{\theta}_n \to \theta_0$ in $\mathbb{P}_{\theta_0}$-probability.

(B6) For any $\delta > 0$ there exists an $\epsilon > 0$ such that

$$\lim_{n \to +\infty} \mathbb{P}_{\theta_0}\Big(\sup_{\{\theta:\, |\theta - \theta_0| > \delta\}} \frac{1}{n}\big(l_n(\theta) - l_n(\theta_0)\big) \leq -\epsilon \Big) = 1.$$

(B7) The prior distribution has density $\pi(\theta)$ (w.r.t. the Lebesgue measure) which is continuous and positive at θ_0.

[a] This Appendix is based on Chapter 1 of Ghosh, J. K. and Ramamoorthi, R.V. *Bayesian Nonparametrics.* Springer-Verlag New York (2013).

Appendix 2: A proof of Theorem 6.3 (continued)

Let $H = \sqrt{n}(\theta - \hat{\theta}_n)$. Then, if θ has density $\pi(\theta|X^{(n)})$ then, by the change of variable formula, H has density $\pi^*(h|X^{(n)})$ defined by

$$\pi^*(h|X^{(n)}) = \frac{\pi\big(\hat{\theta}_n + \frac{h}{\sqrt{n}}\big)\prod_{k=1}^n \tilde{f}\big(X_k|\hat{\theta}_n + \frac{h}{\sqrt{n}}\big)}{\int_{\mathbb{R}} \pi\big(\hat{\theta}_n + \frac{h'}{\sqrt{n}}\big)\prod_{k=1}^n \tilde{f}\big(X_i|\hat{\theta}_n + \frac{h'}{\sqrt{n}}\big)\mathrm{d}h'}, \quad h \in \mathbb{R}.$$

Then, we have the following result.

Theorem 6.5 *Assume $\Theta = \mathbb{R}$ and that (B1)-(B6) hold. Then,*

$$\int_{\mathbb{R}^d} \left|\pi^*(h|X^{(n)}) - \sqrt{\frac{I_1(\theta_0)}{2\pi}} e^{-\frac{h^2 I_1(\theta_0)}{2}}\right| \mathrm{d}h \to 0, \quad \textit{in } \mathbb{P}_{\theta_0}\textit{-probability.}$$

Proof: The proof follows from Lemme 6.5-6.9 stated and proved below.

Remark: The convergence in Theorem 6.5 is for the total variation metric which is stronger than the weak convergence.

Appendix 2: A proof of Theorem 6.3 (continued)

Lemma 6.5 *Assume (B4). Then, a sufficient condition for the result of Theorem 6.5 to hold is that*

$$\int_{\mathbb{R}} \left| \pi\big(\hat{\theta}_n + \frac{h}{\sqrt{n}}\big) e^{l_n(\hat{\theta}_n + h/\sqrt{n}) - l_n(\hat{\theta}_n)} - \pi(\theta_0) e^{-\frac{h^2 I_1(\theta_0)}{2}} \right| \mathrm{d}h \to 0 \qquad (5)$$

in $\mathbb{P}_{\theta_0}$-probability.

Proof: Let $C_n = \int_{\mathbb{R}} \pi\big(\hat{\theta}_n + \frac{h}{\sqrt{n}}\big) e^{l_n(\hat{\theta}_n + h/\sqrt{n}) - l_n(\hat{\theta}_n)} \mathrm{d}h$. Then, noting that $\pi^*(h|X^{(n)})$ can be rewritten as

$$\pi^*(h|X^{(n)}) = \frac{\pi\big(\hat{\theta}_n + \frac{h}{\sqrt{n}}\big) e^{l_n(\hat{\theta}_n + h/\sqrt{n}) - l_n(\hat{\theta}_n)}}{C_n}$$

we have

$$\int_{\mathbb{R}^d} \left| \pi^*(h|X^{(n)}) - \sqrt{\frac{I_1(\theta_0)}{2\pi}} e^{-\frac{h^2 I_1(\theta_0)}{2}} \right| \mathrm{d}h$$
$$= C_n^{-1} \int_{\mathbb{R}} \left| \pi\big(\hat{\theta}_n + \frac{h}{\sqrt{n}}\big) e^{l_n(\hat{\theta}_n + h/\sqrt{n}) - l_n(\hat{\theta}_n)} - C_n \sqrt{\frac{I_1(\theta_0)}{2\pi}} e^{-\frac{h^2 I_1(\theta_0)}{2}} \right| \mathrm{d}h.$$

By (5), $C_n \to \pi(\theta_0)\sqrt{2\pi/I_1(\theta_0)}$ in $\mathbb{P}_{\theta_0}$-probability and thus to prove the lemma it is enough to show that, in $\mathbb{P}_{\theta_0}$-probability,

$$I^{(n)} := \int_{\mathbb{R}} \left| \pi\big(\hat{\theta}_n + \frac{h}{\sqrt{n}}\big) e^{l_n(\hat{\theta}_n + h/\sqrt{n}) - l_n(\hat{\theta}_n)} - C_n \sqrt{\frac{I_1(\theta_0)}{2\pi}} e^{-\frac{h^2 I_1(\theta_0)}{2}} \right| \mathrm{d}h \to 0.$$

Appendix 2: A proof of Theorem 6.3 (continued)

To this end let

$$I_1^{(n)} = \int_{\mathbb{R}} \left| \pi\big(\hat{\theta}_n + \frac{h}{\sqrt{n}}\big) e^{l_n(\hat{\theta}_n + h/\sqrt{n}) - l_n(\hat{\theta}_n)} - \pi(\theta_0) e^{-\frac{h^2 I_1(\theta_0)}{2}} \right| \mathrm{d}h$$

$$I_2^{(n)} = \int_{\mathbb{R}} \left| \pi(\theta_0) e^{-\frac{h^2 I_1(\theta_0)}{2}} - C_n \sqrt{\frac{I_1(\theta_0)}{2\pi}} e^{-\frac{h^2 I_1(\theta_0)}{2}} \right| \mathrm{d}h$$

so that $I \leq I_1 + I_2$ where, by (5), $I_1^{(n)} \to 0$ in $\mathbb{P}_{\theta_0}$-probability.

In addition, $I_2^{(n)}$ can be rewritten as

$$I_2^{(n)} = \left| \pi(\theta_0) - C_n \sqrt{\frac{I_1(\theta_0)}{2\pi}} \right| \int_{\mathbb{R}} e^{-\frac{h^2 I_1(\theta_0)}{2}} \mathrm{d}h$$

and thus, as $C_n \to \pi(\theta_0)\sqrt{2\pi / I_1(\theta_0)}$ in $\mathbb{P}_{\theta_0}$-probability while, under (B4)

$$\int_{\mathbb{R}} e^{-\frac{h^2 I_1(\theta_0)}{2}} \mathrm{d}h < +\infty,$$

we have $I_2^{(n)} \to 0$ in $\mathbb{P}_{\theta_0}$-probability and the proof of Lemma 6.5 is complete.

Appendix 2: A proof of Theorem 6.3 (continued)

Lemma 6.6 *Assume (B1)-(B6) and let* $I_n = -\frac{1}{n}\sum_{i=1}^n \ddot{l}(X_i, \hat{\theta}_n)$. *Then, a sufficient condition for* (5) *to hold is that*

$$\int_{\mathbb{R}} \left| \pi\big(\hat{\theta}_n + \frac{h}{\sqrt{n}}\big) e^{l_n(\hat{\theta}_n + h/\sqrt{n}) - l_n(\hat{\theta}_n)} - \pi(\hat{\theta}_n) e^{-\frac{h^2 I_n}{2}} \right| \mathrm{d}h \to 0$$

in $\mathbb{P}_{\theta_0}$*-probability.*

Proof: Under (B2) and (B5), with $\mathbb{P}_{\theta_0}$-probability tending to one we have (using the mean value theorem)

$$I_n = -\frac{1}{n}\sum_{i=1}^n \ddot{l}(X_i, \theta_0) - (\hat{\theta}_n - \theta_0)\frac{1}{n}\sum_{i=1}^n \dddot{l}(X_i, \theta_n')$$

for some θ_n' between $\hat{\theta}_n$ and θ_0. Hence, with $\mathbb{P}_{\theta_0}$-probability tending to one,

$$\Big| I_n - \Big(-\frac{1}{n}\sum_{i=1}^n \ddot{l}(X_i, \theta_0) \Big) \Big| \leq |\hat{\theta}_n - \theta_0| \frac{1}{n}\sum_{i=1}^n M(X_i)$$

where , under (B2),

$$\lim_{n\to+\infty} \frac{1}{n}\sum_{i=1}^n M(X_i) = \mathbb{E}_{\theta_0}[M(X_1)] < +\infty$$

while, under (B3),

$$\lim_{n\to+\infty} \frac{1}{n}\sum_{i=1}^n \ddot{l}(X_i, \theta_0) = -I_1(\theta_0).$$

Consequently,

$$I_n \to I_1(\theta_0), \qquad \text{in } \mathbb{P}_{\theta_0}\text{-probability.} \tag{6}$$

Appendix 2: A proof of Theorem 6.3 (continued)

Next, remark that

$$\begin{aligned}
&\int_{\mathbb{R}} \left|\pi(\theta_0)e^{-\frac{h^2 I_1(\theta_0)}{2}} - \pi(\hat{\theta}_n)e^{-\frac{h^2 I_n}{2}}\right| \mathrm{d}h \\
&\leq \pi(\theta_0)\int_{\mathbb{R}} \left|e^{-\frac{h^2 I_1(\theta_0)}{2}} - e^{-\frac{h^2 I_n}{2}}\right| \mathrm{d}h + \left|\pi(\theta_0) - \pi(\hat{\theta}_n)\right| \int_{\mathbb{R}} e^{-\frac{h^2 I_n}{2}} \mathrm{d}h \qquad (7)\\
&= \pi(\theta_0)\int_{\mathbb{R}} \left|e^{-\frac{h^2 I_1(\theta_0)}{2}} - e^{-\frac{h^2 I_n}{2}}\right| \mathrm{d}h + \left|\pi(\theta_0) - \pi(\hat{\theta}_n)\right| \sqrt{\frac{2\pi}{I_n}}
\end{aligned}$$

where, using (6) and under (B4), (B5) and (B7), the second term appearing on the r.h.s. of the last inequality sign converges to zero in $\mathbb{P}_{\theta_0}$-probability.

In addition, using the fact that $|e^x - 1| \leq |x|e^{|x|}$ for all $x \in \mathbb{R}$,

$$\begin{aligned}
\int_{\mathbb{R}} \left|e^{-\frac{h^2 I_1(\theta_0)}{2}} - e^{-\frac{h^2 I_n}{2}}\right| \mathrm{d}h &= \int_{\mathbb{R}} e^{-\frac{h^2 I_1(\theta_0)}{2}} \left|1 - e^{-\frac{h^2 (I_n - I_1(\theta_0))}{2}}\right| \mathrm{d}h \\
&\leq \frac{\left|I_n - I_1(\theta_0)\right|}{2} \int_{\mathbb{R}} h^2 e^{-\frac{h^2 (I_1(\theta_0) - |I_n - I_1(\theta_0)|)}{2}} \mathrm{d}h.
\end{aligned}$$

Using (6) and under (B4), with $\mathbb{P}_{\theta_0}$-probability tending to one $I_1(\theta_0) - |I_n - I_1(\theta_0)| > \epsilon$ for some $\epsilon > 0$ so that, with $\mathbb{P}_{\theta_0}$-probability tending to one,

$$\int_{\mathbb{R}} h^2 e^{-\frac{h^2 (I_1(\theta_0) - |I_n - I_1(\theta_0)|)}{2}} \mathrm{d}h \leq C$$

for some $C < +\infty$. Together with (6), this shows that

$$\int_{\mathbb{R}} \left|e^{-\frac{h^2 I_1(\theta_0)}{2}} - e^{-\frac{h^2 I_n}{2}}\right| \mathrm{d}h \to 0, \qquad \text{in } \mathbb{P}_{\theta_0}\text{-probability.}$$

Then, the result follows from (7) and the triangle inequality.

Appendix 2: A proof of Theorem 6.3 (continued)

Lemma 6.7 *Assume (B1)-(B6) and let* $\delta > 0$ *and* $A_1 = \{h \in \mathbb{R} : |h| > \delta/\sqrt{n}\}$*. Then,*

$$\int_{A_1} \left| \pi\big(\hat\theta_n + \frac{h}{\sqrt{n}}\big) e^{l_n(\hat\theta_n + h/\sqrt{n}) - l_n(\hat\theta_n)} - \pi(\hat\theta_n) e^{-\frac{h^2 I_n}{2}} \right| \mathrm{d}h \to 0$$

in $\mathbb{P}_{\theta_0}$*-probability and where* I_n *is as in Lemma 6.6.*

Proof: We have

$$\int_{A_1} \left| \pi\big(\hat\theta_n + \frac{h}{\sqrt{n}}\big) e^{l_n(\hat\theta_n + h/\sqrt{n}) - l_n(\hat\theta_n)} - \pi(\hat\theta_n) e^{-\frac{h^2 I_n}{2}} \right| \mathrm{d}h$$
$$\leq \int_{A_1} \pi\big(\hat\theta_n + \frac{h}{\sqrt{n}}\big) e^{l_n(\hat\theta_n + h/\sqrt{n}) - l_n(\hat\theta_n)} \mathrm{d}h + \pi(\hat\theta_n) \int_{A_1} e^{-\frac{h^2 I_n}{2}} \mathrm{d}h$$

where the first integral converges to zero in $\mathbb{P}_{\theta_0}$-probability under (B6) while under (B4) the second one converges to zero in $\mathbb{P}_{\theta_0}$-probability using (6) and usual tail estimates for the normal distribution.

Appendix 2: A proof of Theorem 6.3 (continued)

Lemma 6.8 *Assume (B1)-(B6) and let* $c > 0$ *and* $A_2 = \{h \in \mathbb{R} : |h| < c\log(\sqrt{n})\}$. *Then,*

$$\int_{A_2} \left| \pi\big(\hat{\theta}_n + \frac{h}{\sqrt{n}}\big) e^{l_n(\hat{\theta}_n + h/\sqrt{n}) - l_n(\hat{\theta}_n)} - \pi(\hat{\theta}_n) e^{-\frac{h^2 I_n}{2}} \right| \mathrm{d}h \to 0$$

in $\mathbb{P}_{\theta_0}$*-probability and where* I_n *is as in Lemma 6.6.*

Proof: Because $\hat{\theta}_n \to \theta_0$ in $\mathbb{P}_{\theta_0}$-probability under (B5), with $\mathbb{P}_{\theta_0}$-probability tending to one we have, under (B2) and using a second order Taylor expansion around $\hat{\theta}_n$,

$$\begin{aligned} l_n(\hat{\theta}_n + h/\sqrt{n}) &= l_n(\hat{\theta}_n) + \frac{1}{2}\Big(\frac{h}{\sqrt{n}}\Big)^2 \sum_{i=1}^n \ddot{l}(X_i, \hat{\theta}_n) \\ &\quad + \frac{1}{6}\Big(\frac{h}{\sqrt{n}}\Big)^3 \sum_{i=1}^n \dddot{l}(X_i, \theta_n') \\ &= -\frac{h^2 I_n}{2} + R_n(h) \end{aligned} \tag{8}$$

for some θ_n' between θ_0 and $\hat{\theta}_n$ and with

$$R_n(h) = \frac{1}{6}\Big(\frac{h}{\sqrt{n}}\Big)^3 \sum_{i=1}^n \dddot{l}(X_i, \theta_n'). \tag{9}$$

To proceed further remark that, for n large enough,

$$\begin{aligned} \sup_{h\in A_2} R_n(h) &= \sup_{h\in A_2} \frac{1}{6}\Big(\frac{h}{\sqrt{n}}\Big)^3 \sum_{i=1}^n \dddot{l}(X_i, \theta_n') \\ &\le \frac{c^3}{6}\frac{(\log(\sqrt{n}))^3}{n}\Big(\frac{1}{\sqrt{n}}\sum_{i=1}^n M(X_i)\Big) \\ &\to 0 \end{aligned} \tag{10}$$

in $\mathbb{P}_{\theta_0}$-probability since, under (B2), $\frac{1}{\sqrt{n}}\sum_{i=1}^n M(X_i) \overset{\text{dist.}}{\Rightarrow} \mathcal{N}_1\big(\mathbb{E}_{\theta_0}[M(X_1)], \mathbb{E}_{\theta_0}[M(X_1)^2] - \mathbb{E}_{\theta_0}[M(X_1)]^2\big)$.

Appendix: A proof of Theorem 6.3 (continued)

Then, using (8),

$$\begin{aligned}
&\int_{A_2} \left|\pi\big(\hat{\theta}_n + \frac{h}{\sqrt{n}}\big)e^{l_n(\hat{\theta}_n + h/\sqrt{n}) - l_n(\hat{\theta}_n)} - \pi\big(\hat{\theta}_n\big)e^{-\frac{h^2 I_n}{2}}\right| \mathrm{d}h \\
&\quad = \int_{A_2} \left|\pi\big(\hat{\theta}_n + \frac{h}{\sqrt{n}}\big)e^{-\frac{h^2 I_n}{2} + R_n(h)} - \pi\big(\hat{\theta}_n\big)e^{-\frac{h^2 I_n}{2}}\right| \mathrm{d}h \\
&\quad \le \int_{A_2} \pi\big(\hat{\theta}_n + \frac{h}{\sqrt{n}}\big)e^{-\frac{h^2 I_n}{2}}\left|e^{R_n(h)} - 1\right| \mathrm{d}h \\
&\quad + \sup_{h \in A_2}\left|\pi\big(\hat{\theta}_n + \frac{h}{\sqrt{n}}\big) - \pi\big(\hat{\theta}_n\big)\right| \int_{A_2} e^{-\frac{h^2 I_n}{2}} \mathrm{d}h.
\end{aligned}$$

Under (B5) and (B7),

$$\sup_{h \in A_2}\left|\pi\big(\hat{\theta}_n + \frac{h}{\sqrt{n}}\big) - \pi\big(\hat{\theta}_n\big)\right| \to 0, \quad \text{in } \mathbb{P}_{\theta_0}\text{-probability}$$

while, using (6) and under (B4), with $\mathbb{P}_{\theta_0}$-probability tending to one $\int_{A_2} e^{-\frac{h^2 I_n}{2}} \mathrm{d}h < C$ for some $C < +\infty$. Hence,

$$\int_{A_2} \pi\big(\hat{\theta}_n + \frac{h}{\sqrt{n}}\big)\left|e^{-\frac{h^2 I_n}{2} + R_n(h)} - e^{-\frac{h^2 I_n}{2}}\right| \mathrm{d}h \to 0$$

in $\mathbb{P}_{\theta_0}$-probability.

In addition, using the fact that $|e^x - 1| \le |x|e^{|x|}$ for all $x \in \mathbb{R}$,

$$\begin{aligned}
&\int_{A_2} \pi\Big(\hat{\theta}_n + \frac{h}{\sqrt{n}}\Big)e^{-\frac{h^2 I_n}{2}}\left|e^{R_n(h)} - 1\right| \mathrm{d}h \\
&\le \sup_{h \in A_2}\Big(|R_n(h)|\pi\big(\hat{\theta}_n + \frac{h}{\sqrt{n}}\big)\Big) \int_{A_2} e^{-\frac{h^2 I_n}{2} + |R_n(h)|} \mathrm{d}h \\
&\to 0
\end{aligned}$$

in $\mathbb{P}_{\theta_0}$-probability using (6) and (10) and under (B5)-(B7). The proof of Lemma 6.7 is complete.

Appendix 2: A proof of Theorem 6.3 (continued)

Lemma 6.9 *Assume (B1)-(B6) and let* $c, \delta > 0$ *and* $A_3 = \{h \in \mathbb{R} : c\log(\sqrt{n}) \leq |h| \leq \delta\sqrt{n}\}$*. Then, for* δ *small enough and* c *large enough,*

$$\int_{A_3} \left| \pi\big(\hat{\theta}_n + \frac{h}{\sqrt{n}}\big) e^{l_n(\hat{\theta}_n + h/\sqrt{n}) - l_n(\hat{\theta}_n)} - \pi(\hat{\theta}_n) e^{-\frac{h^2 I_n}{2}} \right| \mathrm{d}h \to 0$$

in $\mathbb{P}_{\theta_0}$*-probability and where* I_n *is as in Lemma 6.6.*

Proof: We have,

$$\begin{aligned} &\int_{A_3} \left| \pi\big(\hat{\theta}_n + \frac{h}{\sqrt{n}}\big) e^{l_n(\hat{\theta}_n + h/\sqrt{n}) - l_n(\hat{\theta}_n)} - \pi(\hat{\theta}_n) e^{-\frac{h^2 I_n}{2}} \right| \mathrm{d}h \\ &\leq \int_{A_3} \pi\big(\hat{\theta}_n + \frac{h}{\sqrt{n}}\big) e^{-\frac{h^2 I_n}{2} + R_n(h)} \mathrm{d}h + \pi(\hat{\theta}_n) \int_{A_3} e^{-\frac{h^2 I_n}{2}} \mathrm{d}h \end{aligned} \tag{11}$$

with $R_n(h)$ defined in (9).

For the second integral we have (for n large enough)

$$\begin{aligned} \int_{A_3} e^{-\frac{h^2 I_n}{2}} \mathrm{d}h &\leq 2e^{-\frac{I_n c^2 \log(\sqrt{n})^2}{2}} \big(\delta\sqrt{n} - c\log(\sqrt{n})\big) \\ &\leq 2e^{-\frac{I_n c \log(\sqrt{n})}{2}} \big(\delta\sqrt{n} - c\log(\sqrt{n})\big) \\ &= n^{-\frac{c I_n}{4}} \big(\delta\sqrt{n} - c\log(\sqrt{n})\big) \end{aligned}$$

where, by (6) and under (B4), $I_n \to I_1(\theta_0) > 0$ in $\mathbb{P}_{\theta_0}$-probability. Then, by taking c sufficiently large, $\int_{A_3} e^{-\frac{h^2 I_n}{2}} \mathrm{d}h \to 0$ in $\mathbb{P}_{\theta_0}$-probability so that, under (B5) and (B6),

$$\pi(\hat{\theta}_n) \int_{A_3} e^{-\frac{h^2 I_n}{2}} \mathrm{d}h \to 0 \tag{12}$$

in $\mathbb{P}_{\theta_0}$-probability.

Appendix 2: A proof of Theorem 6.3 (end)

Next, let $\gamma > 0$ and remark that for any $h \in A_3$ we have with $\mathbb{P}_{\theta_0}$-probability tending to one and under (B2) and (B5),

$$\begin{aligned}|R_n(h)| &\leq \frac{1}{6}\Big(\frac{h}{\sqrt{n}}\Big)^3 \sum_{i=1}^n \big|\dddot{l}(X_i, \theta_n')\big| \\ &\leq \delta \frac{h^2}{6}\frac{1}{n}\sum_{i=1}^n \big|\dddot{l}(X_i, \theta_n')\big| \\ &\leq \delta \frac{h^2}{6}\frac{1}{n}\sum_{i=1}^n M(X_i) \\ &\leq \delta \frac{h^2}{6}\Big(\mathbb{E}_{\theta_0}[M(X_1)] + \gamma\Big).\end{aligned}$$

Therefore, as $I_n \to I_1(\theta_0) > 0$ in $\mathbb{P}_{\theta_0}$-probability, for $\delta > 0$ sufficiently small we have

$$|R_n(h)| < \frac{h^2 I_n}{4}, \quad \text{with } \mathbb{P}_{\theta_0}\text{-probability tending to one}$$

implying that

$$-\frac{h^2 I_n}{2} + R_n(h) < -\frac{h^2}{4} I_n, \quad \text{with } \mathbb{P}_{\theta_0}\text{-probability tending to one.}$$

Therefore, with $\mathbb{P}_{\theta_0}$-probability tending to one,

$$\begin{aligned}\int_{A_3} \pi\big(\hat{\theta}_n + \frac{h}{\sqrt{n}}\big) e^{-\frac{h^2 I_n}{2} + R_n(h)} \mathrm{d}h &\leq \sup_{h \in A_3} \pi\big(\hat{\theta}_n + \frac{h}{\sqrt{n}}\big) \int_{A_3} e^{-\frac{h^2 I_n}{4}} \mathrm{d}h \\ &\to 0\end{aligned}$$

as $n \to +\infty$. Together with (11)-(12) this complete the proof of Lemma 6.8.

Appendix 3: Proof of Corollary 6.1

The following result is a direct consequence of Theorem 6.5.

Corollary 6.2 *Assume $\Theta = \mathbb{R}$ and that (B1)-(B6) holds. Then,*

$$\log m(X^{(n)}) = l_n(\hat{\theta}_n) - \frac{1}{2}\log n + \mathcal{O}_{\mathbb{P}_{\theta_0}}(1).$$

Proof: By Lemme 6.6-6.9,

$$\int_{\mathbb{R}} \pi\big(\hat{\theta}_n + \frac{h}{\sqrt{n}}\big) e^{l_n(\hat{\theta}_n + h/\sqrt{n}) - l_n(\hat{\theta}_n)} \mathrm{d}h \to \int_{\mathbb{R}} \pi(\theta_0) e^{-\frac{h^2 I_1(\theta_0)}{2}} \mathrm{d}h \qquad (13)$$

in $\mathbb{P}_{\theta_0}$-probability where

$$\int_{\mathbb{R}} \pi(\theta_0) e^{-\frac{h^2 I_1(\theta_0)}{2}} \mathrm{d}h = \pi(\theta_0)\sqrt{\frac{2\pi}{I_1(\theta_0)}}$$

and where (using the change of variable formula)

$$\int_{\mathbb{R}} \pi\big(\hat{\theta}_n + \frac{h}{\sqrt{n}}\big) e^{l_n(\hat{\theta}_n + h/\sqrt{n}) - l_n(\hat{\theta}_n)} \mathrm{d}h = \sqrt{n} e^{-l_n(\hat{\theta}_n)} m(X^{(n)}). \qquad (14)$$

Therefore, since the mapping $x \mapsto \log(x)$ is continuous, (13) implies that

$$\begin{aligned} \log\Big(\int_{\mathbb{R}} \pi\big(\hat{\theta}_n + \frac{h}{\sqrt{n}}\big) e^{l_n(\hat{\theta}_n + h/\sqrt{n}) - l_n(\hat{\theta}_n)} \mathrm{d}h\Big) & \\ &= -l_n(\hat{\theta}_n) + \frac{1}{2}\log(n) + \log m(X^{(n)}) \\ &\to \log \pi(\theta_0) + \frac{1}{2}\log(2\pi) - \frac{1}{2}\log I_1(\theta_0) \end{aligned}$$

as $n \to +\infty$ and in $\mathbb{P}_{\theta_0}$-probability. The proof is complete.

Remark: In the general case $d \geq 1$, (14) becomes

$$\int_{\mathbb{R}} \pi\big(\hat{\theta}_n + \frac{h}{\sqrt{n}}\big) e^{l_n(\hat{\theta}_n + h/\sqrt{n}) - l_n(\hat{\theta}_n)} \mathrm{d}h = n^{d/2} e^{-l_n(\hat{\theta}_n)} m(X^{(n)})$$

and we recover the expansion given in Corollary 6.1.

VII - Introduction to Markov chain Monte Carlo methods

We start this chapter with a brief history of Bayesian statistics:

The emergence of Bayesian statistics has a long and interesting history dating back to 1763 when Thomas Bayes laid down the basic ideas of his new probability theory (Bayes and Price, 1763, published posthumously by Richard Price). It was rediscovered independently by Laplace (Laplace, 1774) and used in a wide variety of contexts, e.g., celestial mechanics, population statistics, reliability, and jurisprudence. However, after that it was largely ignored. A few scientists, like Bruno de Finetti and Harold Jeffreys, kept the Bayesian theory alive in the first half of the 20th century. Harold Jeffreys published the book Theory of Probability (Jeffreys, 1939), which for a long time remained the main reference for using the Bayes theorem. The Bayes theorem was used in the Second World War at Bletchley Park, United Kingdom, for cracking the German Enigma code, but its use remained classified for many years afterwards. From 1950 onwards, the tide turned towards Bayesian methods. However, the lack of proper tools to do Bayesian inference remained a challenge. The frequentist methods in comparison were simpler to implement which made them more popular. (Sharma, S., 2017. Markov Chain Monte Carlo Methods for Bayesian Data Analysis in Astronomy. arXiv preprint arXiv:1706.01629.)

MCMC methods

Markov chain Monte Carlo (MCMC) algorithms allow to generate a Markov chain having the distribution we want to sample from (called the “target distribution”) as invariant distribution.

The existence of MCMC methods dates back to Metropolis et al. (1953) and Hasting (1970). However, it is only the (relatively) recent increases of the computational power that has made these sampling techniques (and hence Bayesian inference) widely applicable.

Nowadays, MCMC methods in general, and the Metropolis-Hastings algorithm in particular, are the most popular tools used in practice to approximate the posterior distributions arising in Bayesian statistics.

In this chapter we present the Metropolis-Hastings (M-H) algorithm and one of its variant, the Gibbs sampler.

We first present theses algorithms and develop the theory in the case where the target distribution has a finite support. Then, we give their general versions and state (without any proofs) the corresponding main theoretical results.

Notation

Let $\mathcal{Y} = \{1, \ldots, m\}$ for some $m \in \mathbb{N}$. Each $i \in \mathcal{Y}$ is called a state and $\mathcal{Y}$ is called the state space. Let $\mathcal{P}(\mathcal{Y})$ be the set of probability distributions on $\mathcal{Y}$.

We say that a matrix $P = (p_{ij},\ i, j \in \mathcal{Y})$ is stochastic if every row is a distribution on $\mathcal{Y}$; that is

$$\sum_{j=1}^{m} p_{ij} = 1, \quad \min_{j \in \mathcal{Y}} p_{ij} \geq 0, \quad \forall i \in \mathcal{Y}.$$

Definition 7.1 *We say that $(Y_t)_{t\geq 0}$ is a Markov chain with initial distribution $\lambda_0 \in \mathcal{P}(\mathcal{Y})$ and transition matrix P if*

1. *Y_0 has distribution λ_0*
2. *For all $t \geq 0$ and $(i_0, \ldots, i_{t+1}) \in \mathcal{Y}^{t+2}$,*

$$\mathbb{P}(Y_{t+1} = i_{t+1} | Y_t = i_t, \ldots, Y_0 = i_0) = \mathbb{P}(Y_{t+1} = i_{t+1} | Y_t = i_t)$$

3. *For all $t \geq 0$ and $(i, j) \in \mathcal{Y}^2$,*

$$\mathbb{P}(Y_{t+1} = j | Y_t = i) = p_{ij}.$$

We say that $(Y_t)_{t\geq 0}$ is Markov(λ_0, P) in short.

For a Markov(λ_0, P) process $(Y_t)_{t\geq 0}$ we use below the shorthand

$$p_{ij}(t) = \mathbb{P}(Y_t = j | Y_0 = i), \quad t \geq 1, \quad (i, j) \in \mathcal{Y}^2$$

and, for $t \geq 0$, we denote by λ_t the marginal distribution of Y_t; that is

$$\lambda_t := \big(\mathbb{P}(Y_t = 1), \ldots, \mathbb{P}(Y_t = m)\big) \in \mathcal{P}(\mathcal{Y}). \tag{1}$$

Lastly, for $t \geq 0$ we let $p_{ij}^{(t)}$ be the element (i, j) of P^t, so that $P^t = (p_{ij}^{(t)})_{i,j=1}^{m}$.

Some key definitions and properties

The following proposition collects two simple results.

Proposition 7.1 *For any $t \geq 1$, $p_{ij}(t) = p_{ij}^{(t)}$ for all $(i,j) \in \mathcal{Y}^2$ and, for any $t \geq 0$, $\lambda_t^T = \lambda_0^T P^t$.*

Proof: Done in class.

Definition 7.2 *P is irreducible if for any $(i,j) \in \mathcal{Y}^2$ there exists a $t \geq 1$ such that $p_{ij}^{(t)} > 0$.*

In words, P is irreducible if it is possible to go to any state j from any state i.

Definition 7.3 *P is aperiodic if, for all $i \in \mathcal{Y}$, we have $p_{ii}^{(t)} > 0$ for all sufficiently large t.*

In words, P is aperiodic if for any i the event $\{Y_t = i\}$ can happen at irregular times.

Definition 7.4 *A probability measure $\mu \in \mathcal{P}(\mathcal{Y})$ is invariant for P if $\mu^T P = \mu^T$.*

Remark: If μ is invariant for P and $(Y_t)_{t\geq 0}$ is Markov(μ, P) then $Y_t \sim \mu$ for all $t \geq 0$.

An invariant distribution μ of P is often called stationary/equilibrium distribution of P because of the following result.

Theorem 7.1 *Assume that, for some $i \in \mathcal{Y}$, $\lim_{t\to+\infty} p_{ij}^{(t)}$ exists for all $j \in \mathcal{Y}$. Then,*

$$\mu := \big(\lim_{t\to+\infty} p_{ij}^{(t)},\, j \in \mathcal{Y}\big)$$

is an invariant distribution of P.

Proof: Done in class.

Convergence of Markov chains to equilibrium

Definitions 7.3 and 7.4 are important because of the following theorem.

Theorem 7.2 *Let P be an irreducible and aperiodic stochastic matrix with invariant distribution $\mu \in \mathcal{P}(\mathcal{Y})$. Let $\lambda_0 \in \mathcal{P}(\mathcal{Y})$ and $(Y_t)_{t\geq 0}$ be Markov(λ_0, P). Then, there exist constants $\rho \in (0,1)$ and $c \in (0,+\infty)$ such that, for all $(i,j) \in \mathcal{Y}^2$ and $t \geq 0$,*

$$|p_{ij}^{(t)} - \mu_j| \leq c\,\rho^t \quad \textit{and} \quad \big|\mathbb{P}(Y_t = j) - \mu_j\big| \leq c\,\rho^t.$$

Proof: See Appendix 1.

Remark: Theorem 7.2 implies that if P is irreducible and aperiodic then P has at most one invariant distribution. In fact, it can be shown that an irreducible and aperiodic stochastic matrix has a unique invariant distribution (see Problem Sheet 5).

Remark: Theorem 7.2 implies that if $(Y_t)_{t\geq 0}$ is Markov(λ_0, P) for some irreducible and aperiodic stochastic matrix P then, as $t \to +\infty$, $Y_t \overset{\text{dist.}}{\Rightarrow} \mu$ where μ is the unique invariant distribution of P.

Corollary 7.1 *Consider the set-up of Theorem 7.2 and let $\varphi : \mathcal{Y} \to \mathbb{R}$. Then,*

$$\lim_{T\to+\infty} \mathbb{E}\Big[\Big(\frac{1}{T}\sum_{t=1}^{T} \varphi(Y_t) - \sum_{i=1}^{m} \varphi(i)\mu_i\Big)^2\Big] = 0.$$

Proof: See Appendix 2.

Remark: It is also possible to show a strong law of large numbers, namely that

$$\lim_{T\to+\infty} \frac{1}{T}\sum_{t=1}^{T} \varphi(Y_t) \to \sum_{i=1}^{m} \varphi(i)\mu_i, \quad \mathbb{P}\text{-almost surely.}$$

Building a Markov chain having the 'right' invariant distribution

So far we took P as given and assumed that P had an invariant distribution μ.

We now consider the converse problem: Given μ, the distribution we are interested in, how can we construct a transition matrix P such that P has μ as invariant distribution?

Before answering this question we need the following simple lemma.

Lemma 7.1 *Let $\mu \in \mathcal{P}(\mathcal{Y})$ and assume that there exists a transition matrix P such that*

$$\mu_i p_{ij} = \mu_j p_{ji}, \quad \forall (i,j) \in \mathcal{Y}^2. \tag{2}$$

Then, μ is an invariant distribution of P.

Proof: Done in class.

Condition (2) is known as the detailed balance condition. It implies that, at equilibrium (i.e. when $Y_t \sim \mu$), the joint probability $\mathbb{P}(Y_t = i, Y_{t+1} = j)$ is symmetric in t and $t+1$. When there exists a $\mu \in \mathcal{P}(\mathcal{Y})$ such that (2) holds we say that the Markov(λ_0, P) process $(Y_t)_{t\geq 0}$ is reversible.

Lemma 7.1 shows that we can construct a Markov chain having μ as invariant distribution provided that we can construct a transition matrix P such that (2) holds.

Surprisingly, not only it is always feasible to construct such a matrix P, but it is (very) easy using the Metropolis-Hastings algorithm.

The Metropolis-Hastings algorithm

Theorem 7.3 *Let $Q = (q_{ij},\, i,j \in \mathcal{Y})$ be a transition matrix such that $q_{ij} > 0$ for all $(i,j) \in \mathcal{Y}^2$, $\mu \in \mathcal{P}(\mathcal{Y})$ be such that $\mu_i > 0$ for all $i \in \mathcal{Y}$ and $P^{\mathrm{MH}} = (p^{\mathrm{MH}}_{ij},\, i,j \in \mathcal{Y})$ be such that, for every $i \in \mathcal{Y}$,*

$$p^{\mathrm{MH}}_{ij} = \begin{cases} q_{ij}\ \min\left\{1, \frac{\mu_j\, q_{ji}}{\mu_i\, q_{ij}}\right\}, & j \neq i \\ 1 - \sum_{k \neq i} q_{ik}\ \min\left\{1, \frac{\mu_k\, q_{ki}}{\mu_i\, q_{ik}}\right\}, & j = i. \end{cases}$$

Then, P^{MH} is irreducible, aperiodic and has μ as unique invariant distribution.

Proof: Done in class.

We are now ready to write down the famous M-H algorithm.

Metropolis-Hastings algorithm (A1)

```
Input: $\mu \in \mathcal{P}(\mathcal{Y})$, $y_0 \in \mathcal{Y}$ and a transition matrix $Q$ on $\mathcal{Y}$
  Set $Y_0 = y_0$
  for $t \geq 1$ do
      $\tilde{Y}_t \sim Q(Y_{t-1}, \mathrm{d}\tilde{y}_t)$
      Set $Y_t = \tilde{Y}_t$ with probability $\alpha(Y_{t-1}, \tilde{Y}_t)$ and $Y_t = Y_{t-1}$ with probability $1 - \alpha(Y_{t-1}, \tilde{Y}_t)$.
  end for
```

Notation: For $y \in \mathcal{Y}$ the notation $\tilde{Y} \sim Q(y, \mathrm{d}\tilde{y})$ means that the random variable $\tilde{Y}$ is such that $\mathbb{P}(\tilde{Y} = j) = q_{yj}$ while the mapping $\alpha : \mathcal{Y} \times \mathcal{Y} \to [0,1]$ is defined by

$$\alpha(i,j) = \min\left\{1, \frac{\mu_j\, q_{ji}}{\mu_i\, q_{ij}}\right\}, \quad i,j \in \mathcal{Y}.$$

Remark: By Theorem 7.3, the M-H algorithm (A1) defines a Markov$(\delta_{y_0}, P^{\mathrm{MH}}_{\mu})$ process having μ as unique invariant distribution and such that the results of Theorem 7.2 and of Corollary 7.1 hold.

The Metropolis-Hastings algorithm on a general state space

The extension of the M-H algorithm (A1) to an arbitrary state space $\mathcal{Y}$ is straightforward:

Metropolis-Hastings algorithm (A2)

```
Input: $\mu \in \mathcal{P}(\mathcal{Y})$, $y_0 \in \mathcal{Y}$ and a transition kernel $Q$ on $\mathcal{Y}$.
    Set $Y_0 = y_0$
    for $t \geq 1$ do
        $\tilde{Y}_t \sim Q(Y_{t-1}, \mathrm{d}\tilde{y}_t)$
        Set $Y_t = \tilde{Y}_t$ with probability
            $\alpha(Y_{t-1}, \tilde{Y}_t) = \min\left\{1, \frac{\mu(\tilde{Y}_t) q(Y_{t-1}|\tilde{Y}_t)}{\mu(Y_{t-1}) q(\tilde{Y}_t|Y_{t-1})}\right\}$
    and $Y_t = Y_{t-1}$ otherwise.
    end for
```

Remark: We abandon the notion of transition matrix to adopt the more general one of transition kernel.

Notation: $q(\tilde{y}|y)$ is the density of $Q(y, \mathrm{d}\tilde{y})$.

Jargon: $Q(y, \mathrm{d}\tilde{y})$ is often called the proposal distribution.

Technical remark: Now $\mathcal{P}(\mathcal{Y})$ is the set of probability distributions on $\mathcal{Y}$ that are absolutely continuous w.r.t. $\mathrm{d}y$.

Invariant distributions, irreducibility and aperiodicity for general state spaces

We start with the general definition of an invariant distribution.

Definition 7.5 *A probability measure $\mu \in \mathcal{P}(\mathcal{Y})$ is an invariant distribution for the transition kernel P if*

$$\int_{\mathcal{Y}} p(y|y')\mu(y')\mathrm{d}y' = \mu(y).$$

Recall that, when $\mathcal{Y}$ is finite, a transition matrix P is irreducible if it is possible to go to any state $j \in \mathcal{Y}$ from any state $i \in \mathcal{Y}$. If $\mathcal{Y}$ is a continuous state space such a requirement is impossible to full-fill and we therefore need to weaken the notion of irreducibility.

Definition 7.6 *Given a measure φ, the Markov chain $(Y_t)_{t\geq 0}$ with transition kernel P is φ-irreducible if, for every $y \in \mathcal{Y}$ and (measurable) set $A \subset \mathcal{Y}$ with $\varphi(A) > 0$, there exists a $t \geq 0$ such that $P^t(y, A) > 0$.*

In words, P is φ-irreducible if it is possible to go to any (measurable) set $A \subset \mathcal{Y}$ with $\varphi(A) > 0$ from any state $y \in \mathcal{Y}$.

Similarly, the notion of aperiodicity needs to be weakened.

Definition 7.7 *A Markov chain $(Y_t)_{t\geq 0}$ with transition kernel P and stationary distribution μ is aperiodic if there do not exist a $p \geq 2$ and disjoints subsets $\mathcal{Y}_1, \dots, \mathcal{Y}_p \subset \mathcal{Y}$ with $P(y, \mathcal{Y}_{i+1}) = 1$ for all $y \in \mathcal{Y}_i$ ($i \in \{1, \dots, p-1\}$), $P(y, \mathcal{Y}_1) = 1$ for all $y \in \mathcal{Y}_p$, and such that $\mu(\mathcal{Y}_1) > 0$ (and hence $\mu(\mathcal{Y}_i) > 0$ for all i).*

A general convergence result for M-H algorithms

We first show that the M-H algorithm (A2) indeed defines a Markov chain having μ as invariant distribution.

Lemma 7.2 *Let $\mu \in \mathcal{P}(\mathcal{Y})$ and assume that there exists a transition kernel P on $\mathcal{Y}$ such that*

$$\mu(y)p(\tilde{y}|y) = \mu(\tilde{y})p(y|\tilde{y}), \quad \forall (y, \tilde{y}) \in \mathcal{Y}^2. \tag{3}$$

Then, μ is an invariant distribution of P.

Proof: Obvious.

Corollary 7.2 *The Markov chain $(Y_t)_{t\geq 0}$ defined by the M-H algorithm (A2) admits μ as invariant distribution.*

Proof: Done in class.

The following result provides a simple way to check the validity of the M-H algorithm (A2)[a]

Theorem 7.4 *Consider the M-H algorithm (A2). Assume that $Q(y, A) > 0$ for all $y \in \mathcal{Y}$ and all (measurable) set $A \subset \mathcal{Y}$ such that $\mu(A) > 0$, and that*

$$\mathbb{P}\left(\frac{\mu(\tilde{Y}_t)q(Y_{t-1}|\tilde{Y}_t)}{\mu(Y_{t-1})q(\tilde{Y}_t|Y_{t-1})} < 1\right) > 0, \quad \forall t \geq 1.$$

Then, the resulting Markov chain $(Y_t)_{t\geq 0}$ is μ-irreducible and aperiodic, and consequently

$$\lim_{t\to+\infty} \mathbb{P}(Y_t \in A) = \mu(A)$$

for any measurable sets $A \subset \mathcal{Y}$.

[a] See e.g. Theorem 7.4, p.274, of Robert, C.P. and Casella, G. *Monte Carlo Statistical Methods. Springer-Verlag New York (2004).*

Central limit theorem for Markov chains

Let $(Y_t)_{t\geq 0}$ and $(Y_t')_{t\geq 0}$ be two Markov chains defined by the M-H algorithm (A2) where the former is such that $Y_0 = y_0$ for some $y_0 \in \mathcal{Y}$ while the latter is such that $Y_0' \sim \mu$ (the proposal distribution $Q(y, \mathrm{d}\tilde{y})$ being the same for the two processes).

Let $\mu(\varphi) := \int_{\mathcal{Y}} \varphi(y)\mu(y)\mathrm{d}y$ for some (measurable) function $\varphi : \mathcal{Y} \to \mathbb{R}$ verifying $\mu(\varphi^2) < +\infty$ and let

$$\hat{\mu}_T(\varphi) = \frac{1}{T}\sum_{t=1}^{T} \varphi(Y_t)$$

be an estimator of $\mu(\varphi)$.

Then, under some conditions, the following central limit theorem holds

$$\sqrt{T}\,\frac{\hat{\mu}_T(\varphi) - \mu(\varphi)}{\sigma} \overset{\text{dist.}}{\Rightarrow} \mathcal{N}_1(0,1), \quad \text{as } T \to +\infty$$

with

$$\sigma^2 = \mathrm{Var}_\mu(\varphi(Y_0')) + 2\sum_{t=1}^{\infty} \mathrm{Cov}\big(\varphi(Y_0'), \varphi(Y_t')\big) = \mathrm{Var}(\varphi(Y_0'))\tau_\varphi$$

and where $\tau_\varphi = 1 + 2\sum_{t=1}^{\infty} \mathrm{Corr}\big(\varphi(Y_0'), \varphi(Y_t')\big)$ is called the integrated auto-correlation time.

Remark: The asymptotic variance depends only on Q and not on y_0 (which is intuitive).

Choosing the proposal distribution $Q(y, \mathrm{d}\tilde{y})$

The choice of the proposal distribution $Q(y, \mathrm{d}\tilde{y})$ is important because

1. It influences the mixing time of the Markov chain, that is the speed at which it converges to its stationary distribution (the constant ρ in Theorem 7.2).
2. It influences the asymptotic variance of the estimator $\hat{\mu}_T(\varphi)$ of $\mu(\varphi)$ through the integrated auto-correlation time τ_φ.

The first point is particularly important because in practice we can only run algorithm (A2) for a finite number T of iterations. If Q is poorly chosen then the distribution of Y_T will be 'far away' from the equilibrium distribution μ and the output of the algorithm will do a poor job at approximating μ.

Finding a good proposal distribution Q is both difficult and problem dependent: a given Q may perform well for some target distributions and very poorly for others.

In practice, the only solution is often to tune Q manually; that is, to try different proposal distributions until the output of the algorithm (A2) suggests that the algorithm has converged (in the sense that Y_T is approximatively distributed according to μ).

Of course, we can never be sure that the M-H algorithm has converged but there are some ways to detect bad choices for Q, and notably the inspection of

1. the acceptance rate
2. the trace plots
3. the autocorrelation functions.

We explain these three complementary approaches in what follows.

Assessing the convergence using the acceptance rate

The first approach that can be used to assess the convergence of the M-H algorithm is to look at the acceptance rate:

$$r_T = \frac{1}{T}\sum_{t=1}^{T} \mathbf{1}(y_t = \tilde{y}_t).$$

Indeed,

- A low value for this quantity indicates that the simulated trajectory $\{y_t\}_{t=1}^{T}$ remains for a long time at a given location before moving to a new state.
- A high acceptance rate usually (but not always) arises when Q is such that, with high probability, $\tilde{Y}_t$ is very ‘close’ to Y_{t-1}.

Hence, both a low and a large value of the acceptance rate is a sign that the mixing time of the Markov chain (i.e. the time needed to be close to equilibrium) is large and that the correlation between Y_t and Y_{t-k} is important even for large values of k.

There exist theoretical results suggesting that the “optimal” acceptance rate is 0.234. In practice, choosing a proposal distribution Q such that $r_T \approx 0.234$ works well.

The main advantage of this approach is its simplicity. However, by summarizing the behaviour of the Markov chain using a single number, the acceptance rate may hide important differences in the mixing times of the chain along its different coordinates.

Assessing the convergence using the trace plots and the autocorrelation functions

Let $Y_{i,t}$ be the i-th coordinate of Y_t.

Then,

- The trace plot represents the simulated trajectory $\{y_{i,t}\}_{t=1}^T$ as a function of t.
- The auto-correlation function (ACF) returns, for integer $k \geq 0$, an estimate $\hat{\gamma}_T(k)$ of $\text{Corr}(Y'_{i,0}, Y'_{i,k})$ where, as per above, the Markov chain $(Y'_t)_{t\geq 0}$ has the same transition as $(Y_t)_{t\geq 0}$ but is such that $Y'_0 \sim \mu$.

  For instance,

  $$\hat{\gamma}_T(k) = \frac{\frac{1}{T-k}\sum_{s=k+1}^T \big(y_{i,t} - \bar{y}_{i,T}\big)\big(y_{i,t-k} - \bar{y}_{i,T}\big)}{\sqrt{\frac{1}{T}\sum_{s=1}^T \big(y_{i,t} - \bar{y}_{i,T}\big)^2 \frac{1}{T-k}\sum_{s=k+1}^T \big(y_{i,t-k} - \bar{y}_{i,T}\big)^2}}$$

  with

  $$\bar{y}_{i,T} = \frac{1}{T}\sum_{t=1}^T y_{i,t}.$$

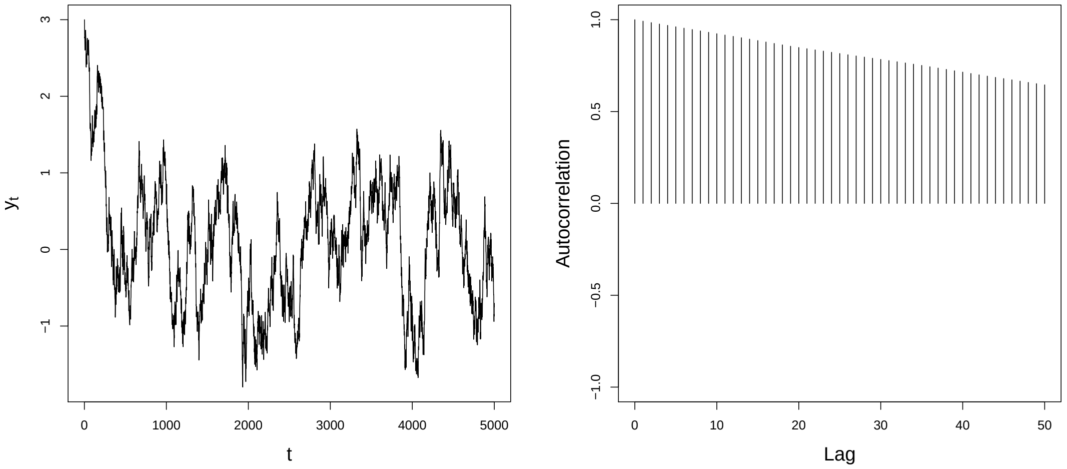
Looking at the trace plots and at the autocorrelation function therefore allows to assess the convergence of the Markov chain coordinate by coordinate.

The M-H algorithm: A simple example

Let $\mathcal{Y} = \mathbb{R}$, $\mu(y)$ be the p.d.f. of the $\mathcal{N}_1(0, 1)$ distribution and $Q_\sigma(y, \mathrm{d}\tilde{y}) = q_\sigma(\tilde{y}|y)\mathrm{d}\tilde{y}$ with $q_\sigma(\tilde{y}|y)$ the p.d.f. of the $\mathcal{N}_1(0, \sigma^2)$ distribution.

Below we use M-H algorithm (A2) based on the transition kernel Q_σ and starting at $y_0 = 3$ to generate a sample that approximates μ.

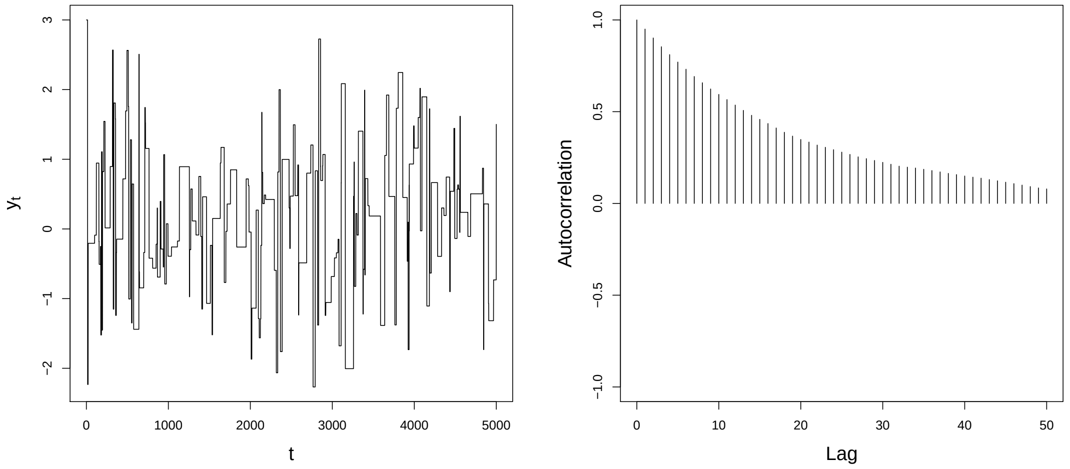
Results for $\sigma = 0.1$:



Trace plot (left) and ACP (right). The acceptance rate is $r_T \approx 0.9736$.

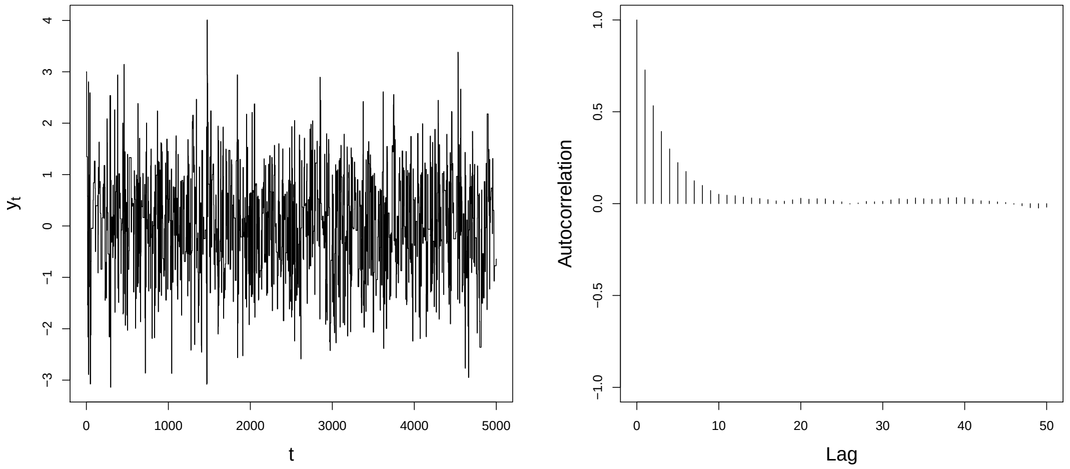
The M-H algorithm: A simple example (end)

Results for $\sigma = 40$:



Trace plot (left) and ACP (right). The acceptance rate is $r_T \approx 0.0364$.

Results for $\sigma = 5$:



Trace plot (left) and ACP (right). The acceptance rate is $r_T \approx 0.2402$.

The Gibbs sampler

The Gibbs sampler, which can be used only when the distribution μ we want to sample from is multivariate, is a particular case of the M-H algorithm (A2) where $Q(Y_{t-1}, \mathrm{d}\tilde{y}_t)$ is such that we have $\mathbb{P}(\alpha(Y_{t-1}, \tilde{Y}_t) = 1) = 1$ for all $t \geq 1$ (see Problem Sheet 5).

Let $\mathcal{Y} = \times_{i=1}^{d} \mathcal{Y}_i$ for some (measurable) sets $\mathcal{Y}_i \subset \mathbb{R}^{d_i}$, $y^{(i)} \in \mathcal{Y}_i$ be the i-th 'block' of $y \in \mathcal{Y}$, so that

$$y = (y^{(1)}, \dots, y^{(d)}).$$

We also define $y^{(k:p)} = (y^{(k)}, \dots, y^{(p)})$ for integers $1 \leq k < p \leq d$ and denote by $y^{(i)}$ the vector y without its i-th block, that is (with obvious conventions when $i \in \{1, d\}$).

$$y^{(i)} = (y^{(1)}, \dots, y^{(i-1)}, y^{(i+1)}, \dots y^{(d)}).$$

For $\mu \in \mathcal{P}(\mathcal{Y})$, $i \in \{1, \dots, d\}$ and $y \in \mathcal{Y}$, we let $\mu^{(i)}(\cdot | y^{(-i)})$ be the p.d.f. on $\mathcal{Y}_i$ defined by

$$\mu^{(i)}(\tilde{y} | y^{(-i)}) = \frac{\mu\big(y^{(1:i-1)}, \tilde{y}, y^{(i+1):d}\big)}{\int_{\mathcal{Y}_i} \mu\big(y^{(1:i-1)}, z, y^{(i+1):d}\big) \mathrm{d}z}, \quad \forall \tilde{y} \in \mathcal{Y}_i$$

In words, $\mu^{(i)}(\cdot | y^{(-i)})$ is the p.d.f. of the distribution of $Y^{(i)}$ under μ, conditional to $Y^{(-i)} = y^{(-i)}$.

The Gibbs sampler on a general state space: The algorithm

Using the above notation the Gibbs sampler on a general state space $\mathcal{Y}$ works as follows.

```
Gibbs sampler (A3)
Input: $\mu \in \mathcal{P}(\mathcal{Y})$, $y_0 \in \mathcal{Y}$
  Set $Y_0 = y_0$
  for $t \geq 1$ do
      for $i = 1, \dots, d$ do
          $Y_t^{(i)} \sim \mu^{(i)}(y_t^{(i)} | Y_t^{(1:i-1)}, Y_{t-1}^{(i+1:d)}) \mathrm{d}y_t^{(i)}$
      end for
  end for
```

The main advantage of the Gibbs sampler is that it does not require to choose a proposal distribution Q; that is, implementing the Gibbs sampler only requires to specify the distribution of interest $\mu \in \mathcal{P}(\mathcal{Y})$ and a starting value $y_0 \in \mathcal{Y}$.

However:

1. The resulting Markov chain may have a large mixing time if the correlation among the different coordinates is important (see the example below).

2. To implement algorithm (A3) we must be able to sample from the full conditional distribution $\mu^{(i)}(\cdot | y^{(-i)})$ for all i, which is rarely the case in practice (see below for a solution to this problem).

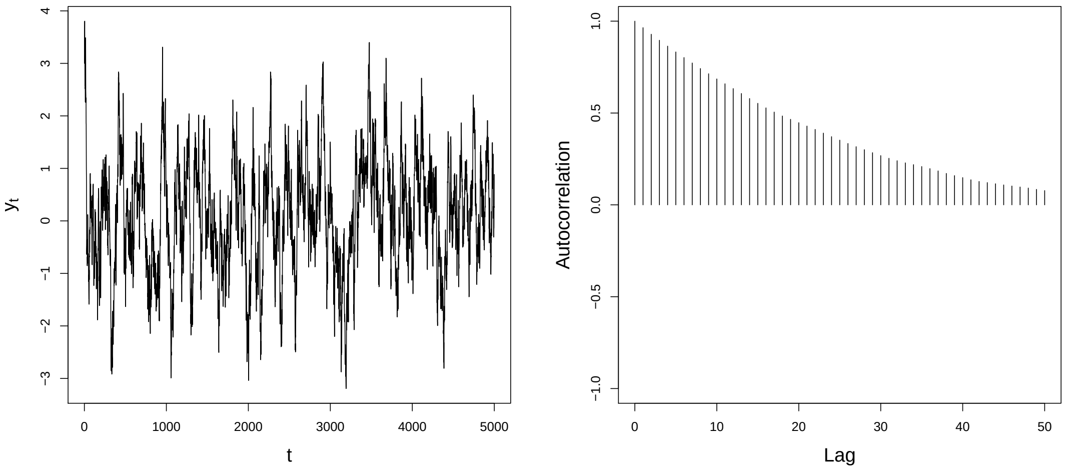
Gibbs sampler: A simple example

Let $\mathcal{Y} = \mathbb{R}^2$ and $\mu(y)$ be p.d.f. of the $\mathcal{N}_2\big(0, \left(\begin{smallmatrix}1 & \rho \\ \rho & 1\end{smallmatrix}\right)\big)$ distribution.

Then, $\mu^{(i)}(\cdot|y^{(-i)})$ is the p.d.f. of the $\mathcal{N}_1(\rho\, y^{(j)}, 1-\rho^2)$ distribution, $i \in \{1, 2\}$.

Below we use the Gibbs sampler (A3) starting at $y_0 = (3, 3)$ to generate a sample that approximates μ for different values of ρ.

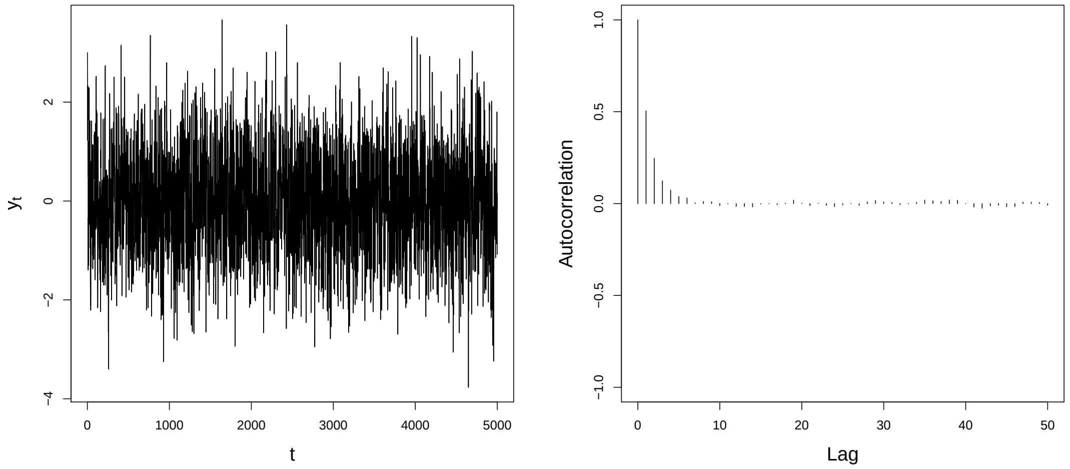
Results for $\rho = 0.98$:



Trace plot (left) and ACP (right) for $(y_t^{(1)})_{t\geq 0}$.

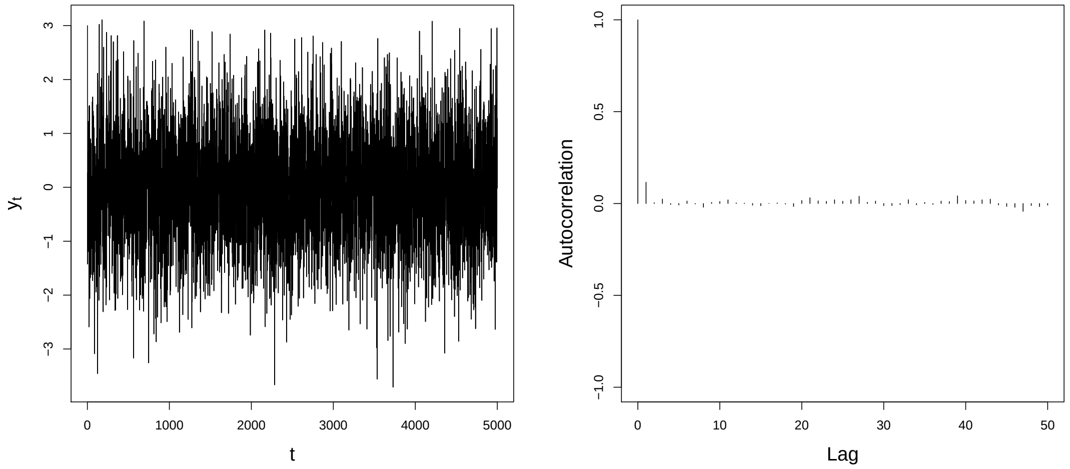
Gibbs sampler: A simple example

Results for $\rho = 0.7$:



Trace plot (left) and ACP (right) for $(y_t^{(1)})_{t\geq 0}$.

Results for $\rho = 0.3$:



Trace plot (left) and ACP (right) for $(y_t^{(1)})_{t\geq 0}$.

“Partial” Gibbs sampling

As mentioned above, the Gibbs sampler (A3) requires to be able to sample from $\mu^{(i)}(\cdot|y^{(-i)})$ for all i, which is rarely the case in practice.

We now assume that we can simulate from $\mu^{(i)}(\cdot|y^{(-i)})$ only for $i = 1, \ldots, d_1$, for some $d_1 < d$.

In this context, the following modified Gibbs sampler can be used.

“Partial” Gibbs sampler (A4)

```
Input: $\mu \in \mathcal{P}(\mathcal{Y})$, $y_0 \in \mathcal{Y}$
    Set $Y_0 = y_0$
    for $t \geq 1$ do
        for $i = 1, \ldots, d_1$ do
            $Y_t^{(i)} \sim \mu^{(i)}(y_t^{(i)}|Y_t^{(1:i-1)}, Y_{t-1}^{(i+1:d)})\mathrm{d}y_t$
        end for
        for $i = d_1 + 1, \ldots, d_2$ do
            $Y_t^{(i)} \sim P_{\mu^{(i)}(y_t^{(i)}|Y_t^{(1:i-1)}, Y_{t-1}^{(i+1:d)})}(Y_{t-1}^{(i)}, \mathrm{d}y_t^{(i)})$
        end for
    end for
```

Notation: We denote by P_η a transition kernel having η as invariant distribution.

Typically, $P_{\mu^{(i)}(\tilde{y}^{(i)}|y^{(1:i-1)}, y_{t-1}^{(i+1:d)})}$ is taken to be a M-H kernel and the resulting algorithm is known as the Metropolis-within-Gibbs algorithm.

Gibbs sampler v.s. Metropolis-within-Gibbs: An example

Let $\mathcal{Y} = \mathbb{R}^2$ and $\mu(y)$ be the p.d.f. of the $\mathcal{N}_2\big(0, \left(\begin{smallmatrix} 1 & \rho \\ \rho & 1 \end{smallmatrix}\right)\big)$ distribution.

Below we use the Gibbs sampler and a Metropolis-within-Gibbs algorithm, both starting at $y_0 \in \mathcal{Y}$, to generate a sample that approximates μ.

The Metropolis-within-Gibbs algorithm we consider is as follows.

Metropolis-within-Gibb for the Bivariate Gaussian example

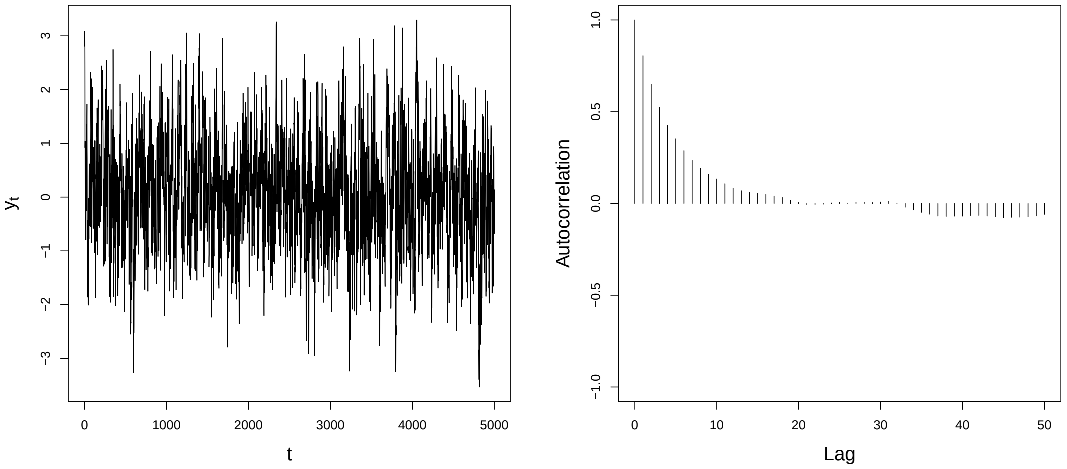
```
**Input:** $y_0 \in \mathcal{Y}$
  Set $Y_0 = y_0$
  **for** $t \geq 1$ **do**
      $Y_t^{(1)} \sim \mathcal{N}_1(\rho Y_{t-1}^{(2)}, 1 - \rho^2)$
      $\tilde{Y}_t^{(2)} \sim \mathcal{N}_1(Y_{t-1}^{(2)}, \sigma^2)$
      Set $Y_t^{(2)} = \tilde{Y}_t^{(2)}$ with probability
          $$\min\left\{1, \frac{\varphi\big(\tilde{Y}_t^{(2)}; \rho Y_t^{(1)}, 1-\rho^2\big)}{\varphi\big(Y_{t-1}^{(2)}; \rho Y_t^{(1)}, 1-\rho^2\big)}\right\}$$
  and $Y_t^{(2)} = Y_{t-1}^{(2)}$ otherwise.
  **end for**
```

Notation: $\varphi(\cdot; \mu, \sigma^2)$ denotes the p.d.f. of the $\mathcal{N}_1(\mu, \sigma^2)$ distribution.

Gibbs sampler v.s. Metropolis-within-Gibbs: An example (end)

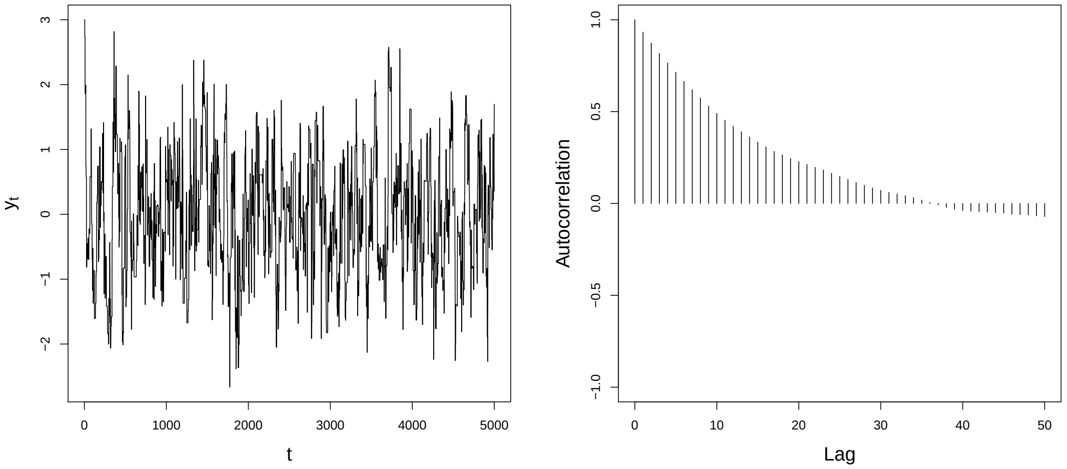
Let $y_0 = (3, 3)$, $\sigma = 2$ and $\rho = 0.9$.

Results for the Gibbs sampler:



Trace plot (left) and ACP (right) for $(y_t^{(2)})_{t \geq 0}$.

Results for the Metropolis-within-Gibbs algorithm:



Trace plot (left) and ACP (right) for $(y_t^{(2)})_{t \geq 1}$. The acceptance rate for $\{y_t^{(2)}\}_{t=1}^T$ is $r_T = 0.2586$.

Appendix 1: Proof of Theorem 7.2

We start with a preliminary result.

Lemma 7.3 *Let P be irreducible and aperiodic. Then, there exists a $t_0 \geq 1$ such that $\min_{(i,j)\in\mathcal{Y}^2} p_{ij}^{(t)} > 0$ for all $t \geq t_0$.*

Proof: Let $(i,j,k) \in \mathcal{Y}^3$ and remark first that, as P is irreducible, there exist a $t_1 \geq 1$ and a $t_2 \geq 1$ such that $p_{ij}^{(t_1)} > 0$ and $p_{ki}^{(t_2)} > 0$. In addition, as P is aperiodic, there exists a $t_3 \geq 1$ such that $p_{ii}^{(t)} > 0$ for all $t \geq t_3$.

Let $(Y_t)_{t\geq 0}$ be Markov(λ_0, P). Then, using the above observations, the Markov property of $(Y_t)_{t\geq 0}$ and Proposition 7.1, for any $t \geq t_3$ we have

$$\begin{aligned}
&p_{kj}^{(t_1+t_2+t)} = \mathbb{P}(Y_{t_1+t_2+t} = j | Y_0 = k)\\
&\geq \mathbb{P}(Y_{t_1+t_2+t} = j, Y_{t_2+t} = i, Y_{t_2} = i | Y_0 = k)\\
&= \mathbb{P}(Y_{t_1+t_2+t} = j | Y_{t_2+t} = i)\mathbb{P}(Y_{t_2+t} = i | Y_{t_2} = i)\mathbb{P}(Y_{t_2} = i | Y_0 = k)\\
&= p_{ij}^{(t_1)}\, p_{ii}^{(t)}\, p_{ki}^{(t_2)}\\
&> 0
\end{aligned}$$

showing that $p_{kj}^{(t)} > 0$ for all $t \geq t_1 + t_2 + t_3$. The result follows.

Appendix 1: Proof of Theorem 7.2 (continued)

We now prove[a] Theorem 7.2 and assume first that $\min_{(i,j)\in\mathcal{Y}^2} p_{ij} > 0$. Let $\mathbf{1} = (1, \dots, 1)$ and $\Pi = \mathbf{1}\mu^T$. Note that

$$P\Pi = \Pi P = \Pi^2 = \Pi. \tag{4}$$

Note also that, because $\min_{(i,j)\in\mathcal{Y}^2} p_{ij} > 0$, there exists an $\alpha \in (0,1)$ such that all the elements of the matrix $P - \alpha\Pi$ are non-negative.

Let

$$\tilde{P} = \frac{1}{1-\alpha}(P - \alpha\Pi)$$

so that, since all the elements of $\tilde{P}$ are non-negative and $\tilde{P}\mathbf{1} = \mathbf{1}$, $\tilde{P}$ is a stochastic matrix. Note also that, by (4),

$$\tilde{P}\Pi = \Pi\tilde{P} = \Pi. \tag{5}$$

Next, let $t \geq 1$. Then,

$$\begin{aligned} P^t &= \Big((1-\alpha)\tilde{P} + \alpha\Pi\Big)^t \\ &= (1-\alpha)^t\tilde{P}^t + \sum_{s=1}^{t}\binom{t}{s}(1-\alpha)^{t-s}\tilde{P}^{t-s}\alpha^s\Pi^s \\ &= (1-\alpha)^t\tilde{P}^t + \Pi\sum_{s=1}^{t}\binom{t}{s}(1-\alpha)^{t-s}\alpha^s \\ &= (1-\alpha)^t\tilde{P}^t + \Pi\big(1 - (1-\alpha)^t\big) \end{aligned} \tag{6}$$

where the second equality uses the Binomial expansion (that holds because the matrices $\tilde{P}$ and Π commute, by (5)), the third equality uses (5) and the last equality uses the fact that the sum is equal to $1 - \mathbb{P}(Z = t)$ where $Z \sim \text{Binomial}(t, 1-\alpha)$.

[a] This proof is due to Prof. Balint Toth

Appendix 1: Proof of Theorem 7.2 (end)

To proceed further for a matrix $A = (a_{ij})$ we let $|A| = (|a_{ij}|)$.

Then, using (6),

$$|P^t - \Pi| = (1-\alpha)^t |\tilde{P}^t - \Pi| \leq (1-\alpha)^t \mathbf{1}\mathbf{1}^T, \quad \forall t \geq 1 \tag{7}$$

where the inequality holds because $\tilde{P}$ and Π are both stochastic matrices.

Since $P^t - \Pi = (p_{ij}^{(t)} - \mu_j)_{i,j=1}^m$ inequality (7) yields

$$|p_{ij}^{(t)} - \mu_j| \leq (1-\alpha)^t, \quad \forall (i,j) \in \mathcal{Y}^2, \quad \forall t \geq 1$$

showing that the conclusion of Theorem 7.2 holds with $c = 1$ and $\rho = (1-\alpha)$ in the special case where $\min_{(i,j)\in\mathcal{Y}^2} p_{ij} > 0$.

Assume now that $\min_{(i,j)\in\mathcal{Y}^2} p_{ij} = 0$. Then, as P is irreducible and aperiodic, there exists by Lemma 7.3 a $t_0 \geq 1$ such that $\min_{(i,j)\in\mathcal{Y}^2} p_{ij}^{(t)} > 0$ for all $t \geq t_0$. Let $P_0 = P^{t_0}$ so that, as per above, there exists an $\alpha_0 \in (0,1)$ such that all the elements of the matrix $P_0 - \alpha_0\Pi$ are non-negative.

Then, repeating the above computations with P replaced by P_0 and α replaced by α_0, we have, noting that P_0^t has elements $(p_{ij}^{(t_0+t)})_{i,j=1}^M$,

$$|p_{ij}^{(t_0+t)} - \mu_j| \leq (1-\alpha_0)^t, \quad \forall (i,j) \in \mathcal{Y}^2, \quad \forall t \geq 1.$$

or, equivalently,

$$|p_{ij}^{(t)} - \mu_j| \leq (1-\alpha_0)^{t-t_0}, \quad \forall (i,j) \in \mathcal{Y}^2, \quad \forall t > t_0.$$

Then, as $|p_{ij}^{(t)} - \mu_j| \leq 1$ for all $(i,j) \in \mathcal{Y}^2$ and all $t \geq 0$, the conclusion of Theorem 7.2 holds with $c = (1-\alpha_0)^{-t_0}$ and $\rho = (1-\alpha_0)$.

Appendix 2: Proof of Corollary 7.1

We first show the following result

Theorem 7.5 *Consider the set-up of Theorem 7.2. Then,*

$$\lim_{T\to+\infty} \mathbb{E}\Big[\Big(\frac{1}{T}\sum_{t=1}^{T}\mathbf{1}_{\{i\}}(Y_t)-\mu_i\Big)^2\Big]=0, \quad \forall i\in\mathcal{Y}.$$

Proof: Fix i. Then,

$$\mathbb{E}\Big[\Big(\frac{1}{T}\sum_{t=1}^{T}\mathbf{1}_{\{i\}}(Y_t)-\mu_i\Big)^2\Big]=\frac{1}{T}\big(v_1(T)+v_2(T)\big)$$

where

$$v_1(T)=\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\big[(\mathbf{1}_{\{i\}}(Y_t)-\mu_i)^2\big]$$

and

$$v_2(T)=\frac{2}{T}\sum_{t=1}^{T-1}\sum_{s=t+1}^{T}\mathbb{E}\big[(\mathbf{1}_{\{i\}}(Y_t)-\mu_i)(\mathbf{1}_{\{i\}}(Y_s)-\mu_i)\big].$$

Then, to prove the result it suffices to show that

$$\limsup_{T\to+\infty} v_1(T)<+\infty, \quad \limsup_{T\to+\infty} |v_2(T)|<+\infty.$$

For $v_1(T)$ we trivially have $v_1(T)\leq 1$ and thus, as required, $\limsup_{T\to+\infty} v_1(T)<+\infty$.

Appendix 2: Proof of Corollary 7.1 (continued)

We now study $v_2(T)$.

To this aim remark first that, by Proposition 7.1 and using the law of iterated expectations, for $1 \leq t < s$,

$$\begin{aligned}
&\mathbb{E}\big[(\mathbf{1}_{\{i\}}(Y_t) - \mu_i)(\mathbf{1}_{\{i\}}(Y_s) - \mu_i)\big] \\
&\qquad\qquad = \mathbb{E}\big[(\mathbf{1}_{\{i\}}(Y_t) - \mu_i)\mathbb{E}[(\mathbf{1}_{\{i\}}(Y_s) - \mu_i)|Y_t]\big] \\
&\qquad\qquad = \mathbb{E}\big[(\mathbf{1}_{\{i\}}(Y_t) - \mu_i)(p_{Y_t i}^{(s-t)} - \mu_i)\big].
\end{aligned}$$

Therefore,

$$\begin{aligned}
|v_2(T)| &\leq \frac{2}{T}\sum_{t=1}^{T-1}\sum_{s=t+1}^{T}\left|\mathbb{E}\Big[(\mathbf{1}_{\{i\}}(Y_t) - \mu_i)(\mathbf{1}_{\{i\}}(Y_s) - \mu_i)\Big]\right| \\
&= \frac{2}{T}\sum_{t=1}^{T-1}\sum_{s=t+1}^{T}\left|\mathbb{E}\Big[(\mathbf{1}_{\{i\}}(Y_t) - \mu_i)(p_{Y_t i}^{(s-t)} - \mu_i)\Big]\right| \\
&\leq \frac{2}{T}\sum_{t=1}^{T-1}\sum_{s=t+1}^{T}\mathbb{E}\Big[|\mathbf{1}_{\{i\}}(Y_t) - \mu_i|\,|p_{Y_t i}^{(s-t)} - \mu_i|\Big] \\
&\leq \frac{2\,c}{T}\sum_{t=1}^{T-1}\sum_{s=t+1}^{T}\rho^{s-t} \\
&\leq \frac{2\,c}{T}\sum_{t=1}^{T-1}\sum_{s=1}^{\infty}\rho^{s} \\
&\leq \frac{2\,c}{1-\rho}
\end{aligned}$$

where the first inequality uses the triangle inequality, the second inequality uses Jensen's inequality and the third inequality the result of Theorem 7.2. The proof of Theorem 7.5 is complete.

Appendix 2: Proof of Corollary 7.1 (end)

We are now ready to prove Corollary 7.1.

Let $\varphi : \mathcal{Y} \to \mathbb{R}$ and $\mu(\varphi) = \sum_{i=1}^m \varphi(i)\mu_i$.

Then ($\mathbb{P}$-a.s.),

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^T \varphi(Y_t) - \sum_{i=1}^m \varphi(i)\mu_i &= \frac{1}{T}\sum_{t=1}^T\sum_{i=1}^m \varphi(i)\big(\mathbf{1}_{\{i\}}(Y_t) - \mu_i\big)\\ &= \sum_{i=1}^m \varphi(i)\frac{1}{T}\sum_{t=1}^T \big(\mathbf{1}_{\{i\}}(Y_t) - \mu_i\big)\end{aligned}$$

so that, using Cauchy-Schwartz's inequality, we have ($\mathbb{P}$-a.s.)

$$\Big(\frac{1}{T}\sum_{t=1}^T \varphi(Y_t) - \sum_{i=1}^m \varphi(i)\mu_i\Big)^2 \le \sum_{i=1}^m \varphi(i)^2 \sum_{i=1}^m \Big(\frac{1}{T}\sum_{t=1}^T \big(\mathbf{1}_{\{i\}}(Y_t) - \mu_i\big)\Big)^2$$

and therefore

$$\mathbb{E}\Big[\Big(\frac{1}{T}\sum_{t=1}^T \varphi(Y_t) - \sum_{i=1}^m \varphi(i)\mu_i\Big)^2\Big] \le \sum_{i=1}^m \varphi(i)^2\mathbb{E}\Big[\Big(\frac{1}{T}\sum_{t=1}^T \mathbf{1}_{\{i\}}(Y_t) - \mu_i\Big)^2\Big]$$

and the result follows from Theorem 7.5.

VIII - Bayesian networks

Bayesian networks belong to the family of probabilistic graphical models.

In a probabilistic graphical model, each node corresponds to a random variable and an edge between two nodes represents the probabilistic dependence between the two corresponding random variables.

Bayesian networks correspond to a type of probabilistic graphical models known as directed acyclic graphs (DAGs), that can be used to provide a compact representation of the conditional independence assumptions for (X, θ).

DAGs are particularly useful

(i) To model complex systems;

(ii) To reduce the complexity of $f(x, \theta) := f(x|\theta)\pi(\theta)$ (and hence to facilitate the approximation of $\pi(\theta|x) \propto f(x, \theta)$) without modifying the conditional independence assumptions for (X, θ).

The main difference between these two points is that in (ii) the statistical model $\{f(\cdot|\theta);\, \theta \in \Theta\}$ is taken as given while in (i) this latter is derived from the DAG.

Informally speaking, the rational behind (ii) is that adding a random variable Ψ may simplify the network (so that $f(x, \theta, \psi)$ is "simpler" than $f(x, \theta)$) without modifying the conditional independence assumptions for (X, θ). This point is explored in more details at the end of the chapter while (i) is discussed in Chapter 9.

Notation and convention

For a set $A \subset \{1, \ldots, m\}$ let $y_A = (y_i,\, i \in A)$ and, for integers $1 \leq i \leq j \leq m$, let $i : j = \{i, \ldots, j\}$.

To simplify the presentation all the random variables that appear below are assumed to be absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^m$ with strictly positive probability density functions.

For a probability density function $p : \mathbb{R}^m \to (0, +\infty)$ and non-empty disjoint sets $A, B \subset 1 : m$, let

$$p_{Y_A}(y_A) = \int_{\mathbb{R}^{m-|A|}} p(y_{1:d}) \mathrm{d}y_{1:d \setminus A}, \quad p_{Y_A|Y_B}(y_A|y_B) = \frac{p_{Y_{A\cup B}}(y_{A\cup B})}{p_{Y_B}(y_B)}.$$

Remark: $p_{Y_A}(y_A|y_B)$ is well-defined as B is non-empty and $p(y) > 0$ for all $y \in \mathbb{R}^m$.

To simplify the notation we will sometimes write $p(y_A)$ instead of $p_{Y_A}(y_A)$, $p(y_A|y_B)$ instead of $p_{Y_A|Y_B}(y_A|y_B)$, etc.

The following simple result will play a key role in what follows.

Theorem 8.1 (Telescopic theorem) *Let* $p : \mathbb{R}^m \to (0, +\infty)$ *be a p.d.f. on* $\mathbb{R}^m$. *Then,*

$$p(y) = p(y_1) \prod_{i=2}^{m} p(y_i|y_{1:(i-1)}), \quad \forall (y_1, \ldots, y_m) \in \mathbb{R}^m.$$

Proof: Left as an exercise.

Directed acyclic graphs (DAGs)

For $i \in 1 : m$ let pa_i be the smallest subset of $0 : (i-1)$ for which

$$p(y) = \prod_{i=1}^{m} p(y_i | y_{\mathrm{pa}_i}) \qquad (1)$$

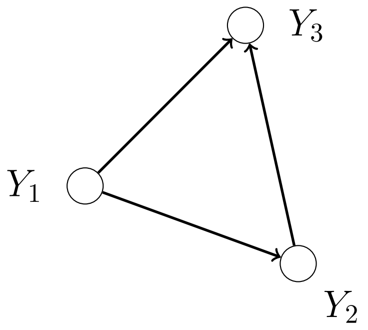
with the convention $p(y_i|y_{\mathrm{pa}_i}) = p(y_i)$ when $\mathrm{pa}_i = \{0\}$.

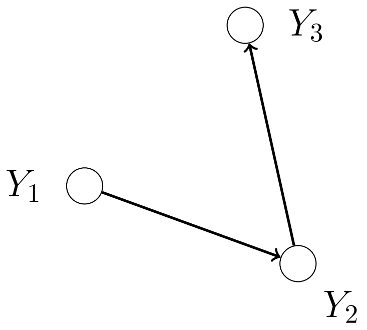
Definition 8.1 pa_i *is the set of parents of* Y_i*.*

A DAG is a graph where each vertex (or node) represents a coordinate of Y and there is an edge from Y_i to Y_j if and only if $i \in \mathrm{pa}_j$.

Remark: A DAG is therefore fully characterized by the pair $(Y, \{\mathrm{pa}_i\}_{i=1}^{m})$.

Example 8.1 *Here are two examples of DAGs (with $m = 3$):*





- *In the left-hand case* $\mathrm{pa}_i = 1 : (i-1)$ *for* $i = 2, 3$ *so that* Y *has p.d.f.*

$$p(y) = p(y_1)p(y_2|y_1)p(y_3|y_1, y_2).$$

- *In the right-hand case,* $\mathrm{pa}_2 = \{1\}$ *and* $\mathrm{pa}_3 = \{2\}$ *so that* Y *has p.d.f.*

$$p(y) = p(y_1)p(y_2|y_1)p(y_3|y_2).$$

DAGs and conditional independence

Definition 8.2 *Let $A, B, C \subset \{1, \dots, m\}$ be non-empty disjoint sets. Then, Y_A is conditionally independent of Y_B given Y_C if*

$$p(y_A, y_B | y_C) = p(y_A | y_C) p(y_B | y_C), \quad \forall y \in \mathbb{R}^m$$

and we write $Y_A \perp\!\!\!\perp Y_B \,|\, Y_C$.

The practical implications of conditional independence are captured in the following result.

Theorem 8.2 *The following statements are equivalent:*

1. $Y_A \perp\!\!\!\perp Y_B \,|\, Y_C$
2. $p(y_A | y_B, y_C) = p(y_A | y_C)$ *for all* $y \in \mathbb{R}^m$.

Proof: Done in class.

For $i \geq 1$ Let $\overline{\mathrm{pa}}_i = \{1, \dots, i-1\} \setminus \mathrm{pa}_i$. Then, Theorem 8.2 implies that (assuming that $\overline{\mathrm{pa}}_i$ is not empty)

$$Y_i \perp\!\!\!\perp Y_{\overline{\mathrm{pa}}_i} \,|\, Y_{\mathrm{pa}_i} \tag{2}$$

and therefore, when building a DAG $(Y, \{\mathrm{pa}_i\}_{i=1}^m)$ for Y, the set $\{\mathrm{pa}_i\}_{i=1}^m$ represents the conditional independence assumptions for Y.

Formal description of a Bayesian network[a]

A Bayesian network B is an annotated directed acyclic graph that represents a joint probability distribution over a set of random variables Y.

The network is defined by a pair $B = (G, P)$, where $G = (Y, \{\mathrm{pa}_i\}_{i=1}^m)$ is the DAG whose nodes $Y_1, \ldots, Y_m$ represent random variables and whose edges (fully characterized by the set $\{\mathrm{pa}_i\}_{i=1}^m$) represent the direct dependencies between these variables.

The graph G encodes independence assumptions, by which each variable Y_i is independent of its nondescendents given its parents in G; that is, $Y_i \perp\!\!\!\perp Y_{\overline{\mathrm{pa}_i}} \,|\, Y_{\mathrm{pa}_i}$.

The second component P denotes the set of parameters of the network. This set contains the "parameter" $p_{Y_i|Y_{\mathrm{pa}_i}}$ for all $i = 1, \ldots m$; that is, $P = \{p_{Y_i|Y_{\mathrm{pa}_i}}\}_{i=1}^m$.

Accordingly, by (1), B defines a unique joint probability distribution for Y, namely:

$$p(y) = \prod_{i=1}^{m} p_{Y_i|Y_{\mathrm{pa}_i}}(y_i|y_{\mathrm{pa}_i}).$$

[a] This definition of Bayesian networks is taken from Ben-Gal I., "Bayesian Networks", in Ruggeri F., Faltin F. & Kenett R., Encyclopedia of Statistics in Quality & Reliability, Wiley & Sons (2007).

DAGs and full conditional distributions

It should be clear that a DAG is constructed with respect to a specific ordering of the components of Y. Hence, for arbitrary $i \in \{1,\dots,m\}$ and non-empty disjoint sets $B, C \subset \{1,\dots,m\} \setminus \{i\}$, a DAG cannot be directly used to figure out if Y_i is conditionally independent of Y_B given Y_C.

To address this kind of questions we need to know $p(y_i|y_{-i})$, the full conditional distribution of Y_i (see Chapter 7).

The next result provides an expression of $p(y_i|y_{-i})$ that depends only on the quantities that need to be specified in a Bayesian network.

Proposition 8.1 *We have*

$$p(y_i|y_{-i}) \propto p_{Y_i|Y_{\mathrm{pa}_i}}(y_i|y_{\mathrm{pa}_i}) \prod_{j\in\{k:\, i\in\mathrm{pa}_k\}} p_{Y_j|Y_{\mathrm{pa}_j}}(y_j|y_{\mathrm{pa}_j}), \quad i=1,\dots,m.$$

Proof: This is a direct consequence of (1).

Remark: This result shows that knowing $\{\mathrm{pa}_i\}_{i=1}^m$ allows to find the smallest set $C_i \subset \{1,\dots,m\}$ such that

$$p(y_i|y_{-i}) \propto \prod_{j\in C_i} p_{Y_j|Y_{\mathrm{pa}_j}}(y_j|y_{\mathrm{pa}_j}), \quad \forall y \in \mathbb{R}^m. \tag{3}$$

Some comments about JAGS

In JAGS the joint distribution we want to sample from is actually specified through a Bayesian network $B=(G,P)$. Proposition 8.1 allows JAGS to efficiently compute $p(y_i|y_{-i})$ as in (3) and, as the elements of P are taken from a given list of probability distributions, to recognize whenever it is possible to sample from $p(y_i|y_{-i})$.

Bayesian networks and prior specification

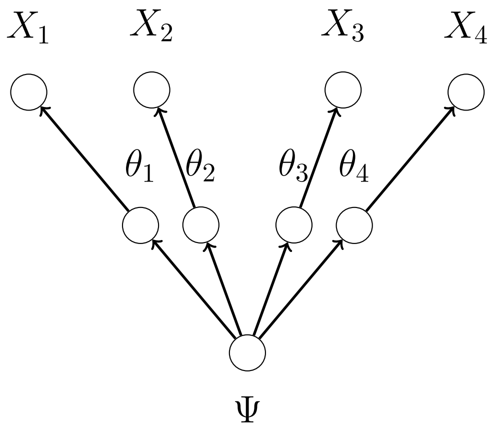
Bayesian networks may be useful to simplify the expression of $\pi(\theta)$ and hence of the joint distribution $f(x, \theta)$.

The idea is to start with the DAG of $Y = (X, \theta)$ and then to add a random variable Ψ to the DAG in order to simplify its structure.

Example 8.2 *Let $k = d = 4$, $m = 2d$ and suppose that*

$$f(x|\theta) = \prod_{i=1}^{d} f(x_i|\theta_i), \quad \pi(\theta) = \prod_{i=1}^{d} \pi_{\theta_i|\theta_{1:(i-1)}}(\theta_i|\theta_{1:(i-1)}) \tag{4}$$

so that there is no conditional independence assumptions on θ. Let Ψ be an additional random variable and consider the following DAG for (X, θ, Ψ)



Then, in the above DAG, the joint distribution of (θ, Ψ) is

$$\tilde{\pi}(\theta, \psi) = \tilde{\pi}_{\Psi}(\psi) \prod_{i=1}^{d} \tilde{\pi}_{\theta_i|\Psi}(\theta_i|\psi). \tag{5}$$

The expression of $\tilde{\pi}(\theta, \psi)$ is simpler than the expression of $\pi(\theta)$ since in (4) $\pi_{\theta_i|\theta_{1:(i-1)}}(\theta_i|\theta_{1:(i-1)})$ is a function of i variables while in (5) $\tilde{\pi}_{\theta_i|\Psi}(\theta_i|\psi)$ depends only on two variables. To simplify further the modelling process it is common to take $\tilde{\pi}_{\theta_i|\Psi} = \tilde{\pi}_{\theta_1|\Psi}$ for all i.

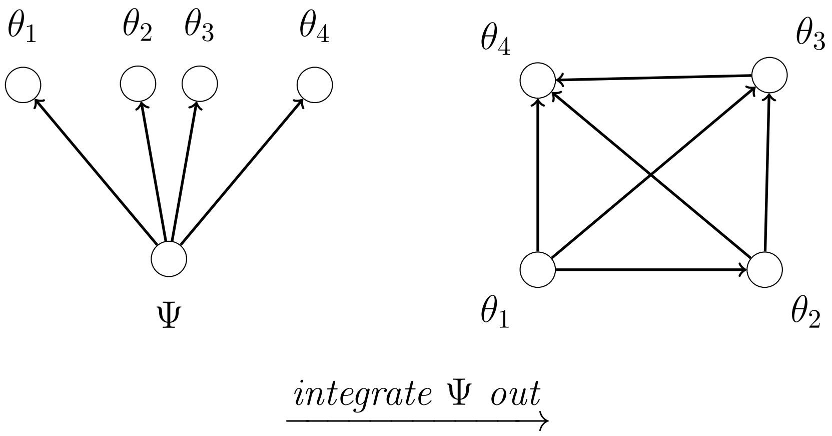
Mutually conditionally independent random variables

Example 8.2 (end) *For this to be useful it remains to show that the joint distribution* (5) *does not modify the conditional independence modelling assumption on* θ. *This is indeed the case since*

$$\tilde{\pi}(\theta_i|\theta_{1:(i-1)}) = \frac{\int \prod_{j=1}^{i} \tilde{\pi}_{\theta_1|\Psi}(\theta_j|\psi)\tilde{\pi}_{\Psi}(\psi)\mathrm{d}\psi}{\int \prod_{j=1}^{i-1} \tilde{\pi}_{\theta_1|\psi}(\theta_j|\psi)\tilde{\pi}_{\Psi}(\psi)\mathrm{d}\psi\mathrm{d}\theta_i} \qquad (6)$$

depends on $\theta_{1:i}$.

Here is what happens to the DAG of (θ, Ψ) *when we integrate* Ψ *out:*



The joint density (5) corresponds to a special form of conditional independence defined in the next definition.

Definition 8.3 *The random variable* Y *is mutually conditionally independent given* Ψ *if* $Y_A \perp\!\!\!\perp Y_B \,|\, \Psi$ *for any disjoint sets* $A, B \subset \{1, \ldots, m\}$. *In this case we write* $\models Y \,|\, \Psi$.

This form of conditional independence is useful in modelling since this is the one that brings the greatest simplification in the expression of $f(x, \theta)$.

DAG of (Y, Ψ) versus DAG of Y

Example 8.2 shows that, for $m = 4$, adding a random variable Ψ such that $\models \theta \,|\, \Psi$ allows to remove two edges in the DAG.

The following proposition generalizes this observation.

Proposition 8.2 *Let* $\mathrm{qa}_i \subset 0 : (i-1)$ *be the set of parents of* Y_i *in the DAG of* (Y, Ψ) *(with the convention* $0 \in \mathrm{qa}_i$ *means that* Ψ *is a parent of* Y_i*) and* $\mathrm{pa}_i \subset 0 : (i-1)$ *be the set of parents of* Y_i *in the DAG of* Y*. Assume that for all* $i = 1, \dots m$*, we have* $\mathrm{qa}_i = \{0\}$ *(so that* $\models Y \,|\, \Psi$*). Then,* $\mathrm{pa}_i = \{1, \dots i-1\}$ *and therefore the DAG of* Y *has* $m(m-1)/2$ *edges while the DAG of* (Y, Ψ) *has* m *edges.*

Proposition 8.2 is a direct consequence of the following theorem that gives the general result for what happens to the DAG of Y when we integrate out Ψ in the DAG of (Y, Ψ).

Theorem 8.3 *Let* $\{\mathrm{qa}_i\}_{i=1}^m$ *and* $\{\mathrm{pa}_i\}_{i=1}^m$ *be as in Proposition 8.2. Then, for* $i = 2, \dots, m$*, we have*

$$\mathrm{pa}_i = \begin{cases} A_{i-1} \cup \left\{ \cup_{j \in A_i} \mathrm{qa}_j \right\} \setminus \{0\}, & 0 \in \mathrm{qa}_i \\ \mathrm{qa}_i, & 0 \notin \mathrm{qa}_i \end{cases}$$

where $A_i = \{j : 1 \leq j \leq i \text{ and } 0 \in \mathrm{qa}_j\}$.

Proof: Done in class.

In words: If Ψ is a parent of Y_i and Ψ is marginalized out then Y_i gets an edge from each Y_j $(i < j)$ for which $0 \in \mathrm{qa}_j$ and also an edge from all the Y_k's that are parents of these Y_j's in the DAG of (Y, Ψ).

Exchangeablility and de Finetti's theorem

In Example 8.2 we describe a way to simplify the expression of the prior distribution $\pi(\theta)$ without imposing any conditional independence assumptions on θ.

However, it is clear that probability density functions of the form

$$\pi(\theta) = \int \prod_{i=1}^{d} \pi(\theta_i|\psi)\pi_\Psi(\psi)\mathrm{d}\psi \tag{7}$$

form only a subset of all the probability density functions on Θ that makes no conditional independence assumptions. Therefore, prior distributions of this form impose conditions on $(\theta_1,\dots,\theta_d)$ that need to be understood.

As shown by de Finetti's theorem (Theorem 8.4), this modelling strategy implicitly assumes that the sequence $(\theta_1,\dots,\theta_d)$ is exchangeable.

Definition 8.4 *A sequence of random variables $Y_1,\dots,Y_m$ is said to be exchangeable if for any permutation σ of $\{1,\dots,m\}$ the joint distribution of $(Y_{\sigma(1)},\dots,Y_{\sigma(m)})$ is the same as the joint distribution of $(Y_1,\dots,Y_m)$.*

Remark: If $Y_1,\dots,Y_m$ are i.i.d. then $(Y_1,\dots,Y_m)$ is exchangeable.

Theorem 8.4 (Hewitt and Savage, 1955) *A sequence of $\mathcal{Y}$-valued random variables $Y_1,\dots,Y_m$ is exchangeable if and only if there exists a unique measure Π on $\mathcal{P}(\mathcal{Y})$ such that, for any measurable sets $B_i\subset\mathcal{Y}$, $i=1,\dots,m$, we have*

$$\mathbb{P}\big(Y_1\in B_1,\dots,Y_m\in B_m\big) = \int_{\mathcal{P}(\mathcal{Y})}\prod_{i=1}^{m} P(B_i)\Pi(\mathrm{d}P). \tag{8}$$

A simpler version of de Finetti's theorem

Theorem 8.4 is the general version of de Finetti's theorem. However, it is not the easiest one to understand as it involves an integral over a space of probability measures.

Although much more restrictive, the next result has the advantage to be easier to understand.

Theorem 8.5 (De Finetti, 1931) *A sequence of* $\{0,1\}$*-valued random variables* $Y_1,\dots,Y_m$ *is exchangeable if and only if there exists a probability measure* $\pi_\Psi\in\mathcal{P}([0,1])$ *such that, for any* $y\in\{0,1\}^m$*, we have*

$$\mathbb{P}\big(Y_1=y_1,\dots,Y_m=y_m\big)=\int_0^1\prod_{i=1}^m\big(\psi^{y_i}(1-\psi)^{1-y_i}\big)\pi_\Psi(\mathrm{d}\psi).$$

In words, if $Y=(Y_1,\dots,Y_m)$ is as in Theorem 8.5, then there exists a random variable Ψ on $[0,1]$ such that $\models Y\mid\Psi$. Moreover, $Y_i|\Psi\sim\text{Bernoulli}(\Psi)$.

Consequently, prior distributions of the form (7) assume that $(\theta_1,\dots,\theta_d)$ is exchangeable.

IX - Hierarchical models

We start with a definition.

Definition 9.1 *A hierarchical Bayes model is a Bayesian model where the prior distribution* $\pi(\theta)$ *is decomposed in conditional distributions*

$$\pi_{\theta|\psi_1}(\theta|\psi_1),\dots,\pi_{\psi_{L-1}|\psi_L}(\psi_{L-1}|\psi_L)$$

and a marginal distribution $\pi_{\psi_L}(\psi_L)$ *such that*

$$\pi(\theta)=\int \pi_{\theta|\psi_1}(\theta|\psi_1)\Big(\prod_{l=1}^{L-1}\pi_{\psi_l|\psi_{l+1}}(\psi_l|\psi_{l+1})\Big)\pi_{\psi_L}(\psi_L)\mathrm{d}(\psi_1,\dots,\psi_L).$$

The parameter ψ_l *is called hyperparameter of level* l*.*

As we will see in this short chapter, hierarchical models are particularly relevant to model observations that are “similar” but not identical, because they allow to easily control the interrelationship between the observations according to known group.

Having similar but not identical observations typically arises when we are interested in making inference on parameters corresponding to different sub-populations (or “units”) of a global population.

Remark: As mentioned in Chapter 3, hierarchical models can also be used to reduce the impact of the prior distribution on the inference.

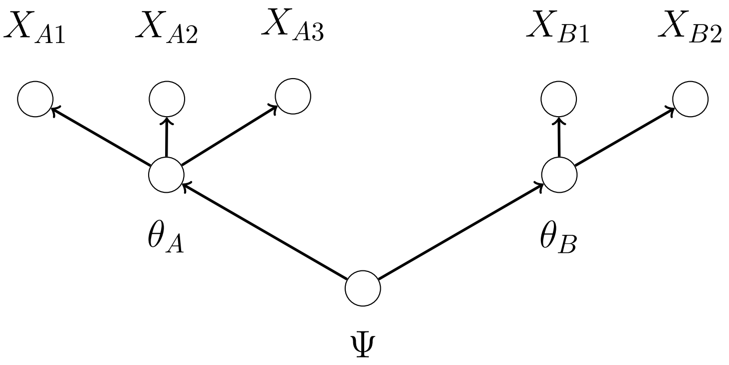
Modelling similar observations: An example

Consider a school with classes A and B. Class $j \in \{A, B\}$ contains n_j students, with $n_A = 3$ and $n_B = 2$. Let X_{ji} be the exam score of student $i \in \{1, \ldots, n_j\}$ in class $j \in \{A, B\}$. All the students are assumed to be "similar" (i.e. same age, similar socio-economic characteristics, etc.).

Assume that we have observed (X_{A1}, X_{A2}, X_{B1}) and that we want to predict X_{B2}. Then, it seems sensible to take into account that

- The distribution of the exam scores in the two classes should be "similar" since the two classes are in the same school. Hence, (X_{A1}, X_{A2}) is relevant to predict X_{B2}.
- However, it is unlikely that the distribution of the exam scores in the two classes is identical since e.g. the teacher is not the same in the two classes. Hence, the prediction of X_{B2} should be more influenced by X_{B1} than (X_{A1}, X_{A2}) .

A possible DAG in this context is therefore:



Remark: In this DAG, the distribution of X_{A1} and of X_{B1} are indeed not identical, since the former depends on θ_A while the latter depends on $\theta_B \neq \theta_A$, but are similar in the sense that they share the same hyperparameter Ψ.

Modelling similar observations: An example (continued)

Let $\theta = (\theta_A, \theta_B)$ and $X = (X_A, X_B)$ where, for $j = A, B$, $X_j = (X_{ji}, \dots, X_{jn_j})$.

Then, from Chapter 8, we know that the above DAG implies that $\models \theta | \Psi$ while $\models X_j | \theta_j$ for $j = A, B$.

Therefore, the above DAG implies the following decomposition for the joint distribution of (X, θ, Ψ):

$$f(x, \theta, \psi) = \pi_{\Psi}(\psi) \prod_{j \in \{A,B\}} \pi_{\theta_j|\Psi}(\theta_j|\psi) \prod_{i=1}^{n_j} f_{X_{ji}|\theta_j}(x_{ji}|\theta_j).$$

Typically, we choose $\pi_{\theta_B|\Psi} = \pi_{\theta_A|\Psi}$ and $f_{X_{ji}|\theta_j} = f_{X_{A1}|\theta_A}$ for all i, j so that, to complete the Bayesian network (i.e. the Bayesian model) we just need to specify

$$\pi_{\Psi}, \quad \pi_{\theta_A|\Psi}, \quad f_{X_{A1}|\theta_A}.$$

Remark: With these three distributions we can handle an arbitrary large number of groups (i.e. classes) and an arbitrary large number of cases per group (i.e. students per class).

Remark: Assuming $\pi_{\theta_B|\Psi} = \pi_{\theta_A|\Psi}$ amounts to assuming that (θ_A, θ_B) is exchangeable (see Theorem 8.4). Therefore, hierarchical models are particularly useful when the units' specific parameters are exchangeable, since in this case the model is built from only a small number of probability density functions.

Remark: In the notation of Definition 9.1, $L = 1$ and $\psi_1 = \psi$.

Modelling similar observations: An example (end)

Remark that if the prior distribution for θ is $\pi(\theta) = \pi_A(\theta_A)\pi_B(\theta_B)$ then the resulting posterior distribution is given by

$$\begin{aligned}\pi(\theta|x) &\propto \left(\pi_A(\theta_A)\prod_{i=1}^{n_A} f_{X_{Ai}|\theta_A}(x_{Ai}|\theta_A)\right)\left(\pi(\theta_B)\prod_{i=1}^{n_B} f_{X_{Bi}|\theta_B}(x_{Bi}|\theta_B)\right)\\ &\propto \pi_A(\theta_A|x_A)\pi_B(\theta_B|x_B)\end{aligned}$$

where

$$\pi_j(\theta_j|x_j) \propto \pi_j(\theta_j)\prod_{i=1}^{n_j} f_{X_{ji}|\theta_j}(x_{ji}|\theta_j), \quad j \in \{A, B\}.$$

Therefore, when choosing $\pi(\theta) = \pi_A(\theta_A)\pi_B(\theta_B)$ the parameters θ_A and θ_B are estimated separately, the former using only x_A and the latter only x_B. Consequently, the resulting posterior distribution for θ_j will be more dispersed (i.e. the estimate of θ_j will be less "precise") than the one obtained with the hierarchical approach used above.

Remark: Using a hierarchical approach however makes sense only if we are ready to assume that X_A and X_B are similar.

Some final comments about hierarchical models

In practice it is quite frequent to encounter situations where it makes sense to assume that $f(x|\theta) = \prod_{i=1}^{d} \tilde{f}(x_i|\theta_i)$ so that we have as many parameters as observations.

In this scenario,

1. Considering a prior distribution of the form $\pi(\theta) = \prod_{i=1}^{d} \pi_i(\theta_i)$ would result in a posterior distribution for θ_i that depends on x_i only (that is, our inference on θ_i would be based on a single observation).
2. Considering a model of the form

$$f(x, \theta, \psi) = \pi_\Psi(\psi) \prod_{i=1}^{d} \tilde{f}(x_i|\theta_i)\pi_{\theta_i|\Psi}(\theta_i|\psi)$$

   would result in a posterior distribution for θ_i that depends on all the observations $(x_1, \dots, x_d)$ and therefore that contains more information about θ_i than in case 1.

An alternative modelling strategy in this context is to specify a model $\{f'(\cdot|, \lambda),\ \lambda \in \Lambda\}$ where λ is a low dimensional parameter (i.e. the dimension of λ is smaller than that of x).

- In a frequentist setting, specifying such a model is far from trivial as this requires to come up with a p.d.f. $f'(x|\lambda)$ that captures the dependence among the different components of x.
- In a Bayesian setting, the hierarchical approach allows to easily specify $f'(x|\lambda)$ as

$$f'(x|\lambda) = \int \prod_{i=1}^{d} \tilde{f}(x_i|\theta_i)\pi_{\theta_i|\Psi}(\theta_i|\lambda)\mathrm{d}(\theta_1, \dots, \theta_d).$$

  Remark: This approach is not allowed in a pure frequentist setting as it requires to treat the parameters as random variables.