Introduction to Bayesian learning
Lecture 1: Bayesian learning: basics

Anne Sabourin, As. Prof., Telecom ParisTech

September 2019

Course mechanism

- 6 sessions of 3.5 hours each
- 2 lab session (sessions 3 and 6)
- Grading : 40% homework (2^{nd} lab report), 60% written exam.
- Course Software : `R`.

Syllabus

1. Bayesian learning : basics
2. Bayesian methods for unsupervised and supervised problems, Bayesian decision theory
3. Lab session I : `R` tutorial, Naive Bayes, Bayesian regression
4. Variational methods
5. Sampling methods : Monte-Carlo Markov Chain
6. Lab session II : variational and sampling methods

Lecture 1, Basics of Bayesian learning : Outline

1. When is a Bayesian approach needed ?
2. The Bayesian framework
3. Construction of estimators
   - Point estimation
   - Interval estimation
4. Prior choices : conjugate priors, exponential family and alternatives
   - Exponential family
   - conjugate priors in exponential families
   - Prior choice
5. A glimpse at Bayesian asymptotics
   - Example : Beta-Binomial model
   - Posterior consistency
   - Asymptotic normality
6. Exercises

1. When is a Bayesian approach needed ?

2. The Bayesian framework

3. Construction of estimators
   - Point estimation
   - Interval estimation

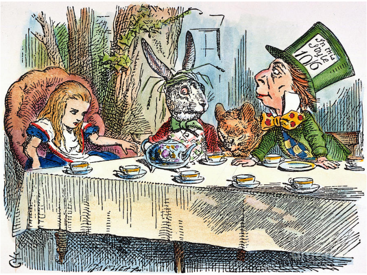
4. Prior choices : conjugate priors, exponential family and alternatives
   - Exponential family
   - conjugate priors in exponential families
   - Prior choice

5. A glimpse at Bayesian asymptotics
   - Example : Beta-Binomial model
   - Posterior consistency
   - Asymptotic normality

6. Exercises

The English lady, the music lover and the drunkard

The English Lady claims that she can tell whether the milk was poured before or after the tea, after one sip.



Ten trials are made. At each trial the milk is randomly poured before or after the tea. The lady's gess is true 9 times over 10.

What is your verdict : can she really tell ?

The English lady, the music lover and the drunkard

A music lover claims that he can tell if a piece is from Haydn or Mozart after listening only ten seconds of it.

Ten trials are made. At each trial a music piece is randomly chosen from Haydn or Mozart. The music lover's guess is true 9 times over 10.

What is your verdict : can he really tell ?

The English lady, the music lover and the drunkard

Your drunken friend claims that he can predict the outcome of a flip of a fair coin.

Ten trials are made. At each trial a coin is flipped. The drunkard's guess is true 9 times over 10.

What is your verdict : can he really tell ?

Issues

- The 3 datasets are the same and the task is similar, however would you give the same answer in the 3 situations ?
- What level of confidence would you have concerning your answer ? Are the asymptotic confidence intervals from the Central Limit Theorem reliable ?

Bayesian statistics provide a formalism to

- Include prior beliefs in the analysis of data.
- Quantify the uncertainty by providing ‘credible intervals’ ($\neq$ classical confidence intervals)

1. When is a Bayesian approach needed ?

2. The Bayesian framework

3. Construction of estimators
   - Point estimation
   - Interval estimation

4. Prior choices : conjugate priors, exponential family and alternatives
   - Exponential family
   - conjugate priors in exponential families
   - Prior choice

5. A glimpse at Bayesian asymptotics
   - Example : Beta-Binomial model
   - Posterior consistency
   - Asymptotic normality

6. Exercises

Probabilistic modeling

- **Dataset** $X = X_{1:n} = (X_1, \dots, X_n)$, $X_i \in \mathcal{X} = \{0, 1\}$
  $X_i = 1$ if right guess, $1 \leq i \leq n$.
  $\mathcal{X}$ is the **sample space**, n is the **sample size**.

- **Statistical model :**
  $X_i \sim \mathrm{P}_\theta = \mathcal{B}er(\theta)$ (Bernoulli distribution) : $\mathrm{P}_\theta\{1\} = \theta$.
  $\theta \in \Theta = [0, 1]$ unknown **parameter.**
  Θ is the **parameter space**.

- *i.i.d.*(**i**ndependent, **i**dentically **d**istributed) data :
  $X = (X_1, \dots, X_n) \sim \mathrm{P}_\theta^{\otimes n}$: product distribution.
- Underlying probability space $(\Omega, \mathcal{A}, \mathbb{P}_\theta)$. $X_i : \Omega \to \mathcal{X}$,
  $\mathrm{P}_\theta = \mathbb{P}_\theta \circ X_i^{-1}$.
  Here $\mathrm{P}_\theta\{1\} = \mathbb{P}_\theta \circ X_i^{-1}(\{1\}) = \mathbb{P}_\theta\{X_i = 1\}$. $\mathrm{P}_\theta = \mathbb{P}_\theta \circ X_i^{-1}$.

Statistical model

Definition : statistical model

A family $\mathcal{P} = \{\mathrm{P}_\theta, \theta \in \Theta\}$ of probability distributions indexed by $\theta \in \Theta$, over a sample space $\mathcal{X}$. Θ is the parameter space.

- 'parametric model' : when $\Theta \subset \mathbb{R}^d$.
- 'non parametric model' : when Θ is infinite-dimensional (example : mixture model with infinitely many components)

Goal

Learn about θ_0 using a dataset $X = X_{1:n} = (X_1, \dots, X_n)$, assuming that $X_i \sim P_{\theta_0}$, $1 \le i \le n$ for some $\theta_0 \in \Theta$.

- $X_i \in \mathbb{R}^p$: unsupervised learning,
  versus
- $X_i = (z_i, Y_i)$: supervised learning (Y_i : label)

What is a Bayesian model ?

- 'Prior knowledge' about θ represented by a probability distribution $\boldsymbol{\pi}$: the **prior distribution**.

- One can define a random variable $\boldsymbol{\theta}$, with $\boldsymbol{\theta} \sim \boldsymbol{\pi}$.

- The X_i's are independent conditionnally to $\boldsymbol{\theta}$.

- We assume that *a single* θ_0 which is a realisation of $\boldsymbol{\theta}$ produces the data, *i.e.* $X_{1:n}$ is distributed according to $\mathrm{P}_{\theta_0}^{\otimes n}$, for some $\theta_0 \in \Theta$.

Definition : Bayesian model

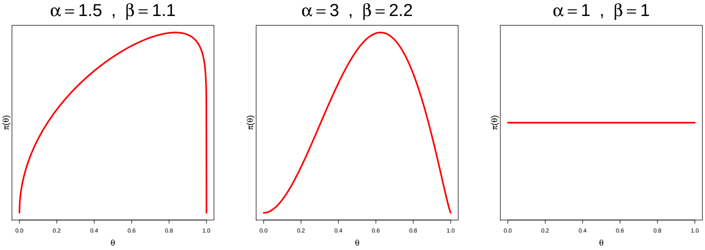
A statistical model $\mathcal{P} = \{\mathrm{P}_\theta, \theta \in \Theta\}$ together with a prior distribution $\boldsymbol{\pi}$ on Θ.

Example : the English lady

- $\theta \in [0,1]$: probability of a right guess.
- Prior knowledge : The true θ_0 is ‘probably’ close to 0.5, maybe higher.
- Prior distribution : a Beta distribution $\mathcal{B}eta(\alpha,\beta)$ on $(0,1)$,

$$\pi(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

- $\mathbb{E}(\boldsymbol{\theta}) = \frac{\alpha}{\alpha+\beta}$, $\mathbb{V}\mathrm{ar}(\boldsymbol{\theta}) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$



3 examples of Beta density

Doing Bayesian inference = conditioning upon the data

- The bayesian model results in a **Joint distribution** over the product space $\Theta \times \mathcal{X}$:

$$Q(A \times B) = \int_{\theta \in A} P_\theta(B) \, \mathrm{d}\boldsymbol{\pi}(\theta) \, , \qquad A \subset \Theta, \, B \subset \mathcal{X}.$$

  P_θ is viewed as a **conditional distribution** of X_i given θ.

- Learning = conditioning prior knowledge about $\boldsymbol{\theta}$ upon data X.

Definition : posterior distribution

conditional distribution of $\boldsymbol{\theta}$ given X

- All the inference (estimation, prediction, ...) is derived from the posterior distribution.

i.i.d. samples : notational conventions

When $X = X_{1:n} = (X_1, \dots, X_n)$, $X_i \overset{\text{i.i.d}}{\sim} P_\theta$, $1 \le i \le n$.

- Then $X : \Omega \to \mathcal{X}^n$ and $X \sim P_\theta^{\otimes n}$ (product measure)
- Joint distribution over $\Theta \times \mathcal{X}^n$,

$$Q(A \times B) = \int_{\theta \in A} P_\theta^{\otimes n}(B) \, \mathrm{d}\boldsymbol{\pi}(\theta), \quad B \subset \mathcal{X}^n$$

- If P_θ has a density $p_\theta(x)$, $x \in \mathcal{X}$, then $P_\theta^{\otimes n}$ has density $p_\theta^{\otimes n}(x) = \prod_{i=1}^n p_\theta(x_i)$, $x = (x_1, \dots, x_n) \in \mathcal{X}^n$.
- For convenience we omit the '${}^{\otimes n}$' sign.

Computing the posterior distribution : Assumptions

- $\boldsymbol{\pi}$ has density π *w.r.t.* reference measure μ (σ-finite), $\frac{\mathrm{d}\boldsymbol{\pi}}{\mathrm{d}\mu} = \pi$.

- Dominated model : $\exists$ reference measure λ on $\mathcal{X}$ such that each P_θ has density p_θ *w.r.t.* λ : $\frac{\mathrm{d}P_\theta}{\mathrm{d}\lambda} = p_\theta$.

- For a given x, $\theta \mapsto p_\theta(x)$ is the **likelihood function**

- Notation : $p(x|\theta) := p_\theta(x)$.

- x : realisation of X : a single r.v. or an *i.i.d.* sample $(X_1, \dots, X_n)$

Computing the posterior distribution : Bayes theorem

Under the previous assumptions :

Bayes Theorem

The posterior distribution has a density *w.r.t.* μ given by

$$\pi(\theta|x) = \frac{p(x|\theta)\pi(\theta)}{\int_\Theta p(x|t)\pi(t)\,\mathrm{d}\mu(t)}$$

$$\text{posterior} \;\propto\; \text{likelihood} \;\times\; \text{prior}$$

For any $x \in \mathcal{X}$ such that the denominator is > 0.

- Denominator : $m(x) = \int_\Theta p(x|t)\pi(t)\,\mathrm{d}\mu(t)$, marginal density of X
- remind $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$ when $P(B) \neq 0$.

Example : The English lady

- Assumptions are met with
  - λ : counting measure on $\mathcal{X} = \{0, 1\}$, $\lambda\{0\} = \lambda\{1\} = 1$.
  - $p_\theta(x) = \theta^x(1-\theta)^{1-x}$
  - μ : Lebesgue measure on $(0, 1)$
  - π : Beta density $\mathcal{B}eta(\alpha, \beta)$
- Computing the posterior density :

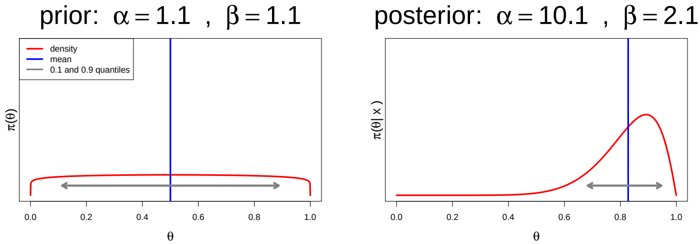
$$\pi(\theta|x) = \frac{p_\theta(x)\pi(\theta)}{\underbrace{m(x)}_{\text{does not depend on } \theta}}$$

$$\propto p_\theta(x)\pi(\theta) \quad (\propto\text{: proportional to})$$

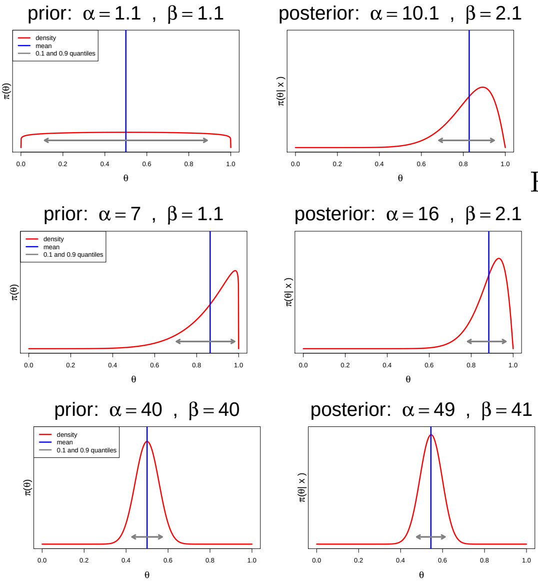
$$= \theta^{\alpha+\sum_{i=1}^n x_i - 1}(1-\theta)^{\beta+n-\sum_{i=1}^n x_i - 1}$$

$$\propto \text{density of } \mathcal{B}eta(\alpha + \sum x_i, \beta + n - \sum x_i)$$

Posterior density for the English lady



Influence of the prior



English lady

music lover

drunkard

Sequential nature of Bayesian learning

Posterior after n *i.i.d.* obs $x_{1:n}$ starting from prior π

$=$

Posterior after the latest obs x_n starting from prior $\pi(\theta|x_{1:n-1})$

Proof

$$\begin{aligned}\pi(\theta|x_{1:n}) &= \frac{\pi(\theta)p(x_{1:n-1}|\theta)p(x_n|\theta)}{\int \pi(t)p(x_{1:n-1}|t)\,\mathrm{d}\mu(t)} \times \frac{\int \pi(t)p(x_{1:n-1}|t)\,\mathrm{d}\mu(t)}{\int \pi(t)p(x_{1:n}|t)\,\mathrm{d}\mu(t)} \\ &= \frac{\pi(\theta|x_{1:n-1})p(x_n|\theta)}{\tilde{m}(x_{1:n})}\end{aligned}$$

with

$$\begin{aligned}\tilde{m}(x_{1:n}) &= \frac{\int \pi(t)p(x_{1:n-1}|t)p(x_n|t)\,\mathrm{d}\mu(t)}{\underbrace{\int \pi(t|x_{1:n-1})\,\mathrm{d}\mu(t)}_{=1} m(x_{1:n-1})} \\ &= \int \pi(t|x_{1:n-1})p(x_n|t)\,\mathrm{d}t\end{aligned}$$

1. When is a Bayesian approach needed ?

2. The Bayesian framework

3. Construction of estimators
   - Point estimation
   - Interval estimation

4. Prior choices : conjugate priors, exponential family and alternatives
   - Exponential family
   - conjugate priors in exponential families
   - Prior choice

5. A glimpse at Bayesian asymptotics
   - Example : Beta-Binomial model
   - Posterior consistency
   - Asymptotic normality

6. Exercises

From posterior probability to estimation

- Raw output of Bayesian analysis : a posterior distribution (represented as a density or as a sample $(\theta_1, \ldots, \theta_N) \sim \pi(\theta | X_{1:n})$, where N is fixed by the user

- In practice : one wants to answer questions of the kind
  - Does $\theta \in \Theta_0 \subset \Theta$?
  - What is your best guess $\widehat{\theta}(X)$ for θ, given data X ? (point estimation)
  - Can you give a region $R \subset \Theta)$ such that $\mathbb{P}(\theta \in R | X_{1:n}) \geq 1 - \alpha$?

1. When is a Bayesian approach needed ?

2. The Bayesian framework

3. Construction of estimators

   Point estimation

   Interval estimation

4. Prior choices : conjugate priors, exponential family and alternatives

5. A glimpse at Bayesian asymptotics

6. Exercises

Bayesian point estimation

most popular estimators of θ : posterior mode and posterior mean.

- Posterior mode $\widetilde{\theta} = \mathrm{argmax}_t \, \pi(t|X_{1:n})$
- Posterior mean $\theta^* = \mathbb{E}_\pi(\theta|X_{1:n}) = \int_\Theta \theta \, \pi(\mathrm{d}\theta|X_{1:n})$.
- generalisation of posterior mean for a quantity of interest $g(\theta)$:

$$g^* = \mathbb{E}_\pi(g(\theta)|X_{1:n}) = \int_\Theta g(\theta) \, \pi(\mathrm{d}\theta|X_{1:n}).$$

- Warning : posterior mode depends on the reference measure

Remark : all three estimators are 'statistics' : functions of $X_{1:n}$.

Discussion : posterior mode

- Intuition : $\widetilde{\theta}$ is the 'center' of the region $\delta\theta$ of measure $\mu(\delta\theta)$ for which the posterior mass $\approx \pi(\widetilde{\theta})\mu(\delta\theta)$ is the highest.

- Warning (main criticism) : $\widetilde{\theta}$ depends on the reference measure.

Discussion : posterior mean

Main justification (for $g : \Theta \to \mathbb{R}$)

$$g^* := \mathbb{E}_\pi(g(\boldsymbol{\theta})|X_{1:n}) = \operatorname*{argmin}_{\gamma\in\mathbb{R}} \underbrace{\mathbb{E}_\pi\left[\left(g(\theta)-\gamma\right)^2|X_{1:n}\right]}_{\varphi(\gamma)}.$$

g^* minimizes he posterior expectancy of the quadratic risk. Indeed,

$$\begin{aligned}\varphi(\gamma) &= \int_\Theta \left(g(\theta)^2 - 2\gamma g(\theta) + \gamma^2\right)\pi(\mathrm{d}\theta|X_{1:n})\\ &= Cste \ - 2\gamma \underbrace{\int g(\theta)\pi(\mathrm{d}\theta|X_{1:n})}_{\mathbb{E}_\pi(g(\boldsymbol{\theta})|X_{1:n})} + \gamma^2\end{aligned}$$

$$\varphi'(\gamma) = 2(-\mathbb{E}_\pi(g(\boldsymbol{\theta})|X_{1:n}) + \gamma)$$

$$\varphi'(\gamma) = 0 \iff \gamma = \mathbb{E}_\pi(g(\boldsymbol{\theta})|X_{1:n})$$

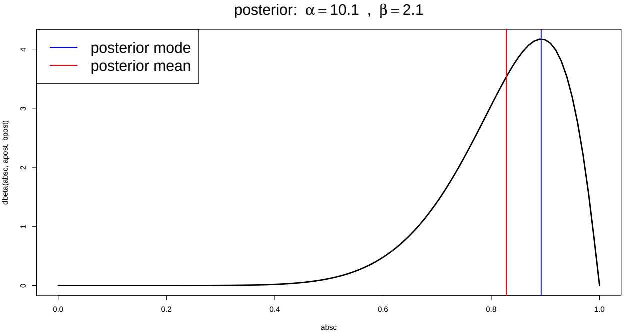
This is indeed a minimum since $\varphi''(\gamma) \equiv 2 > 0$.

Example : Tea Lady

- prior over $]0,1[$: $\boldsymbol{\pi} = \mathcal{B}\mathit{eta}(\alpha = 1.1, \beta = 1.1)$.
- posterior distribution : $\mathcal{B}\mathit{eta}(\alpha' = 10.1, \beta' = 2.1)$.

$$\text{mode} : \tilde{\theta} = \frac{\alpha' - 1}{\alpha' + \beta' - 2} = \frac{9.1}{10.2}$$

$$\text{mean} : \theta^* = \frac{\alpha'}{\alpha' + \beta'} = \frac{10.1}{12.2}$$



1. When is a Bayesian approach needed ?

2. The Bayesian framework

3. Construction of estimators

Point estimation

Interval estimation

4. Prior choices : conjugate priors, exponential family and alternatives

5. A glimpse at Bayesian asymptotics

6. Exercises

Interval / Set estimation

In what reasonable interval/region of Θ do you believe θ to belong ?

Goal : find a region $R \subset \Theta$ with

- High posterior mass
- Moderate 'size' (*w.r.t.* the reference measure)

definition : poterior credible set

Given the data x, A posterior credible set of level α for a quantity of interes $g(\theta)$ is any (measurable) region $R \subset g(\Theta)$ such that $\mathbb{P}_\pi(g(\theta) \in R|x) \geq \alpha$.

⚠ credible sets $\neq$ confidence regions $R_{classic}$ for an estimator (classical setting), such that $\mathbb{P}_\theta(R_{classic} \ni g(\theta)) \geq \alpha$, $\forall\theta$.

- in fact : confidence and credible sets 'approximately' coïncide for large sample sizes (due to Bernstein-Von-Mises theorem, see last section).

Posterior quantiles

- similarly to confidence interval, there is no unique way to define credible sets.

- easy way (for $g(\Theta) \subset \mathbb{R}$) : use posterior quantiles

- remind : if Q is a probability on $\mathbb{R}$, an α-quantile relative to Q is any $q_\alpha \in \mathbb{R}$ s.t. $Q[-\infty, q_\alpha] = \alpha$.

- When $(1-\alpha)/2$ and $(1+\alpha)/2$ quantiles for $\boldsymbol{\pi}(\,\cdot\,|x)$ exist, a credible interval of level alpha is $(q_{\frac{1-\alpha}{2}}, q_{\frac{1+\alpha}{2}}]$.

Minimum volume sets

- It is the solution to the initial requirements (large posterior mass, small reference measure)
- define $R(u) = \{\theta \ : \ \pi(\theta|x) \geq u\}$ ('interior of a density level set)
- The minimum volume set of level α is R_{u_α}, where

$$u_\alpha = \sup\{u \geq 0 : \boldsymbol{\pi}(R(u)|x) \geq \alpha\}$$

- in practice (in general) : hard to compute. Need Monte-Carlo methods, computationally intensive in high dimension.

1. When is a Bayesian approach needed ?

2. The Bayesian framework

3. Construction of estimators
   - Point estimation
   - Interval estimation

4. Prior choices : conjugate priors, exponential family and alternatives
   - Exponential family
   - conjugate priors in exponential families
   - Prior choice

5. A glimpse at Bayesian asymptotics
   - Example : Beta-Binomial model
   - Posterior consistency
   - Asymptotic normality

6. Exercises

Conjugate priors

- In general, computing $\pi(\theta|x)$ is hard.
- Not when the prior is 'conjugate', that is when π and $\pi(\,\cdot\,|x)$ belong to a parametric family with an explicit expression for the posterior.

definition : conjugate prior

A parametric family of priors $\mathcal{F} = \{\pi_\gamma, \gamma \in \Gamma\}$ with $\Gamma \subset \mathbb{R}^d$ is *conjugate* for the model $\{p_\theta, \theta \in \Theta\}$ is for all x, for all $\pi = \pi_\gamma \in \mathcal{F}$, it holds that $\pi(\theta|x) \in \mathcal{F}$, *i.e.* $\exists \gamma'$ such that $\pi(\,\cdot\,|x) = \pi_{\gamma'}$.

- γ parameterizes the prior : it is called the *hyper-parameter*
- Justification for choosing a conjugate prior : computational convenience only.

Example I : Gaussian model with known variance

Setting : $\Theta = \mathbb{R}$, $p_\theta(x) \propto e^{-\sum_i (x_i-\theta)^2/(2\sigma^2)}$.

- Thus $p_\theta(x) \propto \exp\{\text{quadratic function of } \theta\}$.
- If $\pi(\theta) \propto \exp\{\text{quadratic function of } \theta\}$, then also : $\pi(\theta|x) \propto \pi(\theta) p_\theta(x) \propto \exp\{\text{quadratic function of } \theta\}$
- The only densities of the kind $f(\theta) \propto \exp\{\text{quadratic function of } \theta\}$ are Gaussian
- **Conclusion** The prior family $\mathcal{F} = \{\mathcal{N}(\mu, s^2), \mu \in \mathbb{R}, s^2 > 0\}$ is conjugate for the Gaussian model

Example I cont'd

If $\pi(\theta) = \mathcal{N}(x|\mu, s^2)$ and $p_\theta(x) = \mathcal{N}(x|\theta, \sigma^2)$, then

$$\begin{aligned}
\pi(\theta|x_{1:n}) &\propto \exp\{-1/2 \sum_i \frac{(x_i-\theta)^2}{2\sigma^2} + \frac{(\theta-\mu)^2}{s^2}\} \\
&\propto \exp\left\{\theta^2(n/\sigma^2 + 1/s^2) - 2\theta(\sum x_i/\sigma^2 + \mu/s^2) + C\right\} \\
&\propto \mathcal{N}(\theta|\mu_n, s_n^2)
\end{aligned}$$

with

$$\begin{cases} \mu_n & = (s^2+\sigma^2/n)^{-1}(s^2 \frac{1}{n}\sum x_i + (\sigma^2/n)\mu) \\ 1/s_n^2 & = 1/s^2 + n/\sigma^2 \end{cases}$$

N.B. μ_n is the posterior expectectancy, it may be taken as an estimate for θ. It is a weighted average between the maximum likelihood estimator and the prior mean (μ).

Conjugate priors : exercises

1. Gaussian model $\mathcal{N}(\mu, \sigma^2)$)with known mean and unknown variance : Find a conjuate prior for the parameter $\lambda = 1/\sigma^2$.

2. Gaussian model $\mathcal{N}(\mu, \sigma^2)$ with unknown mean and variance : same question for the parameter $\theta = (\mu, \lambda = 1/\sigma^2)$.
   (Hint : write $\pi(\mu, \lambda) = \pi(\mu|\lambda)\pi(\lambda)$ and use the fact that the likelihood writes as $p_{\mu,\lambda}(x_{1:n}) = f_{n,x}(\mu)g_{n,x,\mu}(\lambda)$.

conjugate priors for multivariate normal

- $X \sim \mathcal{N}(\mu, \Lambda^{-1})$, $\mu \in \mathbb{R}^d, \Lambda \in \mathbb{R}^{d \times d}$ positive, definite (precision matrix : inverse of covariance matrix)

1. unknown mean → conjugate prior family on μ : a multivariate Gaussian distributions

2. unknown precision → conjugate prior on Λ : Wishart distributions $\mathcal{W}(\nu, W)$ with ν degrees of freedom ($\nu \in \mathbb{N}^*$) and $W \in \mathbb{R}^{d \times d}$.

Wishart distribution

defined on the cone of positive definite matrices.

- The Wishart distribution $\mathcal{W}(\nu, W)$ has density

$$f_{\mathcal{W}}(\Lambda|\nu, W) = B \det \lambda^{(\nu-d-1)/2} \exp\left\{\frac{-1}{2}\mathrm{Tr}(W^{-1}\Lambda)\right\}$$

  w.r.t. Lebesgue on $\mathbb{R}^{\frac{d(d+1)}{2}}$: $\prod_{i\leq j} \mathrm{d}\Lambda_{(i,j)}$, restricted to the set of positive definite matrices.

- B : a normalizing constant.
- probabilistic representation : let M be a random $\nu \times d$ matrix with *i.i.d.*rows $M_{(i,\cdot)} \sim \mathcal{N}(0, W)$. Then

$$\Lambda \sim \mathcal{W}(\nu, W) \iff \Lambda \overset{\mathrm{d}}{=} M^\top M = \sum_{i=1}^{n} M_{(i,\cdot)}^\top M_{(i,\cdot)}$$

- More details : see *e.g.* *Eaton, Multivariate Statistics : A Vector Space Approach, 2007 (Chapter 8)*

conjugate priors for multivariate normal, Cont'd

3. Unknown mean and precision → conjugate prior family on (μ, Λ) : the Gaussian-Wishart distribution with hyper-parameters (W, ν, m, β)

$$\pi(\mu, \Lambda) = \pi_1(\Lambda)\pi_2(\mu|\lambda)$$

with

$$\boldsymbol{\pi}_2 = \mathcal{W}(W, \nu), \nu \in \mathbb{N}, W \text{ positive definite},$$
$$\boldsymbol{\pi}_2(\,\cdot\,|\Lambda) = \mathcal{N}(m, (\beta\Lambda)^{-1}), \qquad m \in \mathbb{R}^d, \beta > 0$$

1. When is a Bayesian approach needed ?

2. The Bayesian framework

3. Construction of estimators

4. Prior choices : conjugate priors, exponential family and alternatives

   Exponential family

   conjugate priors in exponential families

   Prior choice

5. A glimpse at Bayesian asymptotics

6. Exercises

Definition : exponential family

A dominated parametric model $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$ is an *exponential family* if the densities write

$$p_\theta(x) = C(\theta) h(x) \exp\left\{ \langle T(x), R(\theta) \rangle \right\}$$

for some functions

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

- $C(\theta)$: a normalizing constant
- $R(\theta)$: the *natural parameter* (R : the 'good' re-parametrization)
- If $R(\theta) = \theta$, the family is *natural.*

- Most textbook distributions are from the exponential family !

Example I : Bernoulli model

- $\theta \in \Theta =]0,1[$, $\mathcal{X} = \{0,1\}$
- The model is dominated by $\lambda = \delta_0 + \delta_1$

$$\begin{aligned} p_\theta(x) &= \theta^x (1-\theta)^{1-x} \\ &= \exp\{x \log\theta + (1-x)\log(1-\theta)\} \\ &= (1-\theta)\exp\left\{ \underbrace{x}_{T(x)} \underbrace{\log\frac{\theta}{1-\theta}}_{R(\theta)} \right\} \end{aligned}$$

- The model is an exponential family with
  - $T(x) = x$
  - natural parameter : $\rho = R(\theta) = \log\frac{\theta}{1-\theta}$.
  - normalizing constant $C(\theta) = (1-\theta)$

Example II : Gaussian model

- $\theta = (\mu, \sigma^2) \in \Theta = \mathbb{R} \times \mathbb{R}_+^*$
- the model is dominated by Lebesgue on $\mathcal{X} = \mathbb{R}$.

$$\begin{aligned} p_\theta(x) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{\frac{-(x^2 - 2\mu x + \mu^2)}{2\sigma^2}\right\} \\ &= \underbrace{\frac{1}{\sqrt{2\pi\sigma^2}} \exp\{\frac{-\mu}{2\sigma^2}\}}_{C(\theta)} \exp\left\langle \underbrace{\begin{pmatrix} x \\ x^2 \end{pmatrix}}_{T(x)}, \underbrace{\begin{pmatrix} \mu/\sigma^2 \\ -1/(2\sigma^2) \end{pmatrix}}_{R(\theta)} \right\rangle \end{aligned}$$

- The model is an exponential family with
  - $T(x) = (x, x^2)$
  - natural parameter : $\rho = R(\theta) = (\mu/\sigma^2, -1/2\sigma^2)$.
  - normalizing constant $C(\theta) = (2\pi\sigma^2)^{-1/2}$

likelihood for *i.i.d.* samples in exponential families

$$p_\theta(x_1) = C(\theta)h(x)\exp\left\{ \langle R(\theta), T(x_1)\rangle \right\}$$

$$\Rightarrow$$

$$p_\theta^{\otimes n}(x_{1:n}) = C(\theta)^n \underbrace{\prod_{i=1}^n h(x_i)}_{h_n(x_{1:n})} \exp\left\{\left\langle \underbrace{\sum_{i=1}^n T(x_i)}_{T_n(x_{1:n})}, R(\theta)\right\rangle\right\}.$$

Natural parameter space

- natural parametrization : $\rho = R(\theta)$.
- The density $p_\rho(x) = C(\rho)h(x)\exp\left\langle T(x)\,,\,\rho\right\rangle$ integrates to 1 $\iff \rho \in \mathcal{E}$, the *natural parameter space*, *i.e.*

$$\mathcal{E} = \left\{\rho : \int_{\mathcal{X}} h(x)\exp\left\langle T(x)\,,\,\rho\right\rangle \mathrm{d}\lambda(x) < \infty\right\}$$

- If $\mathcal{E}$ is open : the family is *regular*.

Maximum likelihood in regular exponential families

natural parametrization : $\rho = R(\theta)$. λ : reference measure.

lemma : expression for $\mathbb{E}_\rho\big[T(X)\big]$

$$\mathbb{E}_\rho\big[T(X)\big] = -\nabla_\rho\{\ln C(\rho)\}$$

Proof

$$1 \equiv C(\rho)\int_{\mathcal{X}} h(x)\exp\left\langle T(x)\,,\,\rho\right\rangle \mathrm{d}\lambda(x)$$

(with regularity to exchange $\displaystyle\int$ and ∇) $\Rightarrow$)

$$0 = \nabla_\rho C(\rho)\underbrace{\int_{\mathcal{X}} h(x)\exp\left\langle T(x)\,,\,\rho\right\rangle \mathrm{d}\lambda(x)}_{C(\rho)^{-1}} + C(\rho)\int_{\mathcal{X}} h(x)T(x)\exp\left\langle T(x)\,,\,\rho\right\rangle \mathrm{d}\lambda(x)$$

$$\Rightarrow 0 = \frac{1}{C(\rho)}\nabla_\rho C(\rho) + \mathbb{E}\big(T(X)\big)$$

□

Maximum likelihood in regular exponential families, cont'd

proposition

The MLE estimator $\widehat{\rho}$ in a regular exponential family satisfies

$$\mathbb{E}_{\widehat{\rho}}[T(X)] = \frac{1}{n} \sum_i T(x_i).$$

Proof

$$\begin{aligned} \nabla_\rho \log p_{\widehat{\rho}}(x) = 0 &\iff \nabla_\rho \{ n \log C(\rho) + \langle \sum T(x_i), \rho \rangle \} = 0 \\ &\iff \nabla_\rho \log C(\widehat{\rho}) = \frac{-1}{n} \sum_i T(x_i). \end{aligned}$$

then use the lemma. □

1. When is a Bayesian approach needed ?

2. The Bayesian framework

3. Construction of estimators

4. Prior choices : conjugate priors, exponential family and alternatives
   - Exponential family
   - conjugate priors in exponential families
   - Prior choice

5. A glimpse at Bayesian asymptotics

6. Exercises

Conjugate priors in exponential family

Proposition

A natural exponential family with densities $p_\theta(x) = C(\theta)h(x)\exp\langle\theta, T(x)\rangle$, admits a conjugate prior family $\mathcal{F} = \{\boldsymbol{\pi}_{\lambda,\mu}, \lambda > 0, \mu \in M_\lambda \subset \mathbb{R}^k\}$, with

$$\pi_{\lambda,\mu}(\theta) = K(\mu, \lambda)C(\theta)^\lambda \exp\left\{\langle\theta, \mu\rangle\right\}$$

and $M_\lambda = \{\mu : \int_\Theta \pi(\mu, \lambda)\,\mathrm{d}\theta < \infty\}$.
The posterior for n observation is

$$\pi_{\lambda,\mu}(\theta|x_{1:n}) \propto C(\theta)^{\lambda+n} \exp\left\{\langle\theta, \mu + \sum_i T(x_i)\rangle\right\}$$

so that $\boldsymbol{\pi}_{\lambda,\mu}(\,\cdot\,|x_{1:n}) = \boldsymbol{\pi}_{\lambda_n,\mu_n}(\,\cdot\,)$, with

$$\lambda_n = \lambda + n\,; \qquad \mu_n = \mu + \sum_i T(x_i)$$

proof exercise

Example : Poisson model

$$p_\theta(x) = e^{-\theta}\theta^x/x!, \qquad \mathcal{X} = \mathbb{N}, \theta > 0$$
$$= \frac{1}{x!} e^{-\theta} e^{x \log \theta}$$

→ an exponential family with

$$T(x) = x\,, \qquad \rho = R(\theta) = \log\theta \in \mathbb{R}\,, \qquad C(\rho) = \exp\{-e^\rho\}$$

conjugate prior for ρ :

$$\pi_{a,b}(\rho) \propto \exp\{-be^\rho\}\ \exp\{a\rho\}.$$

Back to θ :

$$\pi(\theta) = \text{"}\frac{\mathrm{d}\boldsymbol{\pi}}{\mathrm{d}\rho}\frac{\mathrm{d}\rho}{\mathrm{d}\theta}\text{"} = \theta^{a-1}\exp\{-b\theta\} \quad \text{(Gamma density)}$$

→ The Gamma family is a conjugate prior for θ.

1. When is a Bayesian approach needed ?

2. The Bayesian framework

3. Construction of estimators

4. Prior choices : conjugate priors, exponential family and alternatives
   - Exponential family
   - conjugate priors in exponential families
   - Prior choice

5. A glimpse at Bayesian asymptotics

6. Exercises

About the choice of a conjugate prior

- A convenient choice only
- One must still choose hyper-parameters (λ, μ)
- This is an issue of *model choice*
- possible to do so via *empirical Bayes* methods, see lecture 2 and lab session.

Other prior choices : non informative priors

- Goal : minimize the bias induced by the prior

- If Θ compact : one can choose $\pi(\theta) = \text{Constant}$
- If Θ non compact, $\int_\theta \pi(\theta)\,\mathrm{d}\theta = \int_\Theta C\,\mathrm{d}\theta = +\infty$
  OK to do so as long as the posterior is well defined, *i.e.* when

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

⚠ uniform only *w.r.t.* the reference measure → not invariant under re-parametrization.
e.g. Flat prior on $]0, 1[$ in a $\mathcal{B}er(\theta)$ model → non flat over $\rho = \log[\theta/(1-\theta)]$

Other prior choices : Jeffreys prior

- For Θ open in $\mathbb{R}^d$. Reasonable with $d = 1$.

- Remind the Fisher information (in a regular model) :

$$I(\theta) = \mathbb{E}_\theta\Big[\Big(\frac{\partial \log p_\theta(X)}{\partial\theta}\Big)^2\Big] = -\mathbb{E}\Big[\frac{\partial^2 \log p_\theta(X)}{\partial\theta^2}\Big].$$

- $I(\theta)$ is the expected curvature of the likelihood around θ.
- Interpretation as a an average information carried by X about θ.
- Idea : grant more prior mass to highly informative θ's

Definition : Jeffreys prior

In a dominated model with densities $p_\theta, \theta \in \Theta$, the Jeffreys prior has densities *w.r.t.* Lebesgue on Θ :

$$\pi(\theta) \propto \sqrt{I(\theta)}.$$

- **exercise** compute the Jeffreys prior in the Bernoulli model, in the location model $\mathcal{N}(\theta, \sigma^2)$, σ^2 known and in the scale model $\mathcal{N}(\mu, \theta^2)$, μ known.

Invariance of the Jeffreys prior

- Change of variable : $h(\theta) = \eta$. Then $p_\theta = p_{h(\theta)}$.
- Let $\boldsymbol{\theta} \sim \boldsymbol{\pi}_{J,\theta}$ the Jeffreys prior. Then $\boldsymbol{\eta} \sim \boldsymbol{\pi}_{J,\theta} \circ h^{-1}$ with density

$$\pi(\eta) \overset{\text{for } \theta = h^{-1}(\eta)}{=} \pi_{J,\theta}(\theta)\frac{\mathrm{d}\theta}{\mathrm{d}\eta} \quad = \quad \frac{\sqrt{I(\theta)}}{h'(\theta)}$$

- On the other hand compute the Jeffreys prior on η :

$$\begin{aligned} \pi_{J,\eta}(\eta) = \sqrt{I_\eta(\eta)} &= \mathbb{E}_\eta\Big[\Big(\frac{\partial \log p_\eta(X)}{\partial \eta}\Big)^2\Big]^{1/2} \\ &\overset{\theta = h^{-1}(\eta)}{=} \mathbb{E}_\theta\Big[\Big(\frac{\partial \log p_\theta(X)}{\partial \theta}\frac{\mathrm{d}\theta}{\partial \eta}\Big)^2\Big]^{1/2} \quad = \quad \frac{\sqrt{I(\theta)}}{h'(\theta)}. \end{aligned}$$

- Same result : the Jeffreys prior in the η parametrization is the image measure of the Jeffreys prior in the θ parametrization.
- In other words the Jeffreys prior is parametrization-invariant.

1. When is a Bayesian approach needed ?

2. The Bayesian framework

3. Construction of estimators
   - Point estimation
   - Interval estimation

4. Prior choices : conjugate priors, exponential family and alternatives
   - Exponential family
   - conjugate priors in exponential families
   - Prior choice

5. A glimpse at Bayesian asymptotics
   - Example : Beta-Binomial model
   - Posterior consistency
   - Asymptotic normality

6. Exercises

Rough overview

as the sample size $n \to \infty$

- The influence of the prior choice vanishes
- The posterior distribution concentrates around the true value θ_0 (almost surely)
- The posterior distribution is asymptotically normal with mean $\widehat{\theta} =$ the maximum likelihood, and variance $n^{-1} I(\theta)^{-1}$ (same as MLE's)

1. When is a Bayesian approach needed ?

2. The Bayesian framework

3. Construction of estimators

4. Prior choices : conjugate priors, exponential family and alternatives

5. A glimpse at Bayesian asymptotics

Example : Beta-Binomial model

Posterior consistency

Asymptotic normality

6. Exercises

Reminder : Beta-Binomial model

- Bayesian model $\begin{cases} \boldsymbol{\theta} \sim \boldsymbol{\pi} = \mathcal{B}eta(a, b) \\ X|\theta \sim \mathcal{B}er(\theta). \end{cases}$
- ${\mathrm{P}_\theta}^\infty$: distribution over $\mathcal{X}^\infty$ of the random sequence $(X_n)_{n\geq 1} \overset{\text{i.i.d}}{\sim} \mathrm{P}_\theta$
- posterior distribution (conjugate prior) :

$$\boldsymbol{\pi}(\,\cdot\,|x_{1:n}) = \mathcal{B}eta(a+s, b+n-s), \quad s = \sum_1^n x_i.$$

Posterior expectation and variance

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\pi}}(\boldsymbol{\theta}|X_{1:n}) &= \frac{a+\sum_1^n X_i}{a+b+n} \\ &= \frac{a/n+\frac{1}{n}\sum_1^n X_i}{(a+b)/n+1} \\ &\xrightarrow[n\to\infty]{a.s.} \theta_0 \quad \text{under } \mathrm{P}_{\theta_0}^{\infty}\end{aligned}$$

$$\begin{aligned}\mathbb{V}\mathrm{ar}_{\boldsymbol{\pi}}(\boldsymbol{\theta}|X_{1:n}) &= \frac{\Big(a+\sum_1^n X_i\Big)\Big(b+n-\sum_1 X_i\Big)}{\Big(a+b+n\Big)^2\Big(a+b+n+1\Big)} \\ &= \frac{1}{n}\frac{\Big(a/n+\frac{1}{n}\sum_1^n X_i\Big)\Big(b/n+1-\frac{1}{n}\sum_1 X_i\Big)}{\Big((a+b)/n+1\Big)^2\Big((a+b+1)/n+1\Big)} \\ &\underset{\mathrm{P}_{\theta_0}^{\infty}-a.s.}{\sim} \frac{\theta_0(1-\theta_0)}{n} \quad \underset{\text{exercise}}{=} \quad (n\,I(\theta_0))^{-1}\end{aligned}$$

Concentration of the posterior distribution

- Write $\boldsymbol{\theta}_n^* = \boldsymbol{\theta}_n^*(X_{1:n}) = \mathbb{E}_{\boldsymbol{\pi}}(\boldsymbol{\theta}|X_{1:n})$.
- Tchebychev inequality ⇒ $\forall \delta > 0, \forall U = (\boldsymbol{\theta}_n^* - \delta, \boldsymbol{\theta}_n^* + \delta)$,

$$\begin{aligned}\mathbb{P}_{\boldsymbol{\pi}}\left(\boldsymbol{\theta} \notin U|X_{1:n}\right) &= \mathbb{P}_{\boldsymbol{\pi}}\left(\left(\boldsymbol{\theta} - \boldsymbol{\theta}_n^*\right)^2 > \delta^2|X_{1:n}\right) \\ &\leq \frac{\mathbb{V}\mathrm{ar}_{\boldsymbol{\pi}}(\boldsymbol{\theta}|X_{1:n})}{n\delta^2} \\ &\underset{\mathrm{P}_{\theta_0}^{\infty}-a.s.}{\sim} \frac{\theta_0(1-\theta_0)}{n\delta^2} \qquad \xrightarrow[n\to\infty]{a.s.} 0.\end{aligned}$$

- summary : $\mathrm{P}_{\theta_0}^{\infty}$ - a.s., we have that
  - The posterior distribution concentrates around the posterior expectancy $\boldsymbol{\theta}_n^*$
  - $\boldsymbol{\pi}((\theta_0 - \delta, \theta_0 + \delta)|X_{1:n}) \to 0$.

1. When is a Bayesian approach needed ?

2. The Bayesian framework

3. Construction of estimators

4. Prior choices : conjugate priors, exponential family and alternatives

5. A glimpse at Bayesian asymptotics
    - Example : Beta-Binomial model
    - Posterior consistency
    - Asymptotic normality

6. Exercises

Posterior consistency

Definition

Let $\{\mathrm{P}_\theta, \theta \in \Theta\}, \boldsymbol{\pi}$ be a Bayesian model and let $\theta_0 \in \Theta$. The posterior is *consistent* at θ_0 if For all neighborhood U of θ_0,

$$\boldsymbol{\pi}(U|X_{1:n}) \xrightarrow[n\to\infty]{} 1, \quad \mathrm{P}_{\theta_0}^{\infty}\text{-a.s.}$$

- In general consistency holds when Θ is finite dimensional if $\boldsymbol{\pi}$ assigns positive mass to θ_0's neighborhoods.
- See *e.g.* [?], Chapter 1.3, 1.4 for details

Doob's theorem

Theorem

If Θ and $\mathcal{X}$ are complete, separable, metric spaces endowed with their Borel σ-field, if $\theta \mapsto P_\theta$ is 1 to 1, then for any prior $\boldsymbol{\pi}$ on Θ, $\exists \Theta_0 \subset \Theta$ with $\boldsymbol{\pi}(\Theta_0) = 1$ such that for all $\theta_0 \in \Theta_0$, the posterior is consistent at θ_0.

- **issue** The $\boldsymbol{\pi}$-negligible set where consistency does not hold may be large.
- Under additional regularity conditions, consistency holds at a given θ_0.

Consistency at a given θ_0.

Theorem([?], Th. 1.3.4)

Let Θ be compact, metric and $\theta_0 \in \Theta$. Let $T(x, \theta) = \log \frac{p_\theta(x)}{p_{\theta_0}(x)}$. Assume

1. $\forall x \in \mathcal{X}$, $\theta \mapsto T(x, \theta)$ is continuous
2. $\forall \theta \in \Theta$, $x \mapsto T(x, \theta)$ is measurable
3. $\mathbb{E}\left(\sup_{\theta \in \Theta} |T(\theta, X_1)|\right) < \infty$.

Then

1. The maximum likelihood estimator is consistent at θ_0 (CV in proba)
2. If $\theta_0 \in \mathrm{Supp}(\pi)$, then the posterior is consistent at θ_0.

1. When is a Bayesian approach needed ?

2. The Bayesian framework

3. Construction of estimators

4. Prior choices : conjugate priors, exponential family and alternatives

5. A glimpse at Bayesian asymptotics
   - Example : Beta-Binomial model
   - Posterior consistency
   - Asymptotic normality

6. Exercises

Bayesian asymptotic normality : Overview

- Tells us about the rate of convergence of $\boldsymbol{\pi}(\,\cdot\,|X_{1:n})$ towards δ_{θ_0}.

- With a $\sqrt{n}$ re-scaling, a Gaussian limit centered at the MLE (under appropriate regularity conditions)

- Good references : [?], [?], [?]

Bernstein - Von Mises Theorem

(stated for $\Theta \subset \mathbb{R}$, similar statements for $\Theta \subset \mathbb{R}^d$).

Theorem

Under appropriate regularity conditions (detailed in [?], Th. 1.4.2),
Let $\mathbf{s} = \sqrt{n}(\boldsymbol{\theta} - \widehat{\theta}_n(X_{1:n}))$, with $\widehat{\theta}(X_{1:n})$ the MLE. Let $\pi^*(s|X_{1:n})$ be the posterior density of $\mathbf{s}$. Then

$$\int_{\mathbb{R}} \left| \pi^*(s|X_{1:n}) - \sqrt{\frac{I(\theta_0)}{2\pi}} e^{\frac{-s^2 I(\theta_0)}{2}} \right| \mathrm{d}s \xrightarrow[n\to\infty]{a.s.} 0 \text{ under } \mathrm{P}_{\theta_0}^{\infty}$$

- Interpretation : as $n \to \infty$,

$$\sqrt{n}(\boldsymbol{\theta} - \widehat{\theta}_n(X_{1:n})) \overset{d}{\approx} \mathcal{N}(0, I(\theta_0)^{-1}), \ i.e.$$

$$\boldsymbol{\theta} \overset{d}{\approx} \mathcal{N}\left(\widehat{\theta}_n, \frac{I(\theta_0)^{-1}}{n}\right)$$

- Multivariate case : similar result with multivariate Gaussian and Fisher information matrix.

Asymptotic normality of the posterior mean

$\theta_n^* = \mathbb{E}_{\boldsymbol{\pi}}[\boldsymbol{\theta}|X_{1:n}]$, $\quad \widehat{\theta}_n$: maximum likelihood.

Theorem

In addition to the assumptions of BVM Theorem, assume $\int_{\mathbb{R}} |\theta|\pi(\theta)\,\mathrm{d}\theta < \infty$. Then under $\mathrm{P}_{\theta_0}^{\infty}$,

1. $\sqrt{n}(\theta_n^* - \widehat{\theta}_n) \xrightarrow[n\to\infty]{} 0$ in probability
2. $\sqrt{n}(\theta_n^* - \theta_0) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, I(\theta_0)^{-1})$.

Regularity conditions for BVM theorem

1. $\{x \in \mathcal{X} : p_\theta(x) > 0\}$ does not depend on θ

2. $L(\theta, x) = \log p_\theta(x)$ is three times differentiable *w.r.t.* θ in a neighborhood of θ_0.

3. $\mathbb{E}_{\theta_0}|\frac{\partial}{\partial\theta} L(\theta_0, X)| < \infty, \mathbb{E}_{\theta_0}|\frac{\partial^2}{\partial\theta^2} L(\theta_0, X)| < \infty$ and $\mathbb{E}_{\theta_0} \sup_{\theta\in(\theta_0-\delta,\theta_0+\delta)} \frac{\partial^3}{\partial\theta^3} L(\theta_0, X)| < \infty$

4. $\int_{\mathcal{X}}$ and ∂_θ may be interchanged.

5. $I(\theta_0) > 0$.

Remark : under these conditions the MLE is asymptotically normal, $\sqrt{n}(\widehat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}(0, I(\theta_0)^{-1}$ as well.

1. When is a Bayesian approach needed ?

2. The Bayesian framework

3. Construction of estimators
   - Point estimation
   - Interval estimation

4. Prior choices : conjugate priors, exponential family and alternatives
   - Exponential family
   - conjugate priors in exponential families
   - Prior choice

5. A glimpse at Bayesian asymptotics
   - Example : Beta-Binomial model
   - Posterior consistency
   - Asymptotic normality

6. Exercises

Simpson Paradox

A university is accused of sexual disrimination beacuse 45% of male applicants are accepted versus only 35% for female applicants.

However, each department (art department and engineering department) accepts more female applicants than male applicants.

How do you explain this ?

hint : use Bayes theorem and the fact that the art department is smaller than the engineering department (fewer applicants) and has a lower overall acceptance rate.

Normalizing constant for the Beta distribution ([?], ex. 2.5)

prove that

$$\int_0^1 \theta^{a-1}(1-\theta)^{b-1}\,\mathrm{d}\theta = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}.$$

Bibliography I

[Berger, 2013] Berger, J. O. (2013).
Statistical decision theory and Bayesian analysis.
Springer Science & Business Media.

[Bishop, 2006] Bishop, C. M. (2006).
Pattern recognition and machine learning.
springer.

[Ghosh and Ramamoorthi, 2003] Ghosh, J. and Ramamoorthi, R. (2003).
Bayesian nonparametrics. 2003.

[Robert, 2007] Robert, C. (2007).
The Bayesian choice : from decision-theoretic foundations to computational implementation.
Springer Science & Business Media.

Bibliography II

[Schervish, 2012] Schervish, M. J. (2012).
Theory of statistics.
Springer Science & Business Media.

[Van der Vaart, 1998] Van der Vaart, A. W. (1998).
Asymptotic statistics, volume 3.
Cambridge university press.