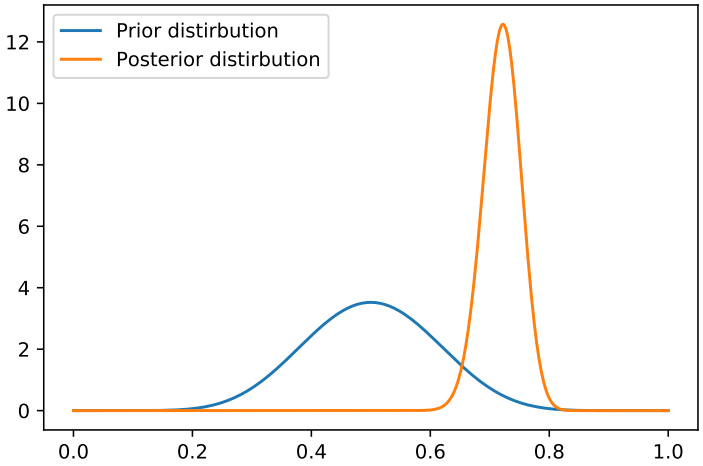
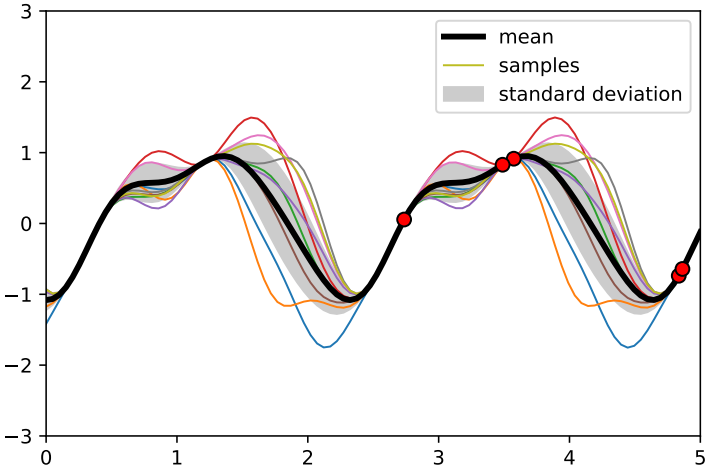
Introduction à l'inférence Bayésienne (Partie I)

Maximilien Servajean

2020-2021





Introduction

- La base de l'étude statistique est la donnée – on notera souvent leur ensemble $\mathcal{D} = \{x_i\}_{i \leq N}$.
- Ces données sont associées à des variables aléatoires associées à un modèle paramétrique f_θ vu comme une approximation d'un phénomène physique duquel proviennent ces dernières.
- Les modèles sont vus comme une approximation d'un mécanisme réel et non comme la vérité.
- Notre objectif est d'extraire de l'information sur θ dans le cadre d'un modèle f_θ après avoir observé $\mathcal{D}$.

Motivations paradigmatiques

- **Motivation 1** : Soit un test T pour une maladie M qui donne la bonne réponse 99% du temps. Je passe le test, ce dernier revient positif. Quelle est la probabilité que j'aie la maladie ?

Motivations paradigmatiques

- **Motivation 1** : Soit un test T pour une maladie M qui donne la bonne réponse 99% du temps. Je passe le test, ce dernier revient positif. Quelle est la probabilité que j'aie la maladie ? Sans plus d'information, aucune idée...

Motivations paradigmatiques

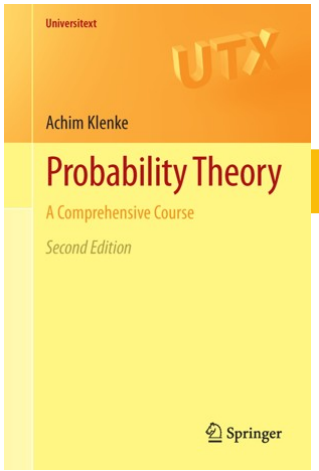
- **Motivation 1** : Soit un test T pour une maladie M qui donne la bonne réponse 99% du temps. Je passe le test, ce dernier revient positif. Quelle est la probabilité que j'aie la maladie ? Sans plus d'information, aucune idée...
- **Motivation 2** : Je fais du *machine learning* et je souhaite résoudre le problème $f^\star = \arg\min_{h \in \mathcal{H}} Re_{\mathcal{S}}(h)$. pour une classe de fonction $\mathcal{H}$. Cependant, $|\mathcal{H}|$ est grand mais je ne souhaite pas le réduire significativement, i.e. je veux considérer des fonctions d'une certaine (ir)régularité.

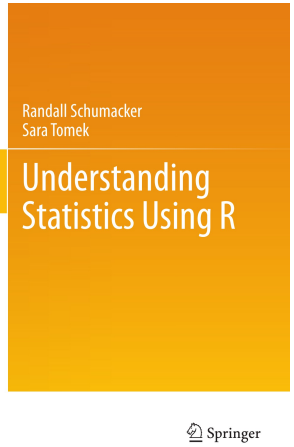
Motivations paradigmatiques

- **Motivation 1** : Soit un test T pour une maladie M qui donne la bonne réponse 99% du temps. Je passe le test, ce dernier revient positif. Quelle est la probabilité que j'aie la maladie ? Sans plus d'information, aucune idée...
- **Motivation 2** : Je fais du *machine learning* et je souhaite résoudre le problème $f^\star = \arg\min_{h \in \mathcal{H}} Re_{\mathcal{S}}(h)$. pour une classe de fonction $\mathcal{H}$. Cependant, $|\mathcal{H}|$ est grand mais je ne souhaite pas le réduire significativement, i.e. je veux considérer des fonctions d'une certaine (ir)régularité.
- **Motivation 3** : Je fais du *machine learning*, mais sachant que $f^\star$ dépend des fluctuations aléatoires sur $\mathcal{S}$, j'aimerais considérer tous les modèles possibles "pondérés" par la "probabilité" qu'ils soient les bons.

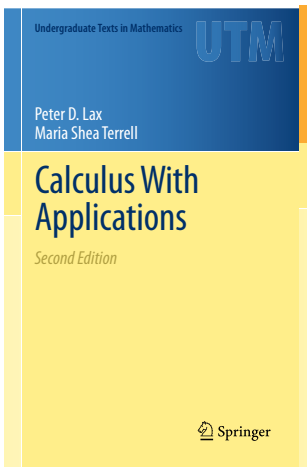
Bibliographie (Rappels)

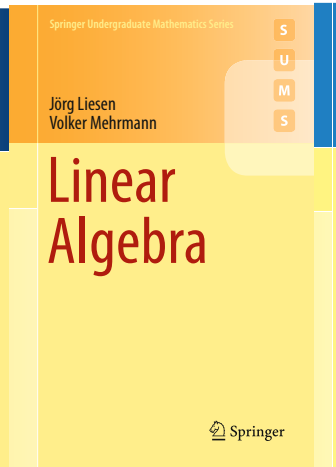
- Proba/stats

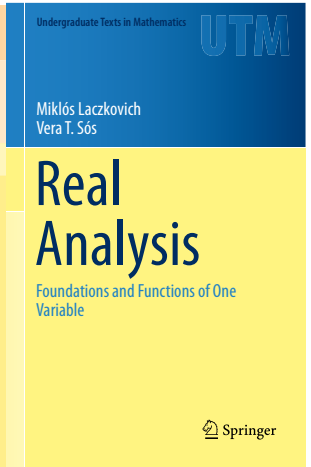




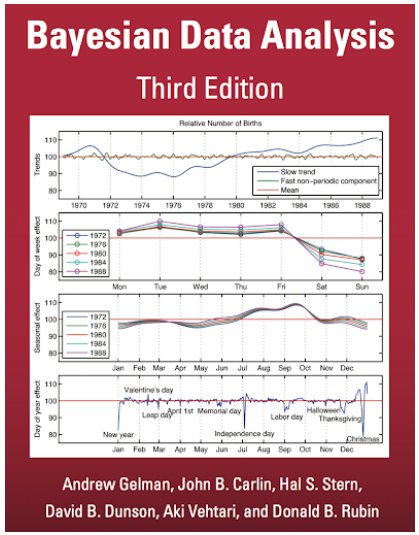
- Math

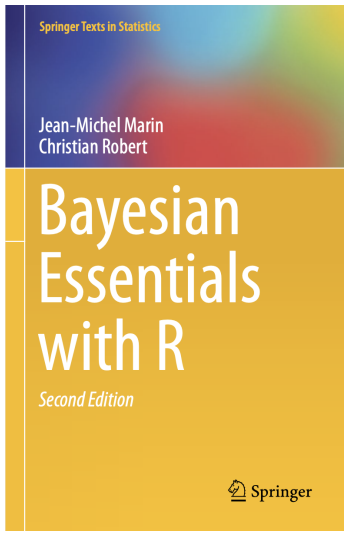


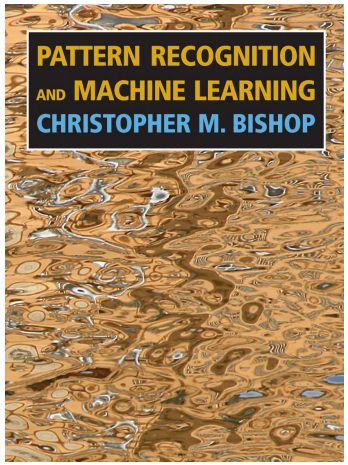




Bibliographie

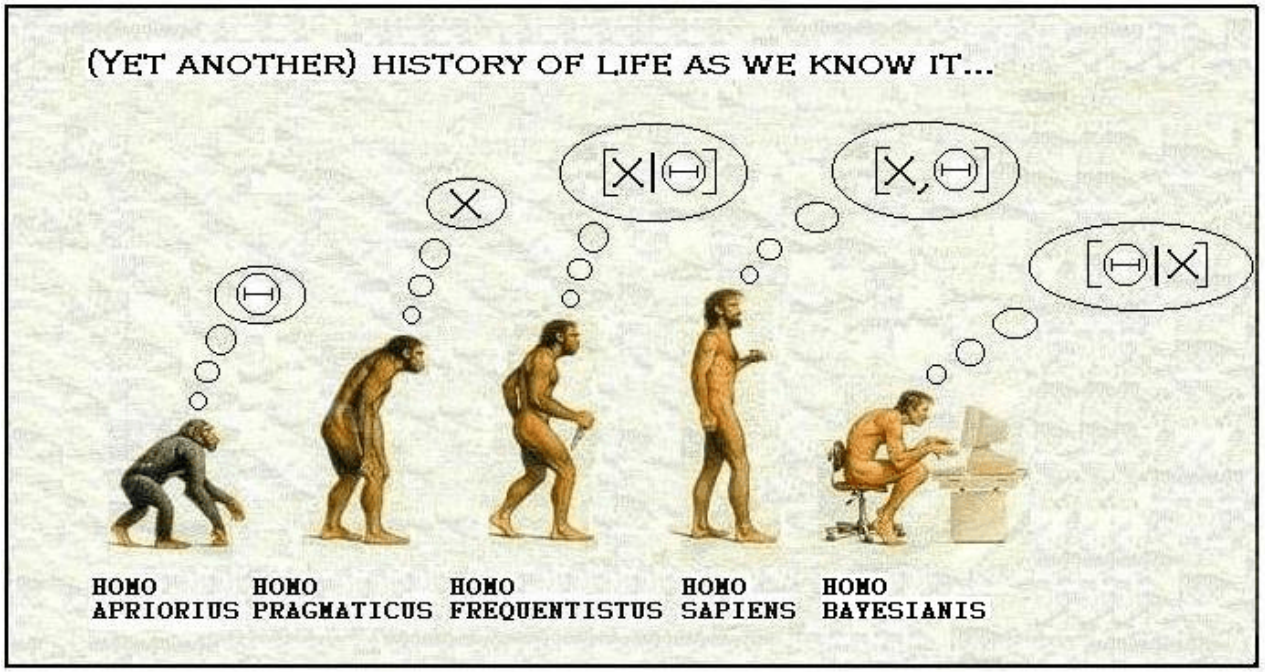






- Livre sur le machine learning avec des références sur l'approche Bayésienne

WE ARE THE BAYESIAN.
YOU WILL BE ASSIMILATED. YOUR TECHNOLOGICAL DISTINCTIVENESS
WILL BE CONSIDERED A SPECIAL CASE OF OUR OWN. RESISTANCE IS FUTILE.


(YET ANOTHER) HISTORY OF LIFE AS WE KNOW IT...
Θ
X
[X|Θ]
[X,Θ]
[Θ|X]
HOMO APRIORIUS
HOMO PRAGMATICUS
HOMO FREQUENTISTUS
HOMO SAPIENS
HOMO BAYESIANIS

Plan

1. Rappels
   - Histoire des probabilités
   - L'étude statistique
   - L'approche fréquentiste
   - Autres outils
2. L'approche Bayésienne
3. Les méthodes MCMC
4. Les méthodes approchées

Histoire des probabilités

- Les probabilités sont l'étude mathématique des choses incertaines,
- la notion de hasard est connue dès l'antiquité (jeux de hasard) mais son étude rigoureuse est beaucoup plus récente,
- le terme probabilité vient du latin *probare* (prouver) qui désigne l'ajout de preuves lors d'un jugement.

Histoire des probabilités (XVIe siècle)

- Premier exposé sur le calcul des probabilités : *Liber de ludo aleae*,
- il étudie la notion d'étique et d'équité dans les jeux (de hasard et de dés en particulier) sous l'angle des probabilités,
- la construction est intuitive : combien de cas favorables par rapport au nombre de cas possibles (notion de répétition de l'expérience).

JérÀŽme Cardan
(1501-1576)

Histoire des probabilités (XVIIe siècle)

- Blaise Pascal (1623-1662) possède une œuvre dans de nombreux domaines,
- dans le cadre des probabilités il se concentre sur la question suivante (avec Pierre de Fermat) :
  - dans une partie entre deux joueurs qui ne peut aller à son terme, comment répartir les gains (problème des partis) ?
  - la solution implique le triangle de Pascal.

$$
\begin{array}{ccccccccccc}
 & & & & & 1 & & & & & \\
 & & & & 1 & & 1 & & & & \\
 & & & 1 & & 2 & & 1 & & & \\
 & & 1 & & 3 & & 3 & & 1 & & \\
 & 1 & & 4 & & 6 & & 4 & & 1 & \\
1 & & 5 & & 10 & & 10 & & 5 & & 1
\end{array}
$$

Histoire des probabilités (XVIIe siècle)

TABLE

DONT IL EST FAIT MENTION DANS LA LETTRE PRÉCÉDENTE.

Si on joue chacun 256, *en*

Il m'appartient sur les 256 pistoles de mon joueur, pour la

	6 Parties.	5 Parties.	4 Parties.	3 Parties.	2 Parties.	1 Partie.
1re Partie.	63	70	80	96	128	256
2e Partie.	63	70	80	96	128	
3e Partie.	56	60	64	64		
4e Partie.	42	40	32			
5e Partie.	24	16				
6e Partie.	8					

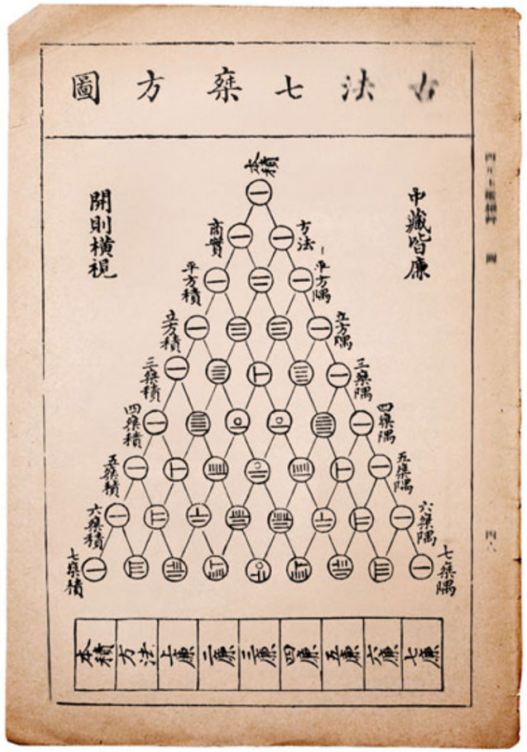
Si on joue 256, *chacun, en*

Il m'appartient sur les 256 de mon joueur, pour les

	6 Parties.	5 Parties.	4 Parties.	3 Parties.	2 Parties.	1 Partie.
La Ire Partie.	63	70	80	96	128	256
2 Ires Parties.	126	140	160	192	256	
3 Ires Parties	182	200	224	256		
4 Ires Parties.	224	240	256			
5 Ires Parties.	248	256				
6 Ires Parties.	256					

Copie d'une lettre de Pascal à Fermat.

Histoire des probabilités (XVIIe siècle)



Le triangle de Pascal, avant Pascal. En Inde connu sous le nom d'Escalier du Mont Meru, en Chine comme le triangle de Yang Hui et en Perse comme le triangle de Khayyam.

Histoire des probabilités (XVIIe siècle)

- Une contribution également philosophique :
  - "La fortune incertaine y est si bien maîtrisée par l'équité du calcul qu'à chacun des joueurs on assigne toujours exactement ce qui s'accorde avec la justice".

Blaise Pascal
(1623-1662)

Histoire des probabilités (XVIIIe et XIXe siècles)

- Les contributions en probabilité continuent :
  - Gauss, Euler, Bernoulli, Laplace, Poisson, etc.
- L'approche intuitive a ses limites : des paradoxes apparaissent...
  - Le paradoxe de Bertrand (probabilité qu'une corde tirée aléatoirement dans un cercle ait une longueur supérieure au cÀŽté du triangle équilatéral inscrit dans le cercle).

Joseph Louis François Bertrand (1822-1900)

Histoire des probabilités (fin XIXe et début du XXe siècle)

- Rejet de l'approche uniquement intuitive et axiomatisation des mathématiques en générales :
  - Théorie des ensembles,
  - Topologie (étude des notions de continuités, de limites, etc.),
  - Théorie de la mesure,
- Maurice Fréchet dira "Si, par exemple, je dis que je veux ajouter les nombres 8 et 3, vous ne m'accuserez pas d'ignorance parce que je n'ai pas spécifié que 8 est pair et 3 impair.....",
- On étudie les cas généraux, abstraits et on cherche la généralisation (et la rigueur).

David Hilbert
(1862-1943)

Histoire des probabilités (après 1930)

- Dans les années 1930, Kolmogorov rattache les probabilités à la théorie de la mesure de Lebesgue et de Borel,
- les probabilités modélisent le hasard, mais ce dernier n'est plus une composante de la théorie,
- les probabilités deviennent centrales en mathématiques et dans les autres domaines.

Henri-Léon Lebesgue (1875-1941)

Une expérience statistique

Une expérience statistique

Définition

Une expérience statistique est le dispositif qui permettra de décrire les données. On se situe en particulier dans le cadre d'un espace probabilisé paramétrique :

- $\{E, \mathcal{T}, P_\theta\}$, *où* E *est l'espace/univers* ($x_i \in E$), $\mathcal{T}$ *une tribu,* P_θ *une mesure de probabilité,*

Une expérience statistique

Définition

Une expérience statistique est le dispositif qui permettra de décrire les données. On se situe en particulier dans le cadre d'un espace probabilisé paramétrique :

- $\{E, \mathcal{T}, P_\theta\}$, *où* E *est l'espace/univers* ($x_i \in E$), $\mathcal{T}$ *une tribu,* P_θ *une mesure de probabilité,*
- *et on appelle modèle une famille de mesures de probabilité sur* $(E, \mathcal{T})$ *:*
  - $\mathcal{M} = \{P_\theta, \theta \in \Theta\}$

  où Θ *est l'espace des paramètres.*

Tribu sur $\mathbb{R}$

La tribu $\mathcal{T}_{\mathbb{R}}$ contient les parties de $\mathbb{R}$ que l'on veut mesurer.

Tribu sur $\mathbb{R}$

La tribu $\mathcal{T}_{\mathbb{R}}$ contient les parties de $\mathbb{R}$ que l'on veut mesurer.

Définition

Soit l'ensemble $\mathbb{R}$*. Une tribu* $\mathcal{T}_{\mathbb{R}}$ *sur* $\mathbb{R}$ *est un ensemble de parties de* $\mathbb{R}$ *satisfaisants :*

- $\mathbb{R} \in \mathcal{T}_{\mathbb{R}}$,

Tribu sur $\mathbb{R}$

La tribu $\mathcal{T}_{\mathbb{R}}$ contient les parties de $\mathbb{R}$ que l'on veut mesurer.

Définition

Soit l'ensemble $\mathbb{R}$. Une tribu $\mathcal{T}_{\mathbb{R}}$ sur $\mathbb{R}$ est un ensemble de parties de $\mathbb{R}$ satisfaisants :

- $\mathbb{R} \in \mathcal{T}_{\mathbb{R}}$,
- *si $A \in \mathcal{T}_{\mathbb{R}}$ alors $A^c \in \mathcal{T}_{\mathbb{R}}$.*

Tribu sur $\mathbb{R}$

La tribu $\mathcal{T}_{\mathbb{R}}$ contient les parties de $\mathbb{R}$ que l'on veut mesurer.

Définition

Soit l'ensemble $\mathbb{R}$. Une tribu $\mathcal{T}_{\mathbb{R}}$ sur $\mathbb{R}$ est un ensemble de parties de $\mathbb{R}$ satisfaisants :

- $\mathbb{R} \in \mathcal{T}_{\mathbb{R}}$,
- *si* $A \in \mathcal{T}_{\mathbb{R}}$ *alors* $A^c \in \mathcal{T}_{\mathbb{R}}$.
- *si* $A_n \in \mathcal{T}_{\mathbb{R}}$ $\forall n \in \mathbb{N}$ *alors* $\bigcup_{n \in \mathbb{N}} A_n \in \mathcal{T}_{\mathbb{R}}$,

Tribu sur $\mathbb{R}$

La tribu $\mathcal{T}_{\mathbb{R}}$ contient les parties de $\mathbb{R}$ que l'on veut mesurer.

Définition

Soit l'ensemble $\mathbb{R}$. Une tribu $\mathcal{T}_{\mathbb{R}}$ sur $\mathbb{R}$ est un ensemble de parties de $\mathbb{R}$ satisfaisants :

- $\mathbb{R} \in \mathcal{T}_{\mathbb{R}}$,
- *si* $A \in \mathcal{T}_{\mathbb{R}}$ *alors* $A^c \in \mathcal{T}_{\mathbb{R}}$.
- *si* $A_n \in \mathcal{T}_{\mathbb{R}}$ $\forall n \in \mathbb{N}$ *alors* $\bigcup_{n \in \mathbb{N}} A_n \in \mathcal{T}_{\mathbb{R}}$,

En conséquent, on a aussi :

Tribu sur $\mathbb{R}$

La tribu $\mathcal{T}_{\mathbb{R}}$ contient les parties de $\mathbb{R}$ que l'on veut mesurer.

Définition

Soit l'ensemble $\mathbb{R}$. Une tribu $\mathcal{T}_{\mathbb{R}}$ sur $\mathbb{R}$ est un ensemble de parties de $\mathbb{R}$ satisfaisants :

- $\mathbb{R} \in \mathcal{T}_{\mathbb{R}}$,
- *si* $A \in \mathcal{T}_{\mathbb{R}}$ *alors* $A^c \in \mathcal{T}_{\mathbb{R}}$.
- *si* $A_n \in \mathcal{T}_{\mathbb{R}}$ $\forall n \in \mathbb{N}$ *alors* $\bigcup_{n \in \mathbb{N}} A_n \in \mathcal{T}_{\mathbb{R}}$,

En conséquent, on a aussi :$\emptyset \in \mathcal{T}_{\mathbb{R}}$,

Tribu sur $\mathbb{R}$

La tribu $\mathcal{T}_{\mathbb{R}}$ contient les parties de $\mathbb{R}$ que l'on veut mesurer.

Définition

Soit l'ensemble $\mathbb{R}$. Une tribu $\mathcal{T}_{\mathbb{R}}$ sur $\mathbb{R}$ est un ensemble de parties de $\mathbb{R}$ satisfaisants :

- $\mathbb{R} \in \mathcal{T}_{\mathbb{R}}$,
- *si $A \in \mathcal{T}_{\mathbb{R}}$ alors $A^c \in \mathcal{T}_{\mathbb{R}}$.*
- *si $A_n \in \mathcal{T}_{\mathbb{R}}$ $\forall n \in \mathbb{N}$ alors $\bigcup_{n\in\mathbb{N}} A_n \in \mathcal{T}_{\mathbb{R}}$,*

En conséquent, on a aussi :$\emptyset \in \mathcal{T}_{\mathbb{R}}$,et si $A_n \in \mathcal{T}_{\mathbb{R}}$ $\forall n \in \mathbb{N}$ alors $\bigcap_{n\in\mathbb{N}} A_n \in \mathcal{T}_{\mathbb{R}}$.

Tribu sur $\mathbb{R}$

La tribu $\mathcal{T}_{\mathbb{R}}$ contient les parties de $\mathbb{R}$ que l'on veut mesurer.

Définition

Soit l'ensemble $\mathbb{R}$. Une tribu $\mathcal{T}_{\mathbb{R}}$ sur $\mathbb{R}$ est un ensemble de parties de $\mathbb{R}$ satisfaisants :

- $\mathbb{R} \in \mathcal{T}_{\mathbb{R}}$,
- *si* $A \in \mathcal{T}_{\mathbb{R}}$ *alors* $A^c \in \mathcal{T}_{\mathbb{R}}$.
- *si* $A_n \in \mathcal{T}_{\mathbb{R}}$ $\forall n \in \mathbb{N}$ *alors* $\bigcup_{n\in\mathbb{N}} A_n \in \mathcal{T}_{\mathbb{R}}$,

En conséquent, on a aussi :$\emptyset \in \mathcal{T}_{\mathbb{R}}$,et si $A_n \in \mathcal{T}_{\mathbb{R}}$ $\forall n \in \mathbb{N}$ alors $\bigcap_{n\in\mathbb{N}} A_n \in \mathcal{T}_{\mathbb{R}}$.

Exemple

La tribu grossière $\mathcal{T} = \{\emptyset, \mathbb{R}\}$ ou encore la tribu qui contient (engendrée par) les segments $[a, b]$, $\forall a, b \in \mathbb{Q}$ (la tribu Borélienne)

Mesure de probabilité

Définition

Soit un espace mesurable (probabilisable) $\{E, \mathcal{T}\}$*, une application* $\mu : \mathcal{T} \mapsto [0, +\infty]$ *est appelée mesure si elle satisfait :*

Mesure de probabilité

Définition

Soit un espace mesurable (probabilisable) $\{E, \mathcal{T}\}$*, une application* $\mu : \mathcal{T} \mapsto [0, +\infty]$ *est appelée mesure si elle satisfait :*

- *l'ensemble vide a une mesure nulle :* $\mu(\varnothing) = 0$,

Mesure de probabilité

Définition

Soit un espace mesurable (probabilisable) $\{E, \mathcal{T}\}$*, une application* $\mu : \mathcal{T} \mapsto [0, +\infty]$ *est appelée mesure si elle satisfait :*

- *l'ensemble vide a une mesure nulle :* $\mu(\varnothing) = 0$*,*
- *l'application* μ *est* σ*-additive : si* A_1*,* A_2*, ..., forment une famille dénombrable de parties de* E *appartenants à* $\mathcal{T}$ *et si ces parties sont deux à deux disjointes, alors la mesure de leur réunion* A *est égale à la somme des mesures des parties :*

$$\mu(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mu(A_i)$$

Mesure de probabilité

Définition

Soit un espace mesurable (probabilisable) $\{E, \mathcal{T}\}$*, une application* $\mu : \mathcal{T} \mapsto [0, +\infty]$ *est appelée mesure si elle satisfait :*

- *l'ensemble vide a une mesure nulle :* $\mu(\varnothing) = 0$*,*
- *l'application* μ *est* σ*-additive : si* A_1*,* A_2*, ..., forment une famille dénombrable de parties de* E *appartenants à* $\mathcal{T}$ *et si ces parties sont deux à deux disjointes, alors la mesure de leur réunion* A *est égale à la somme des mesures des parties :*

$$\mu(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mu(A_i)$$

Si $\mu(E) = 1$*, alors on parle de mesure de probabilité.*

Exemples

Exemples

Exemple

Nous souhaitons étudier un jeu de pile ou face...

Exemples

Exemple

Nous souhaitons étudier un jeu de pile ou face...

- *l'univers est* $\Omega = \{0, 1\}$,

Exemples

Exemple

Nous souhaitons étudier un jeu de pile ou face...

- *l'univers est* $\Omega = \{0, 1\}$,
- *une tribu est* $\mathcal{T} = \{\emptyset, 0, 1, \Omega\}$,

Exemples

Exemple

Nous souhaitons étudier un jeu de pile ou face...

- *l'univers est* $\Omega = \{0, 1\}$,
- *une tribu est* $\mathcal{T} = \{\emptyset, 0, 1, \Omega\}$,
- *le modèle est* $\mathcal{M} = \{Be(\theta), \theta \in [0, 1]\}$ *où Be signifie loi de Bernoulli.*

Exemples

Exemple

Nous souhaitons étudier un jeu de pile ou face...

- *l'univers est* $\Omega = \{0, 1\}$,
- *une tribu est* $\mathcal{T} = \{\emptyset, 0, 1, \Omega\}$,
- *le modèle est* $\mathcal{M} = \{Be(\theta), \theta \in [0, 1]\}$ *où Be signifie loi de Bernoulli.*

Exemple

Dans le cadre du modèle fondamental $\mathcal{M} = \{\mathcal{N}(\theta, 1), \theta \in \mathbb{R}\}$...

Exemples

Exemple

Nous souhaitons étudier un jeu de pile ou face...

- *l'univers est* $\Omega = \{0, 1\}$,
- *une tribu est* $\mathcal{T} = \{\emptyset, 0, 1, \Omega\}$,
- *le modèle est* $\mathcal{M} = \{Be(\theta), \theta \in [0, 1]\}$ *où Be signifie loi de Bernoulli.*

Exemple

Dans le cadre du modèle fondamental $\mathcal{M} = \{\mathcal{N}(\theta, 1), \theta \in \mathbb{R}\}$...

- *l'univers est* $\Omega = \mathbb{R}$,

Exemples

Exemple

Nous souhaitons étudier un jeu de pile ou face...

- *l'univers est* $\Omega = \{0, 1\}$,
- *une tribu est* $\mathcal{T} = \{\emptyset, 0, 1, \Omega\}$,
- *le modèle est* $\mathcal{M} = \{Be(\theta), \theta \in [0, 1]\}$ *où Be signifie loi de Bernoulli.*

Exemple

Dans le cadre du modèle fondamental $\mathcal{M} = \{\mathcal{N}(\theta, 1), \theta \in \mathbb{R}\}$...

- *l'univers est* $\Omega = \mathbb{R}$,
- *la tribu Borélienne de* $\mathbb{R}$

Une variable aléatoire

Une variable aléatoire

- Soit un espace probabilisé $\{\Omega, \mathcal{F}, \mathbb{P}\}$

Une variable aléatoire

- Soit un espace probabilisé $\{\Omega, \mathcal{F}, \mathbb{P}\}$ et un espace probabilisable $\{E, \mathcal{T}\}$.

Une variable aléatoire

- Soit un espace probabilisé $\{\Omega, \mathcal{F}, \mathbb{P}\}$ et un espace probabilisable $\{E, \mathcal{T}\}$.
- Une variable aléatoire X est une fonction mesurable $X : \Omega \mapsto E$.

Une variable aléatoire

- Soit un espace probabilisé $\{\Omega, \mathcal{F}, \mathbb{P}\}$ et un espace probabilisable $\{E, \mathcal{T}\}$.
- Une variable aléatoire X est une fonction mesurable $X : \Omega \mapsto E$.
- On appelle loi de X sur E la mesure de probabilité $\mathbb{P}_X$ définie de la manière suivante :

$$\forall B \in \mathcal{T}, \mathbb{P}_X(B) = \mathbb{P}(X(B)^{-1})$$

Une variable aléatoire

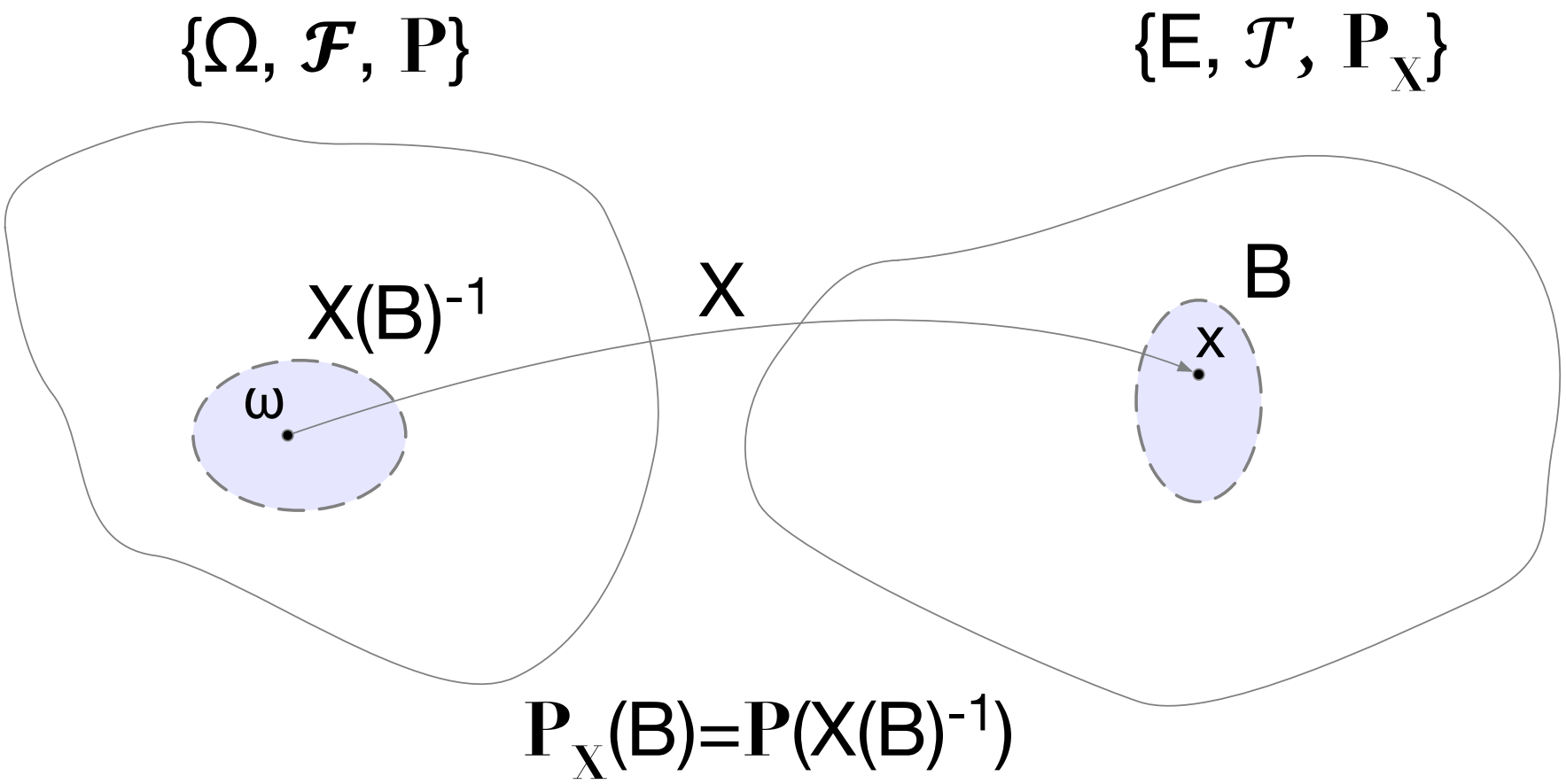
- Soit un espace probabilisé $\{\Omega, \mathcal{F}, \mathbb{P}\}$ et un espace probabilisable $\{E, \mathcal{T}\}$.
- Une variable aléatoire X est une fonction mesurable $X : \Omega \mapsto E$.
- On appelle loi de X sur E la mesure de probabilité $\mathbb{P}_X$ définie de la manière suivante :

$$\forall B \in \mathcal{T}, \mathbb{P}_X(B) = \mathbb{P}(X(B)^{-1})$$

- Dit autrement, il s'agit du *push forward* de $\mathbb{P}$ par X :

$$\mathbb{P}_X = X_\sharp \mathbb{P}.$$

Une variable aléatoire



Exemple

Exemple

- *On souhaite regarder à la sortie d'une usine la proportion de pièces défectueuses,*

Exemple

Exemple

- *On souhaite regarder à la sortie d'une usine la proportion de pièces défectueuses,*
- *pour cela on regarde n pièces et on définit un univers* $\Omega = \{0,1\}^n$ *(on a donc* $\omega = \{\omega_1, ..., \omega_n\}$ *avec* $\omega_i \in \{0,1\}$*).*

Exemple

Exemple

- *On souhaite regarder à la sortie d'une usine la proportion de pièces défectueuses,*
- *pour cela on regarde n pièces et on définit un univers* $\Omega = \{0,1\}^n$ *(on a donc* $\omega = \{\omega_1, ..., \omega_n\}$ *avec* $\omega_i \in \{0,1\}$*).*
- *Soit une fonction de* Ω *dans* $\mathbb{R}$ *définie comme* $X(\omega) = \sum_{i=1}^{n} \frac{\omega_i}{n}$. X *est une variable aléatoire.*

Exemple

Exemple

- *On souhaite regarder à la sortie d'une usine la proportion de pièces défectueuses,*
- *pour cela on regarde n pièces et on définit un univers* $\Omega = \{0,1\}^n$ *(on a donc* $\omega = \{\omega_1, ..., \omega_n\}$ *avec* $\omega_i \in \{0,1\}$*).*
- *Soit une fonction de* Ω *dans* $\mathbb{R}$ *définie comme* $X(\omega) = \sum_{i=1}^{n} \frac{\omega_i}{n}$. X *est une variable aléatoire.*
- *Construisons maintenant la mesure de probabilité des proportions, notée* $\mathbb{P}_X$ *:*

Exemple

Exemple

- *On souhaite regarder à la sortie d'une usine la proportion de pièces défectueuses,*
- *pour cela on regarde n pièces et on définit un univers* $\Omega = \{0,1\}^n$ *(on a donc* $\omega = \{\omega_1, ..., \omega_n\}$ *avec* $\omega_i \in \{0,1\}$*).*
- *Soit une fonction de* Ω *dans* $\mathbb{R}$ *définie comme* $X(\omega) = \sum_{i=1}^{n} \frac{\omega_i}{n}$. X *est une variable aléatoire.*
- *Construisons maintenant la mesure de probabilité des proportions, notée* $\mathbb{P}_X$ *:*

$$\mathbb{P}_X(B) =$$

Exemple

Exemple

- *On souhaite regarder à la sortie d'une usine la proportion de pièces défectueuses,*
- *pour cela on regarde n pièces et on définit un univers* $\Omega = \{0,1\}^n$ *(on a donc* $\omega = \{\omega_1, ..., \omega_n\}$ *avec* $\omega_i \in \{0,1\}$*).*
- *Soit une fonction de* Ω *dans* $\mathbb{R}$ *définie comme* $X(\omega) = \sum_{i=1}^{n} \frac{\omega_i}{n}$. X *est une variable aléatoire.*
- *Construisons maintenant la mesure de probabilité des proportions, notée* $\mathbb{P}_X$ *:*

$$\mathbb{P}_X(B) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \in B\}) \tag{1}$$

Exemple

Exemple

- *Soit* $\Omega = [0, 1]$, $\mathcal{F}$ *la trace de* $\mathcal{B}(\mathbb{R})$ *sur* $[0, 1]$ *et* P *la mesure de probabilité uniforme sur* Ω,

Exemple

Exemple

- *Soit* $\Omega = [0,1]$, $\mathcal{F}$ *la trace de* $\mathcal{B}(\mathbb{R})$ *sur* $[0,1]$ *et* P *la mesure de probabilité uniforme sur* Ω,
- *soit* $E = \mathbb{R}^+$, $\mathcal{T} = \mathcal{B}(\mathbb{R}^+)$ *et* P_X *la mesure image de* P *par la variable aléatoire* X,

Exemple

Exemple

- *Soit* $\Omega = [0,1]$, $\mathcal{F}$ *la trace de* $\mathcal{B}(\mathbb{R})$ *sur* $[0,1]$ *et* P *la mesure de probabilité uniforme sur* Ω,
- *soit* $E = \mathbb{R}^+$, $\mathcal{T} = \mathcal{B}(\mathbb{R}^+)$ *et* P_X *la mesure image de* P *par la variable aléatoire* X,
- *si* $X(\omega) = -\frac{ln(1-\omega)}{\lambda}$, *alors* P_X *est la loi exponentielle.*

Exemple

Exemple

- *Soit* $\Omega = [0,1]$, $\mathcal{F}$ *la trace de* $\mathcal{B}(\mathbb{R})$ *sur* $[0,1]$ *et* P *la mesure de probabilité uniforme sur* Ω,
- *soit* $E = \mathbb{R}^+$, $\mathcal{T} = \mathcal{B}(\mathbb{R}^+)$ *et* P_X *la mesure image de* P *par la variable aléatoire* X,
- *si* $X(\omega) = -\frac{ln(1-\omega)}{\lambda}$, *alors* P_X *est la loi exponentielle.*
  - *Pour simuler selon la loi exponentielle, il suffit ici de simuler entre 0 et 1 puis d'appliquer la fonction/variable aléatoire* X,

Exemple

Exemple

- *Soit* $\Omega = [0,1]$, $\mathcal{F}$ *la trace de* $\mathcal{B}(\mathbb{R})$ *sur* $[0,1]$ *et* P *la mesure de probabilité uniforme sur* Ω,
- *soit* $E = \mathbb{R}^+$, $\mathcal{T} = \mathcal{B}(\mathbb{R}^+)$ *et* P_X *la mesure image de* P *par la variable aléatoire* X,
- *si* $X(\omega) = -\frac{ln(1-\omega)}{\lambda}$, *alors* P_X *est la loi exponentielle.*
  - *Pour simuler selon la loi exponentielle, il suffit ici de simuler entre 0 et 1 puis d'appliquer la fonction/variable aléatoire* X,
  - *si on veut calculer* $P_X(X(\omega) < a)$, *cela revient à* $P(\omega < X^{-1}(a)) =$

Exemple

Exemple

- *Soit* $\Omega = [0,1]$, $\mathcal{F}$ *la trace de* $\mathcal{B}(\mathbb{R})$ *sur* $[0,1]$ *et* P *la mesure de probabilité uniforme sur* Ω,
- *soit* $E = \mathbb{R}^+$, $\mathcal{T} = \mathcal{B}(\mathbb{R}^+)$ *et* P_X *la mesure image de* P *par la variable aléatoire* X,
- *si* $X(\omega) = -\frac{ln(1-\omega)}{\lambda}$, *alors* P_X *est la loi exponentielle.*
  - *Pour simuler selon la loi exponentielle, il suffit ici de simuler entre 0 et 1 puis d'appliquer la fonction/variable aléatoire* X,
  - *si on veut calculer* $P_X(X(\omega) < a)$, *cela revient à* $P(\omega < X^{-1}(a)) = P(\omega < 1 - e^{-\lambda a}) =$

Exemple

Exemple

- *Soit* $\Omega = [0,1]$, $\mathcal{F}$ *la trace de* $\mathcal{B}(\mathbb{R})$ *sur* $[0,1]$ *et* P *la mesure de probabilité uniforme sur* Ω,
- *soit* $E = \mathbb{R}^+$, $\mathcal{T} = \mathcal{B}(\mathbb{R}^+)$ *et* P_X *la mesure image de* P *par la variable aléatoire* X,
- *si* $X(\omega) = -\frac{ln(1-\omega)}{\lambda}$, *alors* P_X *est la loi exponentielle.*
  - *Pour simuler selon la loi exponentielle, il suffit ici de simuler entre 0 et 1 puis d'appliquer la fonction/variable aléatoire* X,
  - *si on veut calculer* $P_X(X(\omega) < a)$, *cela revient à* $P(\omega < X^{-1}(a)) = P(\omega < 1 - e^{-\lambda a}) = 1 - e^{-\lambda a}$,
  - *soit la fonction de répartition.*

Modèle identifiable

Définition

Un modèle est dit identifiable si $\forall A \in \mathcal{T}$*, on a :*

$$P_\theta(A) = P_{\theta'}(A) \Leftrightarrow \theta = \theta'$$

Modèle identifiable

Définition

Un modèle est dit identifiable si $\forall A \in \mathcal{T}$, on a :

$$P_\theta(A) = P_{\theta'}(A) \Leftrightarrow \theta = \theta'$$

- Si un modèle n'est pas identifiable, ses paramètres ne peuvent pas être interprétés significativement (à part scénarios particuliers)...

Mesure dominée

Définition

On dit qu'une mesure μ sur $(E, \mathcal{T})$ est dominée par une mesure ν sur $(E, \mathcal{T})$ si pour tout $A \in \mathcal{T}$, on a

$$\nu(A) = 0 \implies \mu(A) = 0$$

Mesure dominée

Définition

On dit qu'une mesure μ sur $(E, \mathcal{T})$ est dominée par une mesure ν sur $(E, \mathcal{T})$ si pour tout $A \in \mathcal{T}$, on a

$$\nu(A) = 0 \implies \mu(A) = 0$$

et on note cela $\nu \gg \mu$.

Mesure dominée

Définition

On dit qu'une mesure μ sur $(E, \mathcal{T})$ est dominée par une mesure ν sur $(E, \mathcal{T})$ si pour tout $A \in \mathcal{T}$, on a

$$\nu(A) = 0 \implies \mu(A) = 0$$

et on note cela $\nu \gg \mu$. De manière équivalente, on obtient que si μ est dominée par ν, alors μ admet une densité par rapport à ν et cette densité est unique :

$$d\mu = f d\nu$$

Mesure dominée

Définition

On dit qu'une mesure μ sur $(E, \mathcal{T})$ est dominée par une mesure ν sur $(E, \mathcal{T})$ si pour tout $A \in \mathcal{T}$, on a

$$\nu(A) = 0 \implies \mu(A) = 0$$

et on note cela $\nu \gg \mu$. De manière équivalente, on obtient que si μ est dominée par ν, alors μ admet une densité par rapport à ν et cette densité est unique :

$$d\mu = f d\nu$$

Définition

Un modèle $\{P_\theta, \theta \in \Theta\}$ est dit dominé, s'il l'est pour tout $\theta \in \Theta$.

Exemple

- Le modèle fondamental $\mathcal{M} = \{\mathcal{N}(\theta, 1), \theta \in \mathbb{R}\}$ est dominé par la mesure de Lebesgue λ :
  - $dP_\theta = f d\lambda$ avec $f = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\theta)^2}$.

Exemple

- Le modèle fondamental $\mathcal{M} = \{\mathcal{N}(\theta, 1), \theta \in \mathbb{R}\}$ est dominé par la mesure de Lebesgue λ :
  - $dP_\theta = f d\lambda$ avec $f = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\theta)^2}$.
  - On notera en particulier pour tout $A \in \mathcal{T}$:
    - $P_\theta(A) = \int_A dP_\theta$

Exemple

- Le modèle fondamental $\mathcal{M} = \{\mathcal{N}(\theta, 1), \theta \in \mathbb{R}\}$ est dominé par la mesure de Lebesgue λ :
  - $dP_\theta = f d\lambda$ avec $f = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\theta)^2}$.
  - On notera en particulier pour tout $A \in \mathcal{T}$:
    - $P_\theta(A) = \int_A dP_\theta = \int_A 1 \times dP_\theta$

Exemple

- Le modèle fondamental $\mathcal{M} = \{\mathcal{N}(\theta, 1), \theta \in \mathbb{R}\}$ est dominé par la mesure de Lebesgue λ :
  - $dP_\theta = f d\lambda$ avec $f = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\theta)^2}$.
  - On notera en particulier pour tout $A \in \mathcal{T}$:
    - $P_\theta(A) = \int_A dP_\theta = \int_A 1 \times dP_\theta = \int_{x \in A} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\theta)^2} d\lambda(x)$.

Exemple

- Le modèle fondamental $\mathcal{M} = \{\mathcal{N}(\theta, 1), \theta \in \mathbb{R}\}$ est dominé par la mesure de Lebesgue λ :
  - $dP_\theta = f d\lambda$ avec $f = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\theta)^2}$.
  - On notera en particulier pour tout $A \in \mathcal{T}$:
    - $P_\theta(A) = \int_A dP_\theta = \int_A 1 \times dP_\theta = \int_{x \in A} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\theta)^2} d\lambda(x)$.
    - $P_\theta(X \in [\theta - 2, \theta + 2]) =$

Exemple

- Le modèle fondamental $\mathcal{M} = \{\mathcal{N}(\theta, 1), \theta \in \mathbb{R}\}$ est dominé par la mesure de Lebesgue λ :
  - $dP_\theta = f d\lambda$ avec $f = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\theta)^2}$.
  - On notera en particulier pour tout $A \in \mathcal{T}$:
    - $P_\theta(A) = \int_A dP_\theta = \int_A 1 \times dP_\theta = \int_{x \in A} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\theta)^2} d\lambda(x)$.
    - $P_\theta(X \in [\theta - 2, \theta + 2]) = \int_{\theta-2}^{\theta+2} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\theta)^2} dx =$

Exemple

- Le modèle fondamental $\mathcal{M} = \{\mathcal{N}(\theta, 1), \theta \in \mathbb{R}\}$ est dominé par la mesure de Lebesgue λ :
  - $dP_\theta = f d\lambda$ avec $f = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\theta)^2}$.
  - On notera en particulier pour tout $A \in \mathcal{T}$:
    - $P_\theta(A) = \int_A dP_\theta = \int_A 1 \times dP_\theta = \int_{x \in A} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\theta)^2} d\lambda(x)$.
    - $P_\theta(X \in [\theta - 2, \theta + 2]) = \int_{\theta-2}^{\theta+2} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\theta)^2} dx = 0.95$

Exemple

- Le modèle fondamental $\mathcal{M} = \{\mathcal{N}(\theta, 1), \theta \in \mathbb{R}\}$ est dominé par la mesure de Lebesgue λ :
  - $dP_\theta = f d\lambda$ avec $f = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\theta)^2}$.
  - On notera en particulier pour tout $A \in \mathcal{T}$:
    - $P_\theta(A) = \int_A dP_\theta = \int_A 1 \times dP_\theta = \int_{x \in A} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\theta)^2} d\lambda(x)$.
    - $P_\theta(X \in [\theta - 2, \theta + 2]) = \int_{\theta-2}^{\theta+2} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\theta)^2} dx = 0.95$
- Le modèle fondamental n'est cependant pas dominé par la mesure de Dirac δ_x :

Exemple

- Le modèle fondamental $\mathcal{M} = \{\mathcal{N}(\theta, 1), \theta \in \mathbb{R}\}$ est dominé par la mesure de Lebesgue λ :
  - $dP_\theta = f d\lambda$ avec $f = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\theta)^2}$.
  - On notera en particulier pour tout $A \in \mathcal{T}$:
    - $P_\theta(A) = \int_A dP_\theta = \int_A 1 \times dP_\theta = \int_{x \in A} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\theta)^2} d\lambda(x)$.
    - $P_\theta(X \in [\theta - 2, \theta + 2]) = \int_{\theta-2}^{\theta+2} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\theta)^2} dx = 0.95$
- Le modèle fondamental n'est cependant pas dominé par la mesure de Dirac δ_x :
  - La mesure de Dirac est définie comme $\delta_k(x) = \begin{cases} 1 \text{ si } x = k \\ 0 \text{ sinon} \end{cases}$
  - il n'existe aucune densité f telle que $f d\delta$ donnerait le modèle fondamental $\forall x \neq k$.

Exemple

- Le cas discret n'échappe pas à cette règle.
- Soit le modèle de Bernoulli $\mathcal{M} = \{Be(\theta), \theta \in [0, 1]\}$,

Exemple

- Le cas discret n'échappe pas à cette règle.
- Soit le modèle de Bernoulli $\mathcal{M} = \{Be(\theta), \theta \in [0,1]\}$,
- on a $P_\theta = (1-\theta)\delta_0 + \theta\delta_1$, $\theta \in [0,1]$ où δ_0 est la mesure de comptage et vaut 1 lorsque x vaut 0 et vaut 0 ailleurs,

Exemple

- Le cas discret n'échappe pas à cette règle.
- Soit le modèle de Bernoulli $\mathcal{M} = \{Be(\theta), \theta \in [0,1]\}$,
- on a $P_\theta = (1-\theta)\delta_0 + \theta\delta_1$, $\theta \in [0,1]$ où δ_0 est la mesure de comptage et vaut 1 lorsque x vaut 0 et vaut 0 ailleurs,
- le modèle est clairement dominé par la mesure de comptage $\mu = \delta_0 + \delta_1$,

Exemple

- Le cas discret n'échappe pas à cette règle.
- Soit le modèle de Bernoulli $\mathcal{M} = \{Be(\theta), \theta \in [0,1]\}$,
- on a $P_\theta = (1-\theta)\delta_0 + \theta\delta_1$, $\theta \in [0,1]$ où δ_0 est la mesure de comptage et vaut 1 lorsque x vaut 0 et vaut 0 ailleurs,
- le modèle est clairement dominé par la mesure de comptage $\mu = \delta_0 + \delta_1$,
- avec $f_\theta(x) = \exp\big(x\log(\frac{\theta}{1-\theta}) + \log(1-\theta)\big) = \theta^x(1-\theta)^{1-x}$, où $x \in \mathbb{R}$.

Intégration et changement de variables

Définition (Cas univarié)

Soit I une partie de $\mathbb{R}$ et soit $[a, b] \subset I$, soit $\phi : [a, b] \mapsto I$ une fonction dérivable de dérivée intégrable et $f : I \mapsto \mathbb{R}$ une fonction continue. Nous avons l'égalité suivante :

$$\int_{[a,b]} f(\phi(x))|\phi'(x)|dx = \int_{\phi([a,b])} f(x)dx,$$

où $a = \phi^{-1}(\phi(a))$.

Définition (Cas multivarié)

Soit T une partie de $\mathbb{R}^n$, $d\mu = dx \wedge ... \wedge dy$, ϕ notre fonction de changement de variables et f la fonction à intégrer. Nous avons :

$$\int_T f(\phi(\boldsymbol{x}))|detJ_\phi|d\mu = \int_{\phi(T)} f(\boldsymbol{x})d\mu.$$

Python/R

Python ou R fournissent par défaut ou au travers de librairie un accès à certaines loi de probabilité. Exemple avec la loi normale :

$$f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right)$$

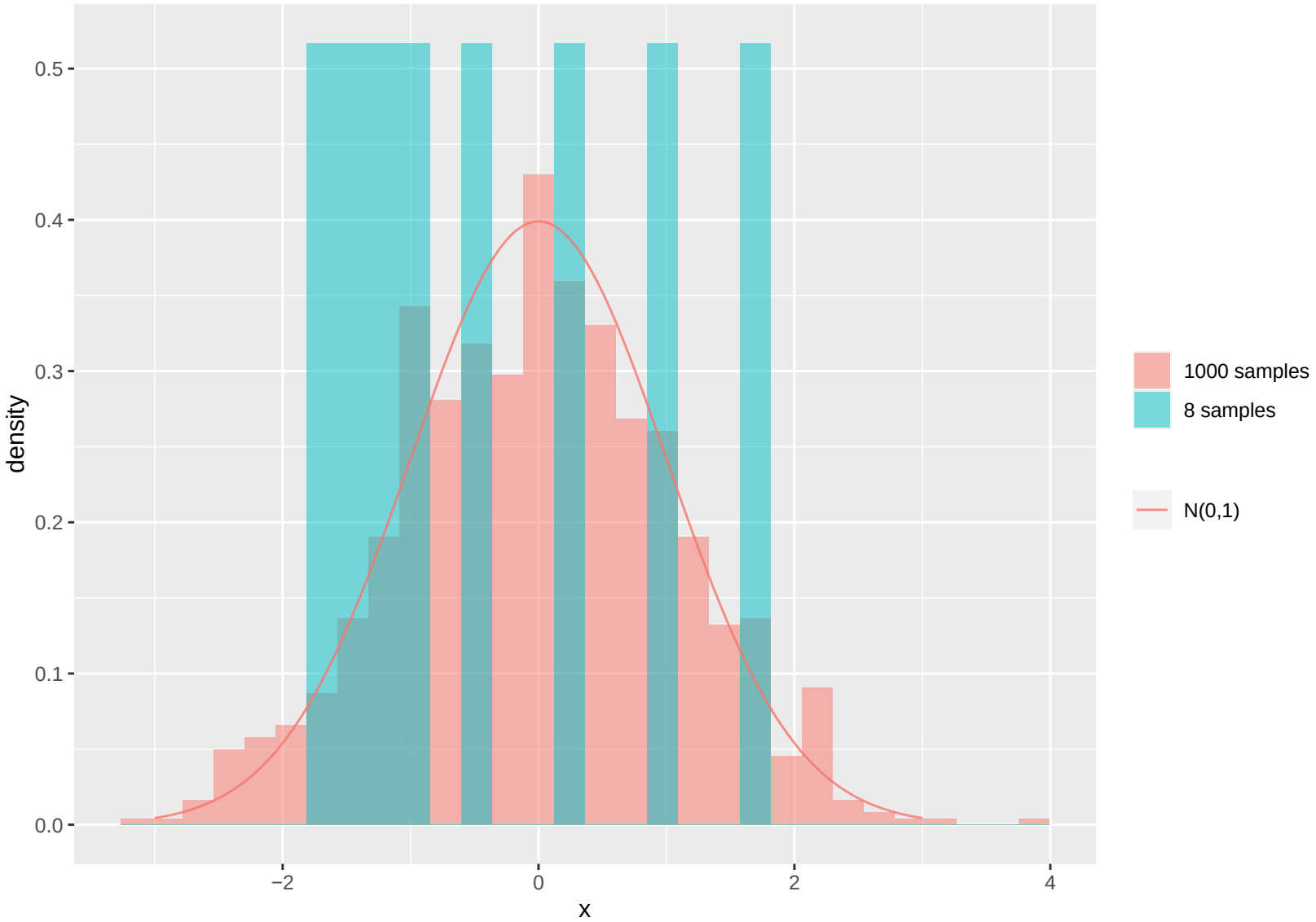
Python :

```
from scipy.stats import norm
import numpy as np
x = norm.rvs(loc=0, scale=1, size=100)
pdf = norm.pdf(np.linspace(-3,3,1000), loc=0, scale=1)
...
```

R :

```
x <- rnorm(n = 100, mean = 0, sd = 1)
pdf <- dnorm(seq(-3, 3, 0.006), mean = 0, sd = 1)
...
```

Échantillonage



Vraisemblance

Définition

Étant donné une mesure de probabilité P_θ et sa densité f_θ (par rapport à une mesure....), on appelle vraisemblance d'une donnée x, sa densité :

$$\mathcal{L}_\theta(x) = f_\theta(x)$$

Vraisemblance

Définition

Étant donné une mesure de probabilité P_θ et sa densité f_θ (par rapport à une mesure....), on appelle vraisemblance d'une donnée x, sa densité :

$$\mathcal{L}_\theta(x) = f_\theta(x)$$

On appelle également vraisemblance d'un échantillon $(x_1, ..., x_N)$ i.i.d. le produit de leur densité :

$$\mathcal{L}_\theta(x_1, ..., x_N) = \prod_{i=1}^{n} f_\theta(x_i)$$

Statistiques et estimateurs

Définition

On appelle statistique $S(\mathcal{D})$ une fonction des données $\mathcal{D}$.

Définition

On appelle estimateur $\hat{\theta}(\mathcal{D})$ une statistique à valeur dans l'espace des paramètres.

Exemple

- Soit un échantillon $\mathcal{D} = (x_1, ..., x_n)$ i.i.d. par exemple selon le modèle fondamental. L'opération $\overline{x} = \frac{1}{n} \sum_i x_i$ est une statistique mais aussi un estimateur.

Statistique exhaustive ou suffisante

Définition

Soit $\mathcal{D} \sim f_\theta^N$. On appelle statistique exhaustive ou suffisante une statistique $S(\mathcal{D})$ telle la probabilité d'observer $\mathcal{D}$ sachant $S(\mathcal{D})$ est indépendante de θ : $\mathbb{P}(\mathcal{D}|S(\mathcal{D}), \theta) = \mathbb{P}(\mathcal{D}|S(\mathcal{D}))$.

Exemple

Soit $f_{\theta=(\mu,\sigma)}(\mathcal{D}) = g(\theta|\mathcal{D})$. On a

$$
\begin{aligned}
g(\theta|\mathcal{D}) &= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} exp\Big(-\frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2}\Big) \\
&\propto exp\Big(-\frac{\sum_{i=1}^{N}(x_i-\mu)^2}{2\sigma^2}\Big)/\sigma^N \\
&\propto exp\Big(-\frac{N\mu^2 - 2\mu N\overline{x} + \sum_i x_i^2}{2\sigma^2}\Big)/\sigma^N \\
&\propto exp\Big(-\frac{N(\mu-\overline{x})^2 + s^2}{2\sigma^2}\Big)/\sigma^N, s^2 = \sum_i (x_i - \overline{x})^2
\end{aligned}
$$

où $\overline{x}$ et s^2 sont suffisantes.

L'approche fréquentiste

- L'approche fréquentiste suppose qu'il existe une "vraie" valeur du paramètre θ telle que $X \sim P_\theta$.

L'approche fréquentiste

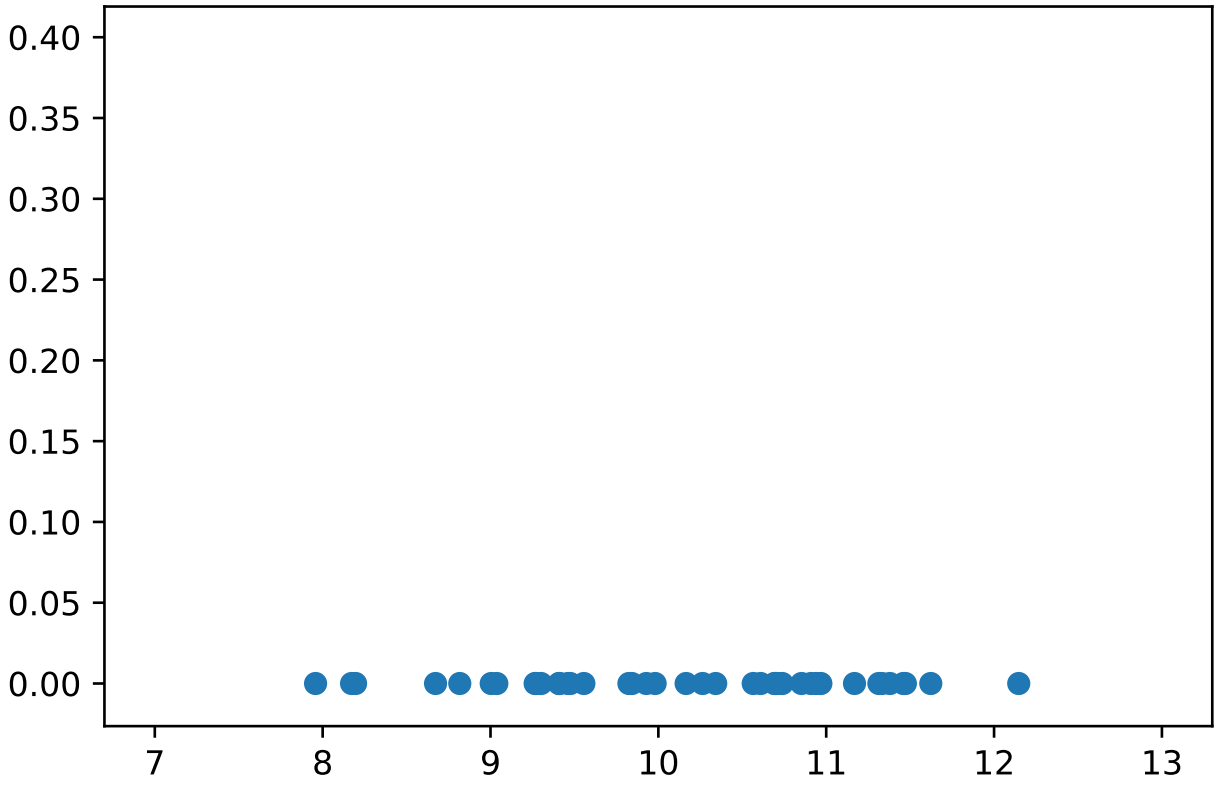
- L'approche fréquentiste suppose qu'il existe une "vraie" valeur du paramètre θ telle que $X \sim P_\theta$.
- L'objectif consiste donc à construire un bon estimateur du "vrai" θ.

L'approche fréquentiste

- L'approche fréquentiste suppose qu'il existe une "vraie" valeur du paramètre θ telle que $X \sim P_\theta$.
- L'objectif consiste donc à construire un bon estimateur du "vrai" θ.
- Il existe plusieurs stratégies comme la méthode des moments ou le maximum de vraisemblance

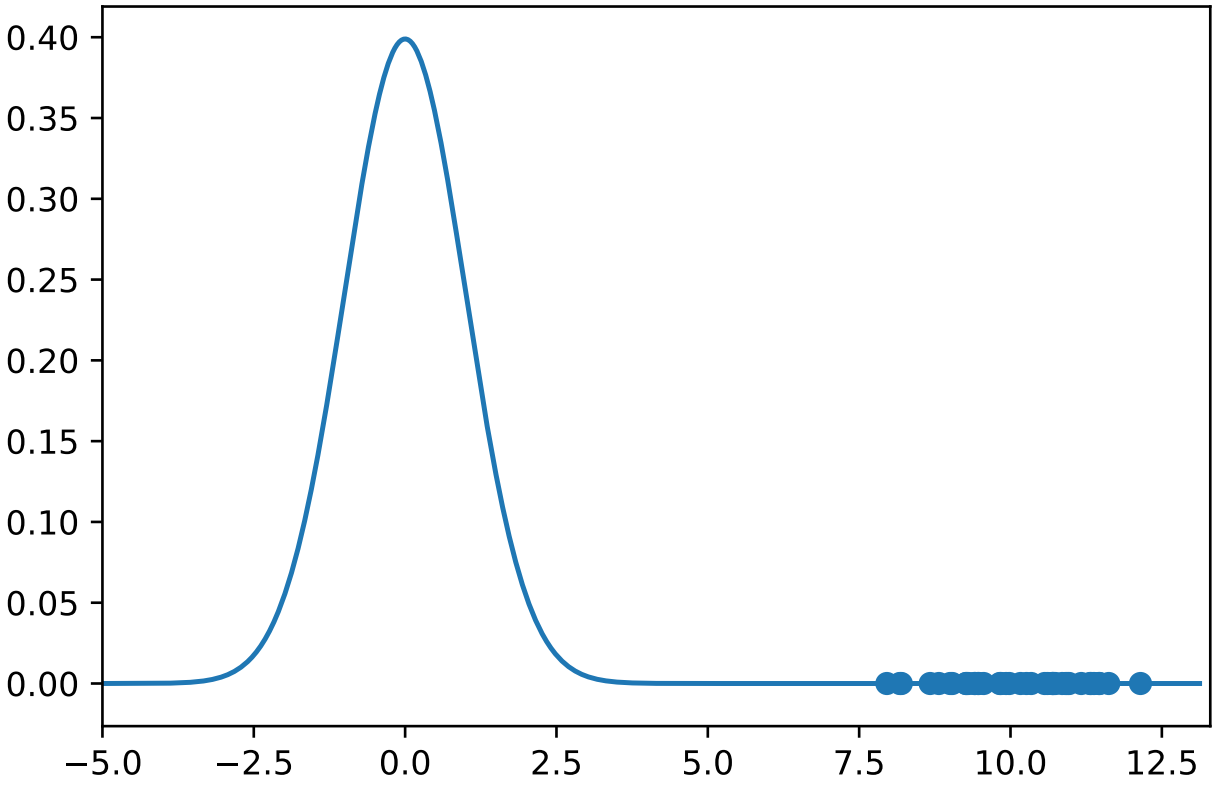
Exemple

Pour le modèle fondamental :



Exemple

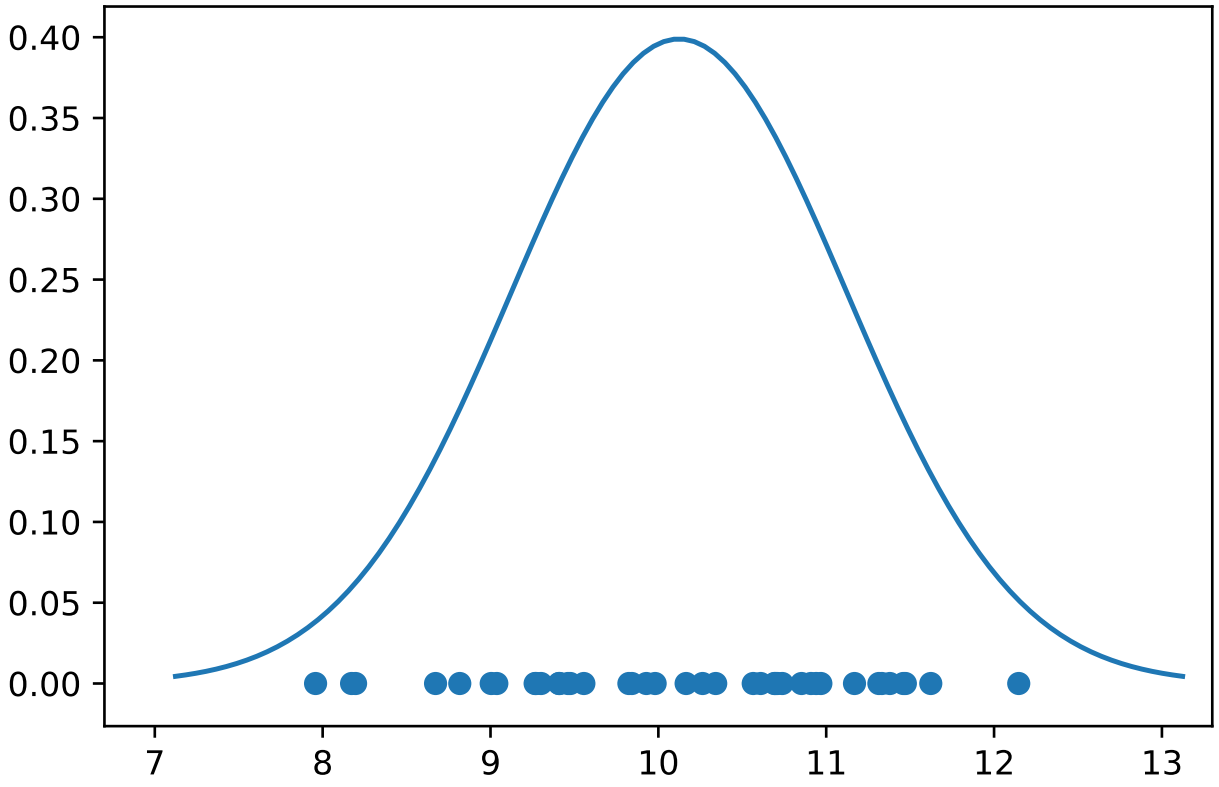
Pour le modèle fondamental :



$$\hat{\theta} = 0$$

Exemple

Pour le modèle fondamental :



$$\hat{\theta} = \overline{X}$$

Exemple

- À votre avis et intuitivement, quel peut-être un bon estimateur d'une loi de Bernoulli pour un jeu de pile ou face ?

Exemple

- À votre avis et intuitivement, quel peut-être un bon estimateur d'une loi de Bernoulli pour un jeu de pile ou face ?
  - $\frac{\#piles}{\#piles+\#faces}$

Loi conditionnelle

- Étant donnée deux variables aléatoires discrètes X, Y à valeur dans E, la loi conditionnelle de X par rapport à Y est la loi de X sachant la réalisation de Y et est définie par :

Loi conditionnelle

- Étant donnée deux variables aléatoires discrètes X, Y à valeur dans E, la loi conditionnelle de X par rapport à Y est la loi de X sachant la réalisation de Y et est définie par :

$$P(X|Y) = \frac{P(X \cap Y)}{p(Y)}$$

Loi conditionnelle

- Étant donnée deux variables aléatoires discrètes X, Y à valeur dans E, la loi conditionnelle de X par rapport à Y est la loi de X sachant la réalisation de Y et est définie par :

$$P(X|Y) = \frac{P(X \cap Y)}{p(Y)}$$

- dans le cas continu (à densité), la loi conditionnelle est définie comme :

$$f_{X|Y=y}(x) = \frac{f_{(X,Y)}(x,y)}{f_Y(y)}$$

Somme de variables aléatoires discrètes

- Soit X et Y deux variables aléatoires indépendentes définies sur $I = [a, b] \subset \mathbb{N}$ et soit $Z = X + Y$,

Somme de variables aléatoires discrètes

- Soit X et Y deux variables aléatoires indépendentes définies sur $I = [a, b] \subset \mathbb{N}$ et soit $Z = X + Y$,
- Z est définie sur $[2a, 2b]$.

Somme de variables aléatoires discrètes

- Soit X et Y deux variables aléatoires indépendentes définies sur $I = [a, b] \subset \mathbb{N}$ et soit $Z = X + Y$,
- Z est définie sur $[2a, 2b]$.
- On a par construction :

$$\mathbb{P}(Z = z) = \sum_{x \in I} \mathbb{P}(X = x)\mathbb{P}(Y = z - x)$$

Somme de variables aléatoires discrètes

- Soit X et Y deux variables aléatoires indépendentes définies sur $I = [a, b] \subset \mathbb{N}$ et soit $Z = X + Y$,
- Z est définie sur $[2a, 2b]$.
- On a par construction :

$$\mathbb{P}(Z = z) = \sum_{x \in I} \mathbb{P}(X = x)\mathbb{P}(Y = z - x)$$

- On se doute que $\mathbb{P}(Z = z)$ est maximum lorsque la valeur z rend la superposition de $\mathbb{P}(X = x)$ et $\mathbb{P}(y = z - x)$ la plus parfaite possible.

Somme de variables aléatoires continues (1/4)

- Soit X et Y deux variables aléatoires indépendantes de densités f et g définies sur $\mathbb{R}$ et $Z = X + Y$,
- On cherche la loi de Z

Somme de variables aléatoires continues (1/4)

- Soit X et Y deux variables aléatoires indépendantes de densités f et g définies sur $\mathbb{R}$ et $Z = X + Y$,
- On cherche la loi de Z et on commence par la densité cumulée de la manière suivante :

$$\begin{aligned} H(z) &= \int\int_{x+y\leq z} f(x)g(y)dydx \\ &= \int_{-\infty}^{\infty} f(x) \int_{-\infty}^{z-x} g(y)dydx \\ &= \int_{-\infty}^{\infty} f(x)G(z-x)dx \end{aligned}$$

Somme de variables aléatoires continues (2/4)

- Afin de retrouver la densité il suffit de dériver la densité cumulée :

Somme de variables aléatoires continues (2/4)

- Afin de retrouver la densité il suffit de dériver la densité cumulée :

$$h(z) = \int_{-\infty}^{\infty} f(x)g(z-x)dx$$

Somme de variables aléatoires continues (2/4)

- Afin de retrouver la densité il suffit de dériver la densité cumulée :

$$h(z) = \int_{-\infty}^{\infty} f(x)g(z-x)dx$$

- On reconnaît l'opérateur de convolution $*$ défini comme :

$$(g * f)(x) = \int_{-\infty}^{\infty} f(t)g(x-t)dt.$$

Somme de variables aléatoires continues (2/4)

- Afin de retrouver la densité il suffit de dériver la densité cumulée :

$$h(z) = \int_{-\infty}^{\infty} f(x)g(z-x)dx$$

- On reconnaît l'opérateur de convolution $*$ défini comme :

$$(g * f)(x) = \int_{-\infty}^{\infty} f(t)g(x-t)dt.$$

- On observe intuitivement que $g * f$ atteint sa valeur max lorsque g et f sont bien superposées,

Somme de variables aléatoires continues (2/4)

- Afin de retrouver la densité il suffit de dériver la densité cumulée :

$$h(z) = \int_{-\infty}^{\infty} f(x)g(z-x)dx$$

- On reconnaît l'opérateur de convolution $*$ défini comme :

$$(g * f)(x) = \int_{-\infty}^{\infty} f(t)g(x-t)dt.$$

- On observe intuitivement que $g * f$ atteint sa valeur max lorsque g et f sont bien superposées,
- En remplaçant dt par une "mesure" discrète, nous retombons sur le cas discret.

Somme de variables aléatoires continues (2/4)

- Afin de retrouver la densité il suffit de dériver la densité cumulée :

$$h(z) = \int_{-\infty}^{\infty} f(x)g(z-x)dx$$

- On reconnaît l'opérateur de convolution $*$ défini comme :

$$(g * f)(x) = \int_{-\infty}^{\infty} f(t)g(x-t)dt.$$

- On observe intuitivement que $g * f$ atteint sa valeur max lorsque g et f sont bien superposées,
- En remplaçant dt par une "mesure" discrète, nous retombons sur le cas discret.
- C'est la convolution utilisée en *deep learning* : un neurone convolutif possède une activation maximale si son "filtre" et bien superposé au motif de l'image.

Somme de variables aléatoires continues (3/4)

Exemple

- *Soit deux variables* X *et* Y *de densité* f *et* g *égales à* $1/n$ *sur* $[0, n]$ *et* 0 *ailleurs*

Somme de variables aléatoires continues (3/4)

Exemple

- *Soit deux variables X et Y de densité f et g égales à $1/n$ sur $[0, n]$ et 0 ailleurs (il s'agit bien entendu d'une loi uniforme),*

Somme de variables aléatoires continues (3/4)

Exemple

- *Soit deux variables X et Y de densité f et g égales à $1/n$ sur $[0, n]$ et 0 ailleurs (il s'agit bien entendu d'une loi uniforme),*
- *Convoluons les deux densités :*

$$(g * f)(x) = \int_{-\infty}^{\infty} f(t)g(x-t)dt = \frac{1}{n}\int_{0}^{n} g(x-t)dt$$

Somme de variables aléatoires continues (3/4)

Exemple

- *Soit deux variables X et Y de densité f et g égales à $1/n$ sur $[0, n]$ et 0 ailleurs (il s'agit bien entendu d'une loi uniforme),*
- *Convoluons les deux densités :*

$$(g * f)(x) = \int_{-\infty}^{\infty} f(t)g(x-t)dt = \frac{1}{n}\int_{0}^{n} g(x-t)dt$$

- *Évidemment, si $x \notin [0, 2n]$, nous avons $g(x-t) = 0$,*

Somme de variables aléatoires continues (3/4)

Exemple

- *Soit deux variables X et Y de densité f et g égales à $1/n$ sur $[0, n]$ et 0 ailleurs (il s'agit bien entendu d'une loi uniforme),*
- *Convoluons les deux densités :*

$$(g * f)(x) = \int_{-\infty}^{\infty} f(t)g(x-t)dt = \frac{1}{n}\int_{0}^{n} g(x-t)dt$$

- *Évidemment, si $x \notin [0, 2n]$, nous avons $g(x-t) = 0$,*
- *De même si $t \notin [x-n, x]$, alors $g(x-t) = 0$,*

Somme de variables aléatoires continues (3/4)

Exemple

- *Soit deux variables X et Y de densité f et g égales à $1/n$ sur $[0, n]$ et 0 ailleurs (il s'agit bien entendu d'une loi uniforme),*
- *Convoluons les deux densités :*

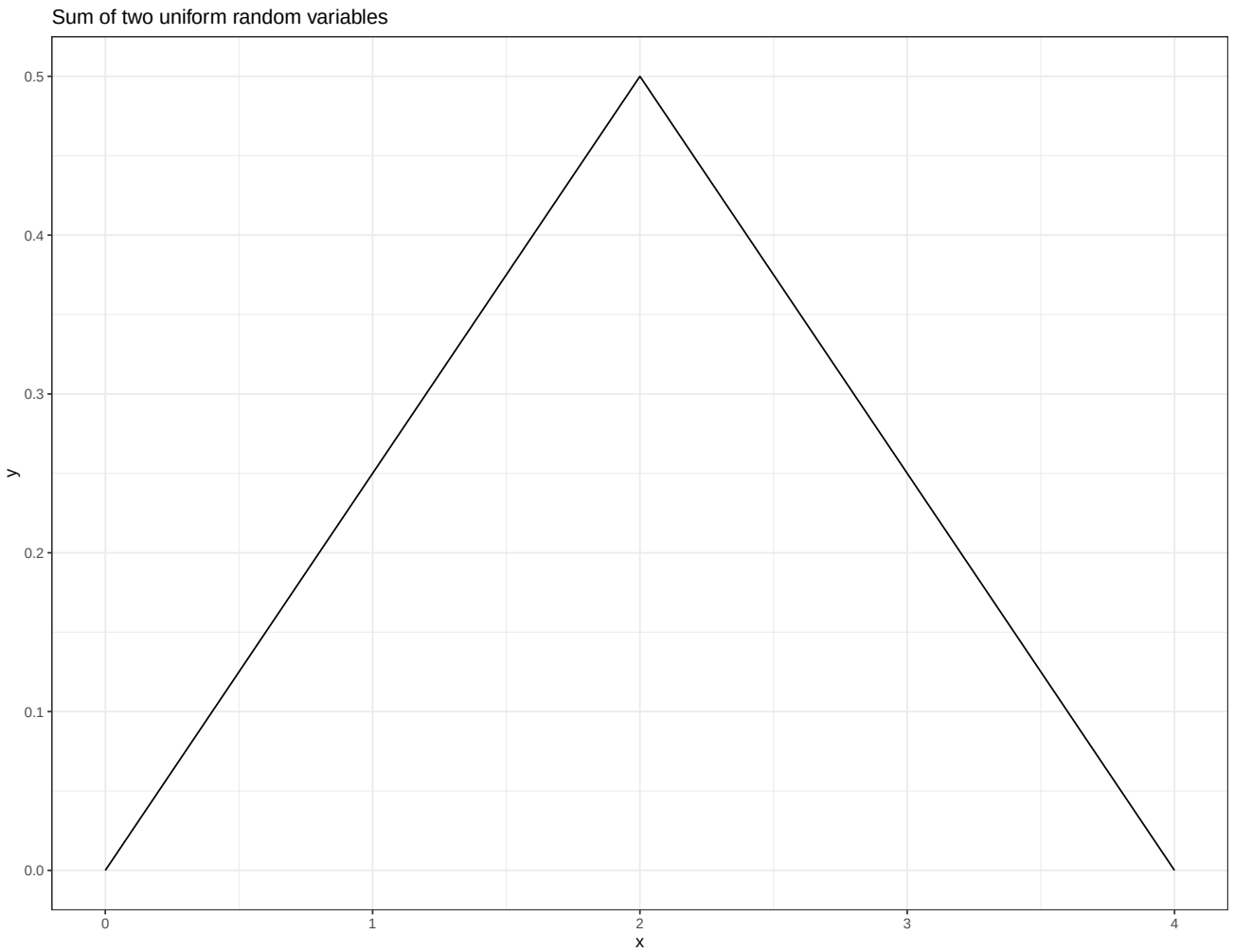
$$(g * f)(x) = \int_{-\infty}^{\infty} f(t)g(x-t)dt = \frac{1}{n}\int_0^n g(x-t)dt$$

- *Évidemment, si $x \notin [0, 2n]$, nous avons $g(x-t) = 0$,*
- *De même si $t \notin [x-n, x]$, alors $g(x-t) = 0$,*
- *On a ainsi :*

$$(g * f)(x) = \frac{1}{n^2}\int_0^x dt = \frac{x}{n^2} \ si \ x \leq n \ et$$
$$(g * f)(x) = \frac{1}{n^2}\int_{x-n}^n dt = \frac{2n-x}{n^2} \ sinon$$

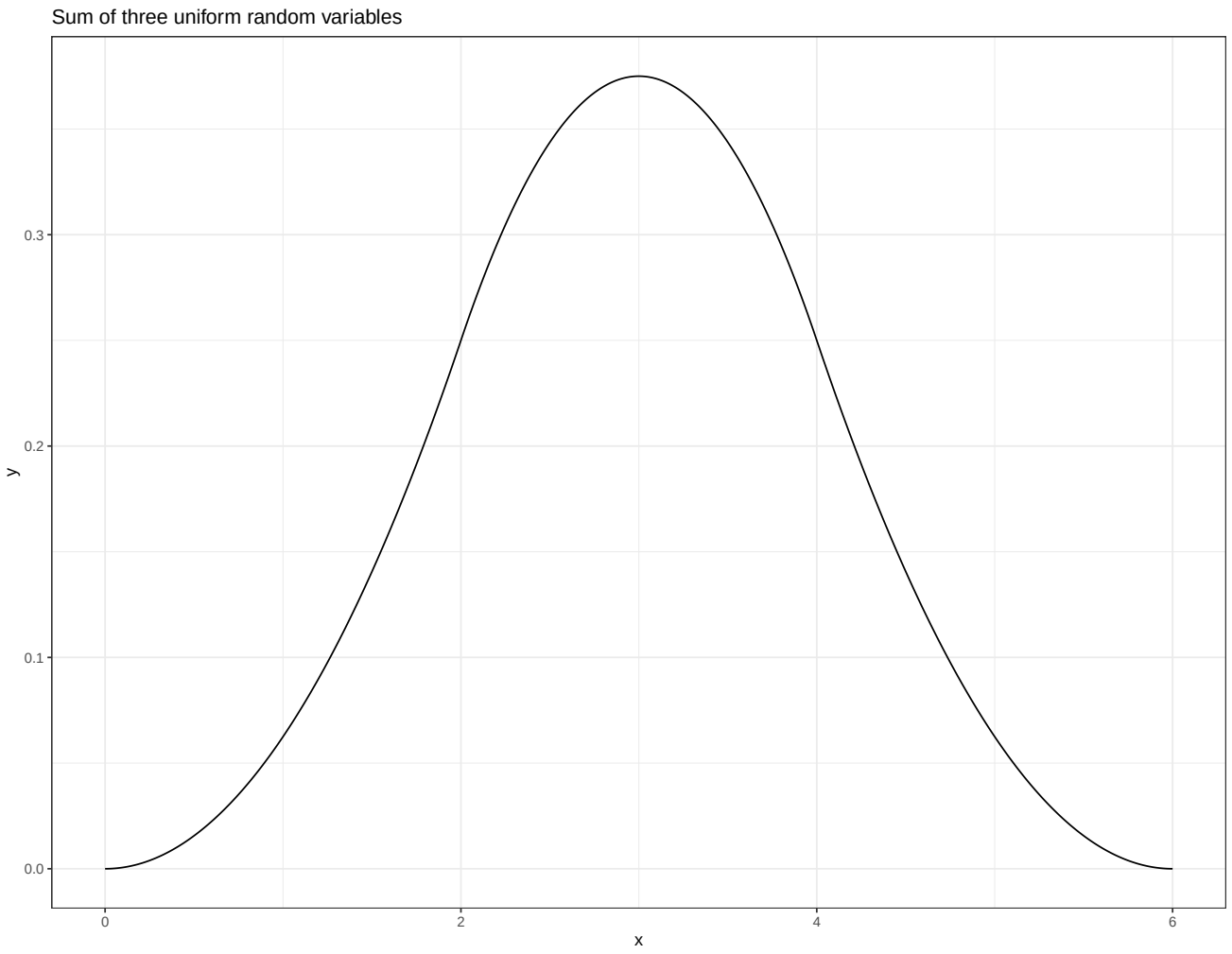
Somme de variables aléatoires continues (4/4)

Soit $X, Y \in [0, 2]$.



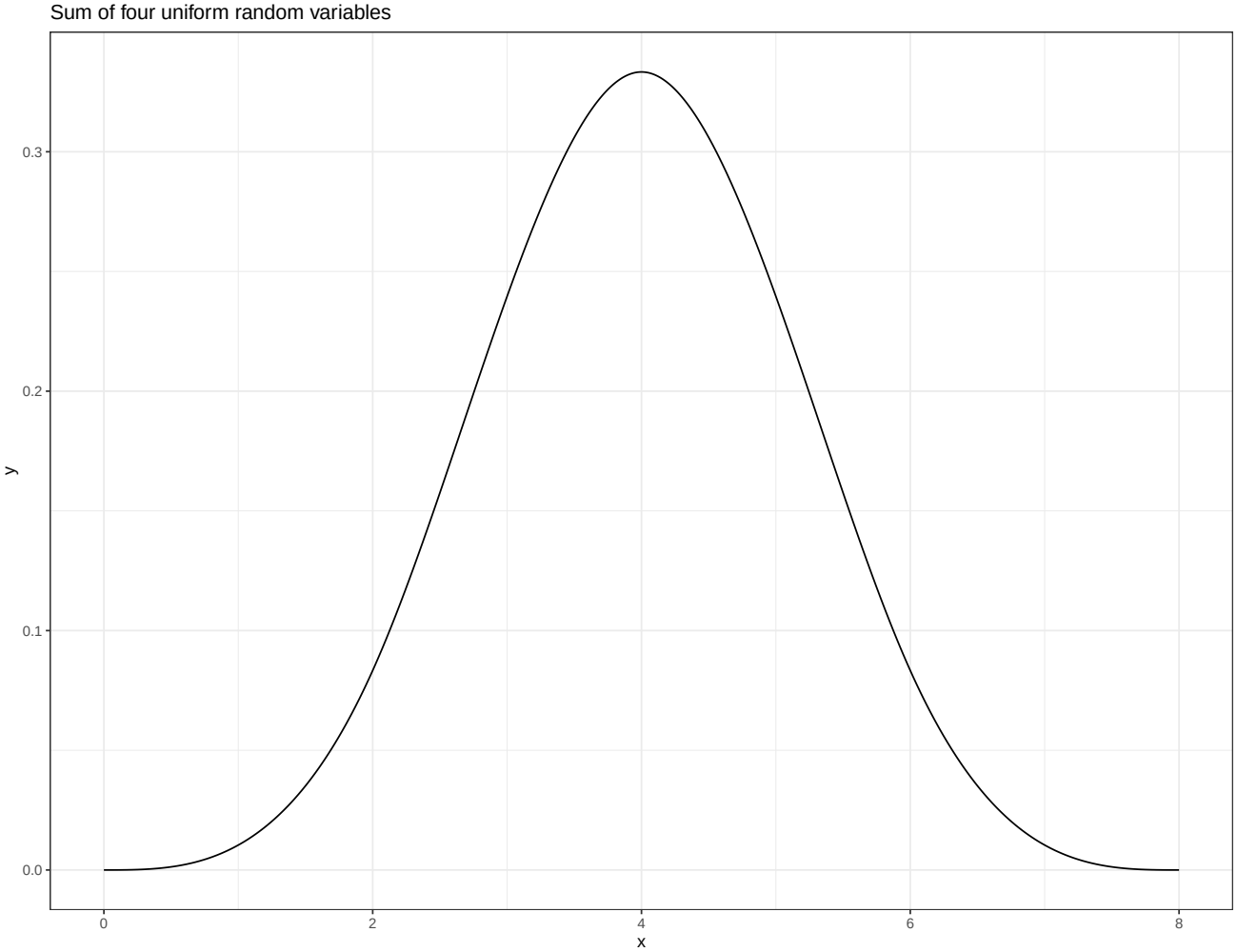
Somme de variables aléatoires continues (4/4)

Soit $X, Y, Z \in [0, 2]$.



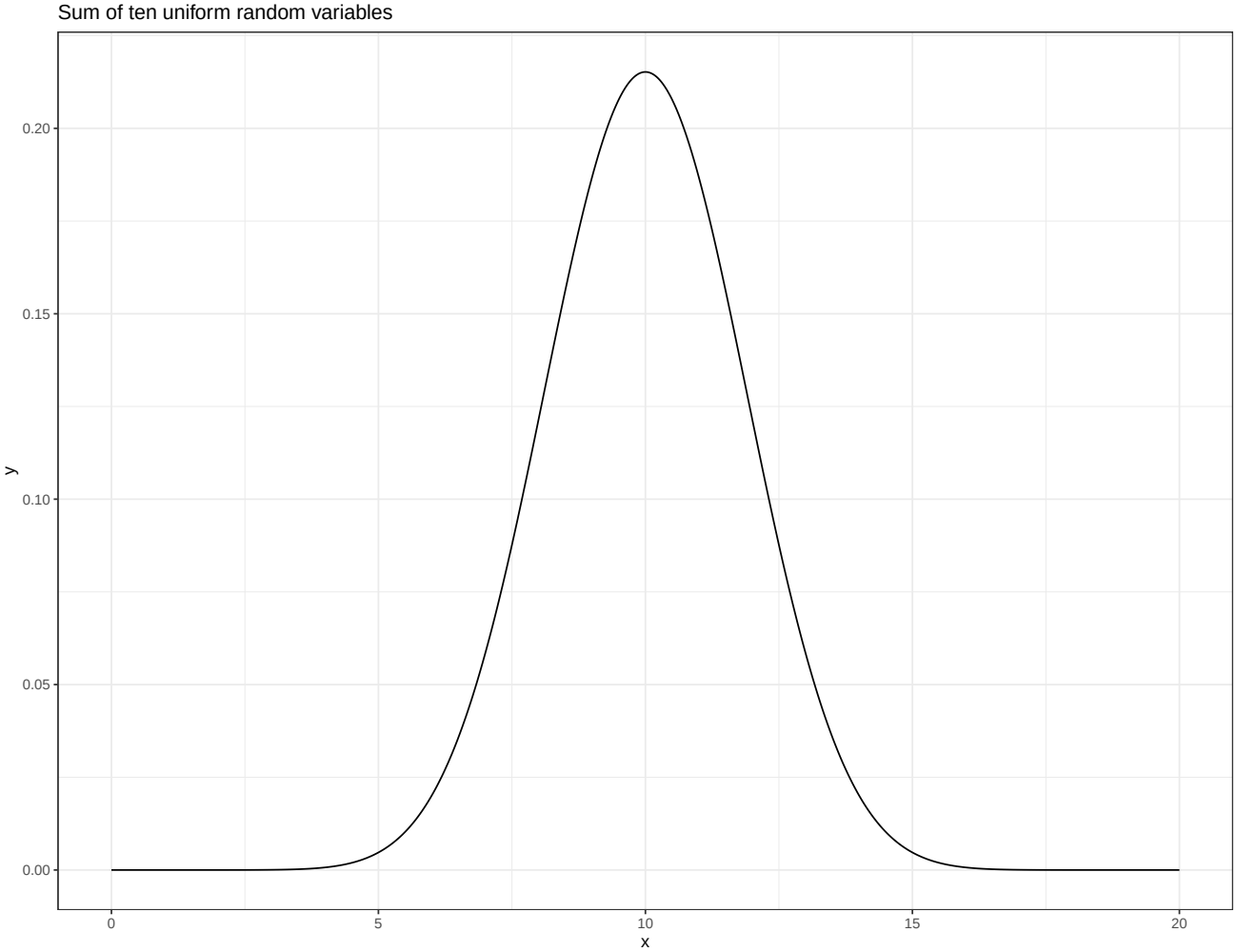
Somme de variables aléatoires continues (4/4)

Soit $X, Y, Z, \ldots \in [0, 2]$.



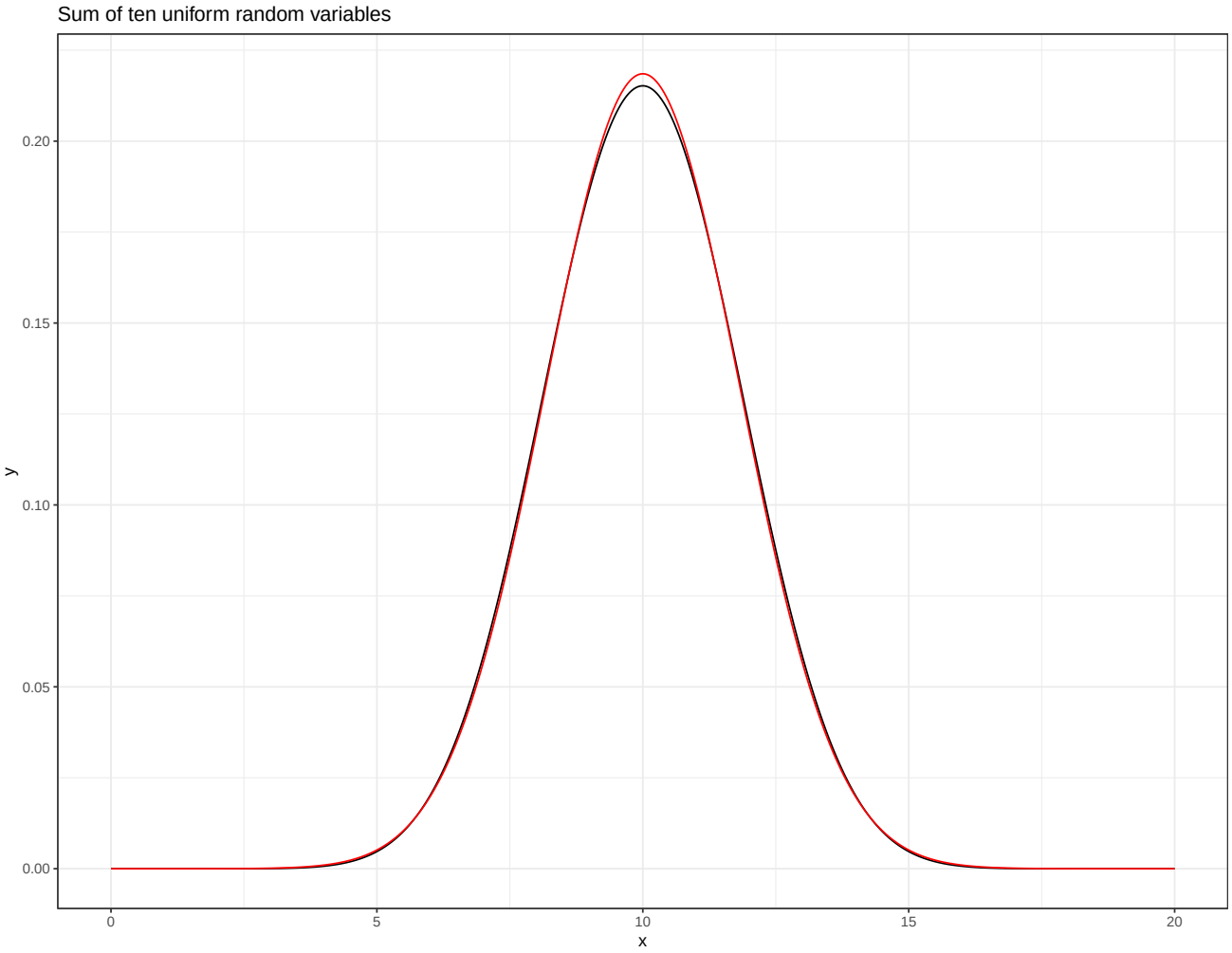
Somme de variables aléatoires continues (4/4)

Soit $X, Y, Z, \ldots \in [0, 2]$.



Somme de variables aléatoires continues (4/4)

Soit $X, Y, Z, ... \in [0, 2]$ et
$V \sim \mathcal{N}(\mu = 10(2+0)/2, \sigma^2 = \frac{1}{12}(2-0)^2 10)$.



Valeur-p 1/3

Définition

Soit H_0 l'hypothèse nulle et x la donnée observée, la valeur-p définit la probabilité d'observer sous l'hypothèse nulle une donnée au moins aussi extrême que x.

- Il ne s'agit pas de la probabilité de l'hypothèse H_0.

Valeur-p 2/3

Exemple

La pièce est-elle truquée ?

Valeur-p 2/3

Exemple

La pièce est-elle truquée ? La valeur-p est la probabilité d'obtenir d'obtenir nos observations ou des observations encore plus extrêmees dans le cadre d'une hypothèse nulle H_0 :
$p = \mathbb{P}(X > |z| \cup X < -|z| \big| H_0)$ *(⚠$p \neq \mathbb{P}(H_0 \big| z)$).*

Valeur-p 2/3

Exemple

La pièce est-elle truquée ? La valeur-p est la probabilité d'obtenir d'obtenir nos observations ou des observations encore plus extrêmees dans le cadre d'une hypothèse nulle H_0 :
$p = \mathbb{P}(X > |z| \cup X < -|z| \big| H_0)$ *(⚠$p \neq \mathbb{P}(H_0 \big| z)$).*

1. *on choisit un seuil de significativité et si la valeur-p est inférieure à ce seuil, on conclut qu'il est "improbable" que les données suivent l'hypothèse nulle.*
2. *on calcule p et on conclut.*

Valeur-p 2/3

Exemple

La pièce est-elle truquée ? La valeur-p est la probabilité d'obtenir d'obtenir nos observations ou des observations encore plus extrêmees dans le cadre d'une hypothèse nulle H_0 :
$p = \mathbb{P}(X > |z| \cup X < -|z| \big| H_0)$ *(⚠$p \neq \mathbb{P}(H_0 \big| z)$).*

1. *on choisit un seuil de significativité et si la valeur-p est inférieure à ce seuil, on conclut qu'il est "improbable" que les données suivent l'hypothèse nulle.*
2. *on calcule p et on conclut.*

Exercice : Avec un seuil fixé à 5% et l'hypothèse nulle est "la pièce est bien équilibrée" et l'hypothèse H_1 $p > \frac{1}{2}$. Pouvons nous rejeter H_0 avec 4 piles et 0 face ?

Valeur-p 2/3

Exemple

La pièce est-elle truquée ? La valeur-p est la probabilité d'obtenir d'obtenir nos observations ou des observations encore plus extrêmees dans le cadre d'une hypothèse nulle H_0 :
$p = \mathbb{P}(X > |z| \cup X < -|z| \big| H_0)$ *(⚠$p \neq \mathbb{P}(H_0 \big| z)$).*

1. *on choisit un seuil de significativité et si la valeur-p est inférieure à ce seuil, on conclut qu'il est "improbable" que les données suivent l'hypothèse nulle.*
2. *on calcule p et on conclut.*

Exercice : Avec un seuil fixé à 5% et l'hypothèse nulle est "la pièce est bien équilibrée" et l'hypothèse H_1 $p > \frac{1}{2}$. Pouvons nous rejeter H_0 avec 4 piles et 0 face ?
Non... $0.5^4 = 0.0625 > 5\%$.

Valeur-p 3/3

Exemple

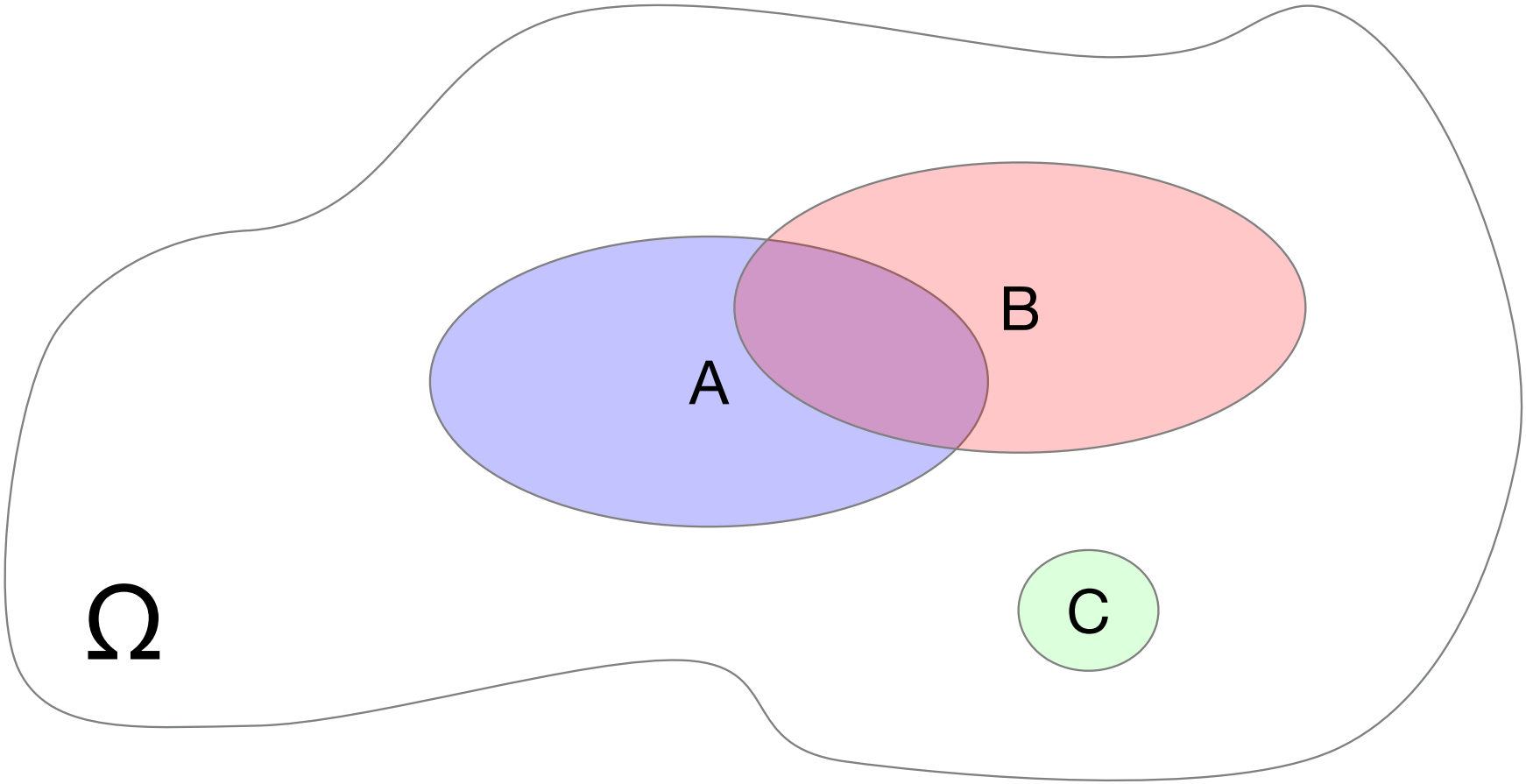
- *Le test du χ^2 a notamment pour objectif de tester l'indépendance de deux variables aléatoires.*
- *La statistique T du test suit une loi du χ^2 sous l'hypothèse H_0.*
- *Le nombre de degré de liberté et la valeur-p souhaitée nous permettent de trouver notre seuil dans la table du χ^2.*
- *la statistique T est calculée sur les données observées et est comparée au seuil.*

Indépendance de variables aléatoires

- Deux variables aléatoires X, Y à valeur dans E sont dites indépendantes si la réalisation de l'une d'entre elles n'apporte pas d'information quant à la seconde :

$$P(X|Y) = P(X) \Leftrightarrow P(X \cap Y) = P(X)P(Y)$$

Exemple



Loi marginale

Définition

Étant donné un vecteur aléatoire $(X_1, ..., X_n)$ *de densité jointe* $f_{(X_1,...,X_n)}(x_1, ..., x_n)$*, la loi de* X_i *uniquement est appelée sa loi marginale. En particulier, on peut noter :*

$$f_{X_i}(x) = \int ... \int f_{(X_1,...,X_n)}(x_1, ..., x_n) dX_1 ... dX_{i-1} dX_{i+1} ... dX_n$$

Exemple

- Soit X et Y deux variables aléatoires telles que :

	Y=0	Y=1
X=0	0.25	0.35
X=1	0.3	0.1

Exemple

- Soit X et Y deux variables aléatoires telles que :

	Y=0	Y=1
X=0	0.25	0.35
X=1	0.3	0.1

- $P(X = 0) =$

Exemple

- Soit X et Y deux variables aléatoires telles que :

	Y=0	Y=1
X=0	0.25	0.35
X=1	0.3	0.1

- $P(X = 0) = \sum_i P(X = 0 \cap Y = i)$

Exemple

- Soit X et Y deux variables aléatoires telles que :

	Y=0	Y=1
X=0	0.25	0.35
X=1	0.3	0.1

- $P(X = 0) = \sum_i P(X = 0 \cap Y = i) = 0.6$

Exemple

- Soit un vecteur $V = (X, Y)$ de densité (par rapport à la mesure de ....) :

$$f_{(X,Y)}(v) = \frac{1}{2\pi}\exp\big(-\frac{1}{2}({}^t vv)\big),$$

soit une loi normale sur $\mathbb{R}^2$ de moyenne 0 et de matrice de variance-covariance identitée.

Exemple

- Soit un vecteur $V = (X, Y)$ de densité (par rapport à la mesure de ....) :

$$f_{(X,Y)}(v) = \frac{1}{2\pi}\exp\big(-\frac{1}{2}({}^t vv)\big),$$

soit une loi normale sur $\mathbb{R}^2$ de moyenne 0 et de matrice de variance-covariance identitée. La loi marginale de X est :

Exemple

- Soit un vecteur $V = (X, Y)$ de densité (par rapport à la mesure de ....) :

$$f_{(X,Y)}(v) = \frac{1}{2\pi}\exp\big(-\frac{1}{2}({}^t vv)\big),$$

soit une loi normale sur $\mathbb{R}^2$ de moyenne 0 et de matrice de variance-covariance identitée. La loi marginale de X est :

$$f_X(x) = \int f_{(X,Y)}((x,y))dy$$

Exemple

- Soit un vecteur $V = (X, Y)$ de densité (par rapport à la mesure de ....) :

$$f_{(X,Y)}(v) = \frac{1}{2\pi}\exp\big(-\frac{1}{2}({}^t vv)\big),$$

soit une loi normale sur $\mathbb{R}^2$ de moyenne 0 et de matrice de variance-covariance identitée. La loi marginale de X est :

$$\begin{aligned} f_X(x) &= \int f_{(X,Y)}((x,y))dy \\ &= \frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}x^2)\left[\int \frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}y^2)dy\right] \end{aligned}$$

Exemple

- Soit un vecteur $V = (X, Y)$ de densité (par rapport à la mesure de ....) :

$$f_{(X,Y)}(v) = \frac{1}{2\pi}\exp\big(-\frac{1}{2}({}^t vv)\big),$$

soit une loi normale sur $\mathbb{R}^2$ de moyenne 0 et de matrice de variance-covariance identitée. La loi marginale de X est :

$$\begin{aligned} f_X(x) &= \int f_{(X,Y)}((x,y))dy \\ &= \frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}x^2)\left[\int \frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}y^2)dy\right] \\ &= \frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}x^2) = \mathcal{N}(0,1) \end{aligned}$$

Convergence en probabilité

Convergence en probabilité

Définition

Soit $(X_i)_{i\geq 1}$ *une suite de variables aléatoires dans* $\mathbb{R}^d$, $d \in \mathbb{N}^*$ *définies sur le même espace probabilisé. Celle-ci converge en probabilité vers* X *si*

$$\forall \epsilon > 0, P(||X_n - X|| > \epsilon) \to 0, n \to \infty,$$

et on note cela $X_n \xrightarrow{P} X$.

Convergence en loi

Convergence en loi

Définition

Soit $(X_i)_{i\geq 1}$ une suite de variables aléatoires dans $\mathbb{R}^d$, $d \in \mathbb{N}^*$, on dit que la suite X_n converge en loi vers une variable aléatoire X si pour toute fonction $f : \mathbb{R}^d \to \mathbb{R}$ continue bornée on a

$$\mathbb{E}\big[f(X_n)\big] \to \mathbb{E}\big[f(X)\big], n \to \infty,$$

et on note cela $X_n \xrightarrow{\mathcal{L}} X$.

Loi des grands nombres

Loi des grands nombres

Proposition

Soit $(X_1, ..., X_n)$ *une suite de variable aléatoire i.i.d. à valeurs dans* $\mathbb{R}^d$, $d \in \mathbb{N}^*$ *avec* $\mathbb{E}\big[\|X_i\|\big] < \infty$. *On obtient alors :*

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{P} \mathbb{E}\big[X\big]$$

Théorème central limite

Théorème central limite

Proposition

Soit $(X_1, ..., X_n)$ *une suite de variable aléatoire i.i.d. à valeurs dans* $\mathbb{R}^d$, $d \in \mathbb{N}^*$ *avec* $\mathbb{E}\left[\|X_i\|^2\right] < \infty$. *On obtient alors :*

$$\sqrt{n}(\overline{X} - \mu) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma)$$

avec $\mu = \mathbb{E}\left[X\right]$ *et* $\Sigma = \mathbb{E}\Big[(X - \mathbb{E}[X])^t(X - \mathbb{E}[X])\Big]$, *soit une loi normale centrée sur* $\mathbb{R}^d$ *de variance-covariance* Σ.

Intervalle de confiance

Pouvons-nous faire confiance à notre estimateur ?

Intervalle de confiance

Pouvons-nous faire confiance à notre estimateur ?
Le théorème central limite nous permet de calculer un intervalle de confiance :

$$I_c = \left[\overline{x} - t_\alpha \frac{\sigma}{\sqrt{n}}, \overline{x} + t_\alpha \frac{\sigma}{\sqrt{n}}\right]$$

où $\overline{x}$ est la moyenne de notre échantillon, σ son écart type observé et n sa taille.

Intervalle de confiance

Pouvons-nous faire confiance à notre estimateur ?
Le théorème central limite nous permet de calculer un intervalle de confiance :

$$I_c = \left[\overline{x} - t_\alpha \frac{\sigma}{\sqrt{n}}, \overline{x} + t_\alpha \frac{\sigma}{\sqrt{n}}\right]$$

où $\overline{x}$ est la moyenne de notre échantillon, σ son écart type observé et n sa taille.
⚠L'intervalle de confiance α ne signifie pas que nous avons α chance que le "vrai" paramètre soit situé dans le segment !

Plan

1 Rappels

2 L'approche Bayésienne
- Introduction
- Conjugaison
- Posterior prédictive
- Régression linéaire Bayésienne

3 Les méthodes MCMC

4 Les méthodes approchées

Introduction

Pierre-Simon de Laplace
(1749-1827)

Thomas Bayes (1702-1761)

- On ne considère plus qu'il existe une "vraie" valeur de θ,
- mais que θ est lui même une variable aléatoire,
- l'objet d'intérêt est la loi de probabilité de θ.

Construction du modèle

- De manière plus formelle, on souhaite obtenir une loi sur θ après avoir observé des données $\mathcal{D}$:

$$\pi(\theta|\mathcal{D})$$

Construction du modèle

- De manière plus formelle, on souhaite obtenir une loi sur θ après avoir observé des données $\mathcal{D}$:

$$\pi(\theta|\mathcal{D})$$

- On retrouve quelque chose de la forme :

$$P(A|B)$$

Construction du modèle

- De manière plus formelle, on souhaite obtenir une loi sur θ après avoir observé des données $\mathcal{D}$:

$$\pi(\theta|\mathcal{D})$$

- On retrouve quelque chose de la forme :

$$P(A|B)P(B) =$$

Construction du modèle

- De manière plus formelle, on souhaite obtenir une loi sur θ après avoir observé des données $\mathcal{D}$:

$$\pi(\theta|\mathcal{D})$$

- On retrouve quelque chose de la forme :

$$P(A|B)P(B) = P(A, B) =$$

Construction du modèle

- De manière plus formelle, on souhaite obtenir une loi sur θ après avoir observé des données $\mathcal{D}$:

$$\pi(\theta|\mathcal{D})$$

- On retrouve quelque chose de la forme :

$$P(A|B)P(B) = P(A,B) = P(B|A)P(A)$$

Construction du modèle

- De manière plus formelle, on souhaite obtenir une loi sur θ après avoir observé des données $\mathcal{D}$:

$$\pi(\theta|\mathcal{D})$$

- On retrouve quelque chose de la forme :

$$P(A|B)P(B) = P(A,B) = P(B|A)P(A)$$

- En appliquant la même logique à notre modèle :

$$\pi(\theta|\mathcal{D}) =$$

Construction du modèle

- De manière plus formelle, on souhaite obtenir une loi sur θ après avoir observé des données $\mathcal{D}$:

$$\pi(\theta|\mathcal{D})$$

- On retrouve quelque chose de la forme :

$$P(A|B)P(B) = P(A,B) = P(B|A)P(A)$$

- En appliquant la même logique à notre modèle :

$$\pi(\theta|\mathcal{D}) = \mathcal{L}_\theta(\mathcal{D})$$

Construction du modèle

- De manière plus formelle, on souhaite obtenir une loi sur θ après avoir observé des données $\mathcal{D}$:

$$\pi(\theta|\mathcal{D})$$

- On retrouve quelque chose de la forme :

$$P(A|B)P(B) = P(A,B) = P(B|A)P(A)$$

- En appliquant la même logique à notre modèle :

$$\pi(\theta|\mathcal{D}) = \frac{\mathcal{L}_\theta(\mathcal{D})\pi(\theta)}{}$$

Construction du modèle

- De manière plus formelle, on souhaite obtenir une loi sur θ après avoir observé des données $\mathcal{D}$:

$$\pi(\theta|\mathcal{D})$$

- On retrouve quelque chose de la forme :

$$P(A|B)P(B) = P(A,B) = P(B|A)P(A)$$

- En appliquant la même logique à notre modèle :

$$\pi(\theta|\mathcal{D}) = \frac{\mathcal{L}_\theta(\mathcal{D})\pi(\theta)}{p(\mathcal{D})}$$

Construction du modèle

- De manière plus formelle, on souhaite obtenir une loi sur θ après avoir observé des données $\mathcal{D}$:

$$\pi(\theta|\mathcal{D})$$

- On retrouve quelque chose de la forme :

$$P(A|B)P(B) = P(A,B) = P(B|A)P(A)$$

- En appliquant la même logique à notre modèle :

$$\pi(\theta|\mathcal{D}) = \frac{\mathcal{L}_\theta(\mathcal{D})\pi(\theta)}{p(\mathcal{D})}$$

- On retrouve la vraisemblance des données sachant θ ainsi que $\pi(\theta)$ et $p(\mathcal{D})$.

Construction du modèle

- $\pi(\theta)$ est la densité associée à la loi *a priori* sur θ :

$$d\Pi = \pi d\nu$$

Construction du modèle

- $\pi(\theta)$ est la densité associée à la loi *a priori* sur θ :

$$d\Pi = \pi d\nu$$

- Il s'agit d'une loi sur nos paramètres qui doit décrire l'état actuel des connaissances sur ces derniers.

Construction du modèle

- $\pi(\theta)$ est la densité associée à la loi *a priori* sur θ :

$$d\Pi = \pi d\nu$$

- Il s'agit d'une loi sur nos paramètres qui doit décrire l'état actuel des connaissances sur ces derniers.
- *L'a priori* est une brique nécessaire à l'inférence Bayésienne.
- $p(\mathcal{D})$ semble inconnu, mais peut se retrouver par désintégration :

Construction du modèle

- $\pi(\theta)$ est la densité associée à la loi *a priori* sur θ :

$$d\Pi = \pi d\nu$$

- Il s'agit d'une loi sur nos paramètres qui doit décrire l'état actuel des connaissances sur ces derniers.
- *L'a priori* est une brique nécessaire à l'inférence Bayésienne.
- $p(\mathcal{D})$ semble inconnu, mais peut se retrouver par désintégration :

$$p(\mathcal{D}) = \int_{\Theta} \mathcal{L}_{\theta}(\mathcal{D})\pi(\theta)d\theta$$

Construction du modèle

- $\pi(\theta)$ est la densité associée à la loi *a priori* sur θ :

$$d\Pi = \pi d\nu$$

- Il s'agit d'une loi sur nos paramètres qui doit décrire l'état actuel des connaissances sur ces derniers.
- *L'a priori* est une brique nécessaire à l'inférence Bayésienne.
- $p(\mathcal{D})$ semble inconnu, mais peut se retrouver par désintégration :

$$p(\mathcal{D}) = \int_{\Theta} \mathcal{L}_{\theta}(\mathcal{D})\pi(\theta)d\theta$$

- En inférence Bayésienne, on note la vraisemblance des données $\mathcal{L}(\mathcal{D}|\theta)$ et la vraisemblance jointe $\mathcal{L}(\mathcal{D}, \theta) = \mathcal{L}(\mathcal{D}|\theta)\pi(\theta)$.

Exemple

- Reprenons l'exemple de pile ou face. La loi de X choisie était la Bernoulli :

$$P_\theta(X = x) = \theta^x(1-\theta)^{1-x}, x \in \{0, 1\}$$

Exemple

- Reprenons l'exemple de pile ou face. La loi de X choisie était la Bernoulli :

$$P_\theta(X = x) = \theta^x (1-\theta)^{1-x}, x \in \{0, 1\}$$

- une loi *a priori* possible est la loi beta, de densité :

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

  - où $\Gamma(\cdot)$ est la fonction gamma et α et β sont appelés hyper paramètres.

Exemple

- Reprenons l'exemple de pile ou face. La loi de X choisie était la Bernoulli :

$$P_\theta(X = x) = \theta^x (1-\theta)^{1-x}, x \in \{0, 1\}$$

- une loi *a priori* possible est la loi beta, de densité :

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

  - où $\Gamma(\cdot)$ est la fonction gamma et α et β sont appelés hyper paramètres.
- La loi $\mathcal{L}((X, \theta))$ obtenue est donc :

$$\mathcal{L}((X, \theta)) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)+\Gamma(\beta)} \theta^{\alpha+x-1}(1-\theta)^{\beta-x}$$

La loi de Bayes

Définition

La loi a posteriori est la loi conditionnelle de θ sachant X, notée $\Pi(\theta|X)$. Elle est à densité $f_{X=x}(\theta)$ par rapport à la mesure ν et s'obtient en appliquant la formule de Bayes :

$$f_{X=x}(\theta) = \frac{f_\theta(x)\pi(\theta)}{\int f_\theta(x) d\Pi(\theta)}$$

La loi de Bayes

Définition

La loi a posteriori est la loi conditionnelle de θ sachant X, notée $\Pi(\theta|X)$. Elle est à densité $f_{X=x}(\theta)$ par rapport à la mesure ν et s'obtient en appliquant la formule de Bayes :

$$f_{X=x}(\theta) = \frac{f_\theta(x)\pi(\theta)}{\int f_\theta(x)d\Pi(\theta)} = \frac{f_\theta(x)\pi(\theta)}{\int f_\theta(x)\pi(\theta)d\nu(\theta)}$$

La loi de Bayes

Définition

La loi a posteriori est la loi conditionnelle de θ sachant X, notée $\Pi(\theta|X)$. Elle est à densité $f_{X=x}(\theta)$ par rapport à la mesure ν et s'obtient en appliquant la formule de Bayes :

$$f_{X=x}(\theta) = \frac{f_\theta(x)\pi(\theta)}{\int f_\theta(x)d\Pi(\theta)} = \frac{f_\theta(x)\pi(\theta)}{\int f_\theta(x)\pi(\theta)d\nu(\theta)}$$

- $\int f_\theta(x)\pi(\theta)d\nu(\theta)$ est la loi marginale de X. Elle est constante par rapport à θ et on reformule la formule de Bayes comme :

$$f_{X=x}(\theta) \propto f_\theta(x)\pi(\theta) = \mathcal{L}((x,\theta)),$$

La loi de Bayes

Définition

La loi a posteriori est la loi conditionnelle de θ sachant X, notée $\Pi(\theta|X)$. Elle est à densité $f_{X=x}(\theta)$ par rapport à la mesure ν et s'obtient en appliquant la formule de Bayes :

$$f_{X=x}(\theta) = \frac{f_\theta(x)\pi(\theta)}{\int f_\theta(x)d\Pi(\theta)} = \frac{f_\theta(x)\pi(\theta)}{\int f_\theta(x)\pi(\theta)d\nu(\theta)}$$

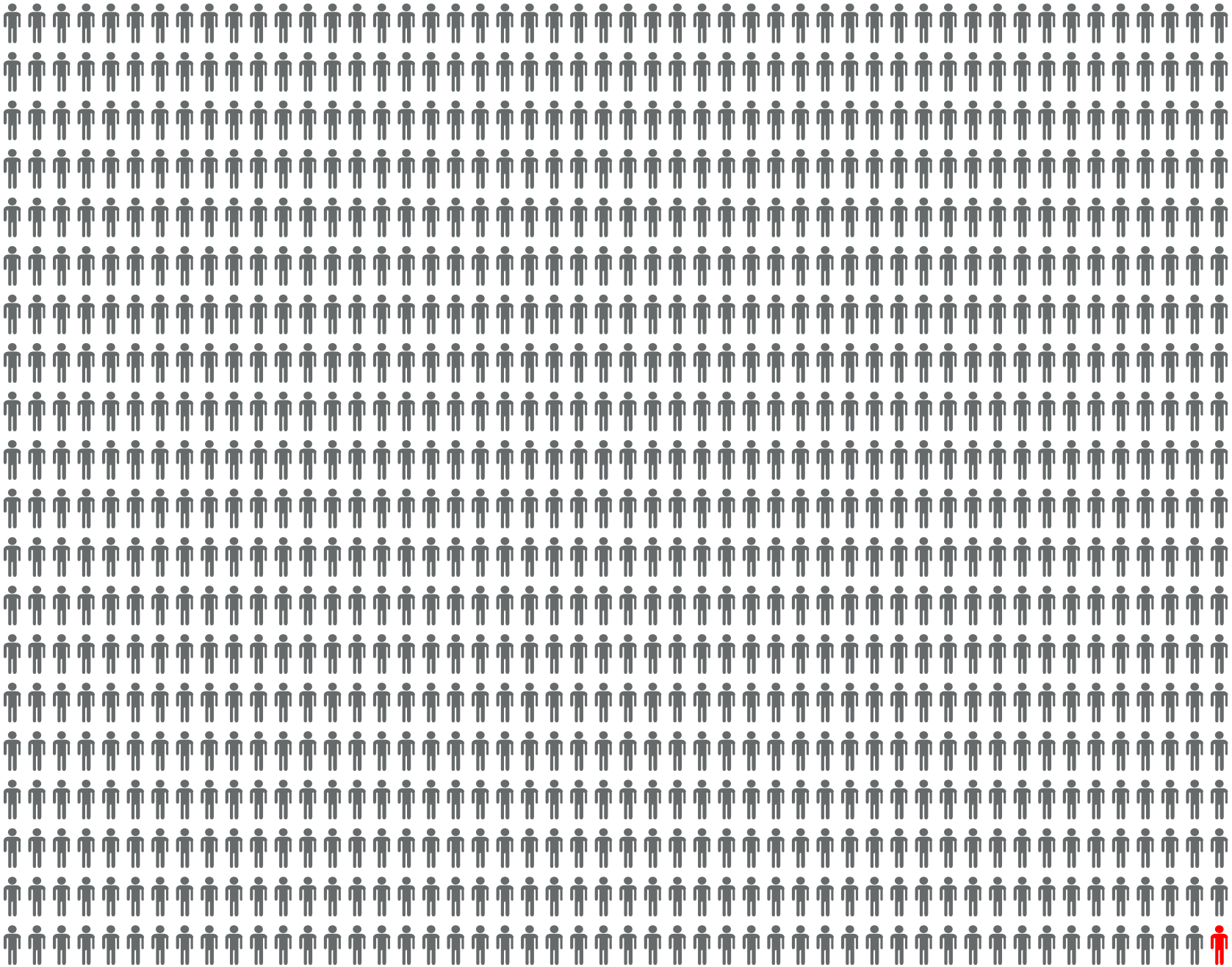
- $\int f_\theta(x)\pi(\theta)d\nu(\theta)$ est la loi marginale de X. Elle est constante par rapport à θ et on reformule la formule de Bayes comme :

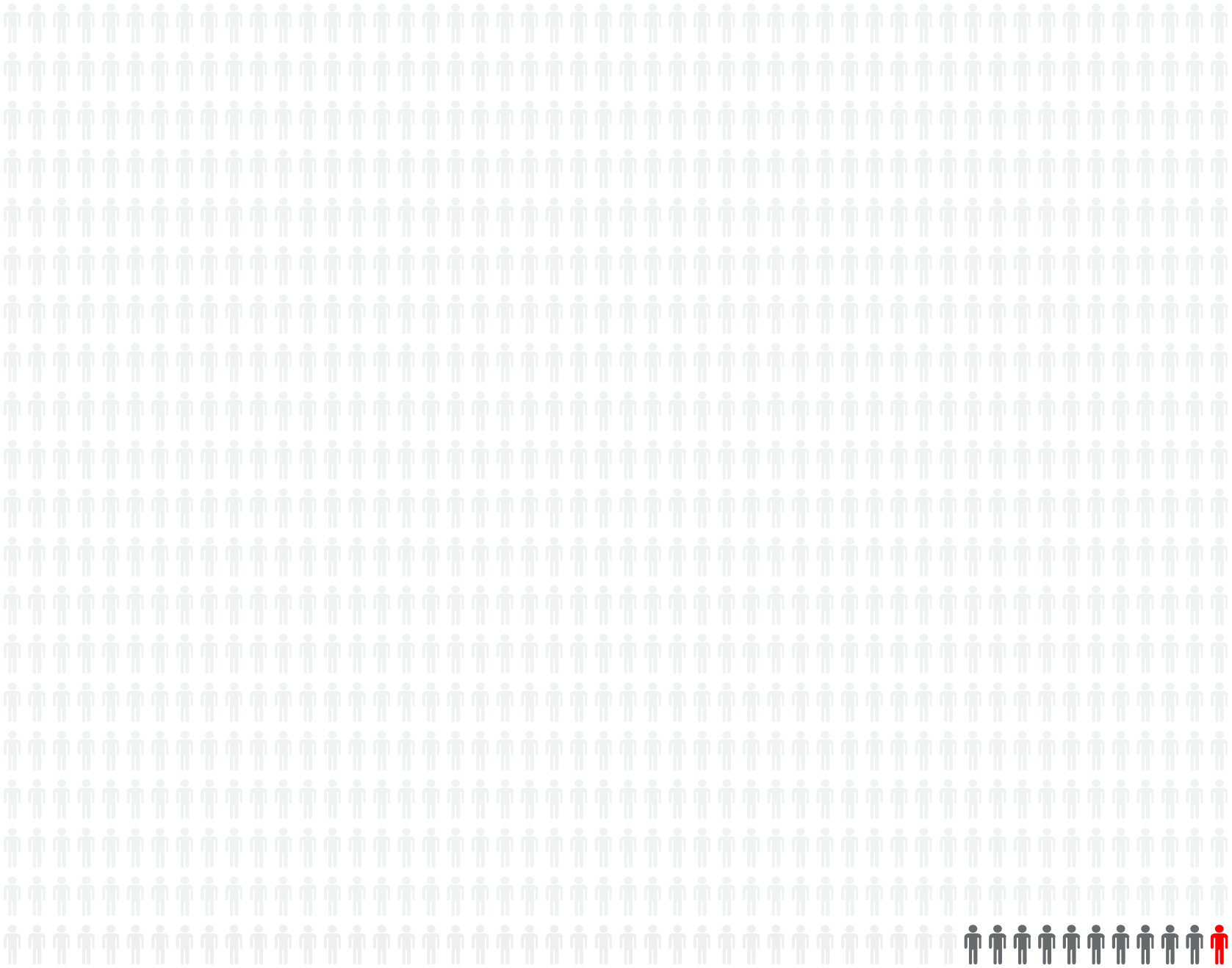
$$f_{X=x}(\theta) \propto f_\theta(x)\pi(\theta) = \mathcal{L}((x,\theta)),$$

- il s'agit, en quelque sorte, d'affiner notre mesure sur $\{\Theta, \mathcal{F}\}$ après avoir observé des données.

Exemple naïf

- Soit une maladie *xyz* qui touche une personne sur mille.
- On propose un dépistage via un test T possédants les caractéristiques suivantes :
  - $\mathbb{P}(T = 1|M = 1) = 99\%$
  - $\mathbb{P}(T = 0|M = 1) = 1\%$
  - $\mathbb{P}(T = 0|M = 0) = 99\%$
  - $\mathbb{P}(T = 1|M = 0) = 1\%$





La solution

- Formalisons cet exemple. Soit $\delta[M = T]$ une fonction qui vaut 1 lorsque la condition est vérifiée et 0 sinon, i.e. le test se trompe ou le test ne se trompe pas.

La solution

- Formalisons cet exemple. Soit $\delta[M = T]$ une fonction qui vaut 1 lorsque la condition est vérifiée et 0 sinon, i.e. le test se trompe ou le test ne se trompe pas.
- On construit la vraisemblance des données sachant la maladie :

$$\mathbb{P}(T|M) = 0.99^{\delta[M=T]} 0.01^{\delta[T \neq M]}$$

La solution

- Formalisons cet exemple. Soit $\delta[M = T]$ une fonction qui vaut 1 lorsque la condition est vérifiée et 0 sinon, i.e. le test se trompe ou le test ne se trompe pas.
- On construit la vraisemblance des données sachant la maladie :

$$\mathbb{P}(T|M) = 0.99^{\delta[M=T]} 0.01^{\delta[T \neq M]}$$

- On construit notre *a priori* :

$$\mathbb{P}(M) = 0.001^{\delta[M]} 0.999^{\delta[M-1]}$$

La solution

- Formalisons cet exemple. Soit $\delta[M = T]$ une fonction qui vaut 1 lorsque la condition est vérifiée et 0 sinon, i.e. le test se trompe ou le test ne se trompe pas.
- On construit la vraisemblance des données sachant la maladie :

$$\mathbb{P}(T|M) = 0.99^{\delta[M=T]} 0.01^{\delta[T \neq M]}$$

- On construit notre *a priori* :

$$\mathbb{P}(M) = 0.001^{\delta[M]} 0.999^{\delta[M-1]}$$

- On construit "l'évidence" :

$$\mathbb{P}(T) = \mathbb{P}(T|M = 1)\mathbb{P}(M = 1) + \mathbb{P}(T|M = 0)\mathbb{P}(M = 0)$$

La solution

- Formalisons cet exemple. Soit $\delta[M = T]$ une fonction qui vaut 1 lorsque la condition est vérifiée et 0 sinon, i.e. le test se trompe ou le test ne se trompe pas.
- On construit la vraisemblance des données sachant la maladie :

$$\mathbb{P}(T|M) = 0.99^{\delta[M=T]} 0.01^{\delta[T \neq M]}$$

- On construit notre *a priori* :

$$\mathbb{P}(M) = 0.001^{\delta[M]} 0.999^{\delta[M-1]}$$

- On construit "l'évidence" :

$$\mathbb{P}(T) = \mathbb{P}(T|M = 1)\mathbb{P}(M = 1) + \mathbb{P}(T|M = 0)\mathbb{P}(M = 0)$$

- Enfin, on construit notre posterior :

$$\mathbb{P}(M|T) = \mathbb{P}(T|M)\mathbb{P}(M)/\mathbb{P}(T)$$
$$\mathbb{P}(M = 1|T = 1) = 0.99 \times 0.001/(0.99 \times 0.001 + 0.01 \times 0.999) \approx 0.09$$
$$\mathbb{P}(M = 0|T = 1) \approx 0.91$$

La loi de Bayes

Définition

Avec plusieurs variables observées, la formule de Bayes donne :

$$f_{X_1,\ldots,X_N}(\theta) = \frac{\prod_{i=1}^{N} f_\theta(x_i)\pi(\theta)}{\int \prod_{i=1}^{N} f_\theta(x_i)\pi(\theta)d\nu(\theta)}$$

Conjugaison

Définition

Une famille $\mathcal{F}$ de lois a priori (e.g. beta, gamma, bernoulli) est dite conjuguée à la famille choisie pour le modèle $\mathcal{M} = \{P_\theta, \theta \in \Theta\}$ si la loi a posteriori $\Pi(\cdot|X)$ appartient également à la famille $\mathcal{F}$.

Conjugaison

Loi *a priori*	Vraisemblance (des données)	Loi *a posteriori*
Beta $\mathcal{B}(\alpha, \beta)$	Bernoulli/binomiale $\mathcal{B}e(p)$	beta $\mathcal{B}(\alpha + n - x, \beta + x)$
Dirichlet $\mathcal{D}(\alpha_i)$	Multinomiale $\mathcal{M}(p_i)$	Dirichlet $\mathcal{D}(\alpha_i + n_i)$
Normale $\mathcal{N}(\mu_0, \sigma_0^2)$	Normale (variance fixe) $\mathcal{N}(\mu, \sigma^2)$	Normale $\mathcal{N}(.....)$
Inv gamma	Normale avec moyenne connue	Inv gamma
Normale-inv-Wishart	Normale multivariée	Normale-inv-Wishart
...	...	...

Exemple

- Reprenons l'exemple de pile ou face. La famille de lois du modèle est la Bernoulli, et pour *prior*, la loi beta.

Exemple

- Reprenons l'exemple de pile ou face. La famille de lois du modèle est la Bernoulli, et pour *prior*, la loi beta.
- Appliquons la formule de Bayes :

$$f_{X=x}(\theta) \propto \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)+\Gamma(\beta)}\theta^{\alpha+x-1}(1-\theta)^{\beta-x}$$

Exemple

- Reprenons l'exemple de pile ou face. La famille de lois du modèle est la Bernoulli, et pour *prior*, la loi beta.
- Appliquons la formule de Bayes :

$$\begin{aligned} f_{X=x}(\theta) &\propto \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)+\Gamma(\beta)}\theta^{\alpha+x-1}(1-\theta)^{\beta-x} \\ &\propto \theta^{\alpha+x-1}(1-\theta)^{\beta-x} \end{aligned}$$

Exemple

- Reprenons l'exemple de pile ou face. La famille de lois du modèle est la Bernoulli, et pour *prior*, la loi beta.
- Appliquons la formule de Bayes :

$$\begin{aligned} f_{X=x}(\theta) &\propto \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)+\Gamma(\beta)}\theta^{\alpha+x-1}(1-\theta)^{\beta-x} \\ &\propto \theta^{\alpha+x-1}(1-\theta)^{\beta-x} \end{aligned}$$

- on observe facilement que $f_X(\theta)$ est proportionnelle à une loi beta de paramètres $\alpha + x$ et $\beta + 1 - x$: La loi beta est conjuguée à la loi bernoulli.

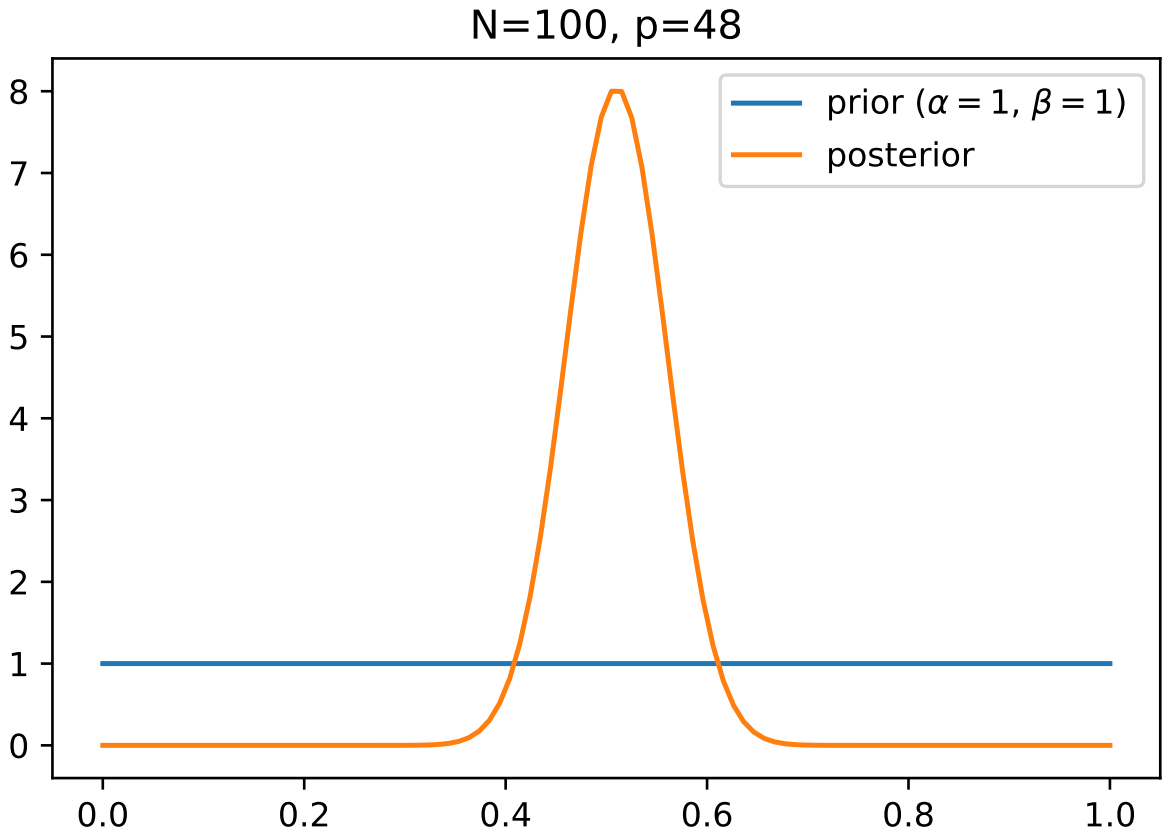
Exemple

- Reprenons l'exemple de pile ou face. La famille de lois du modèle est la Bernoulli, et pour *prior*, la loi beta.
- Appliquons la formule de Bayes :

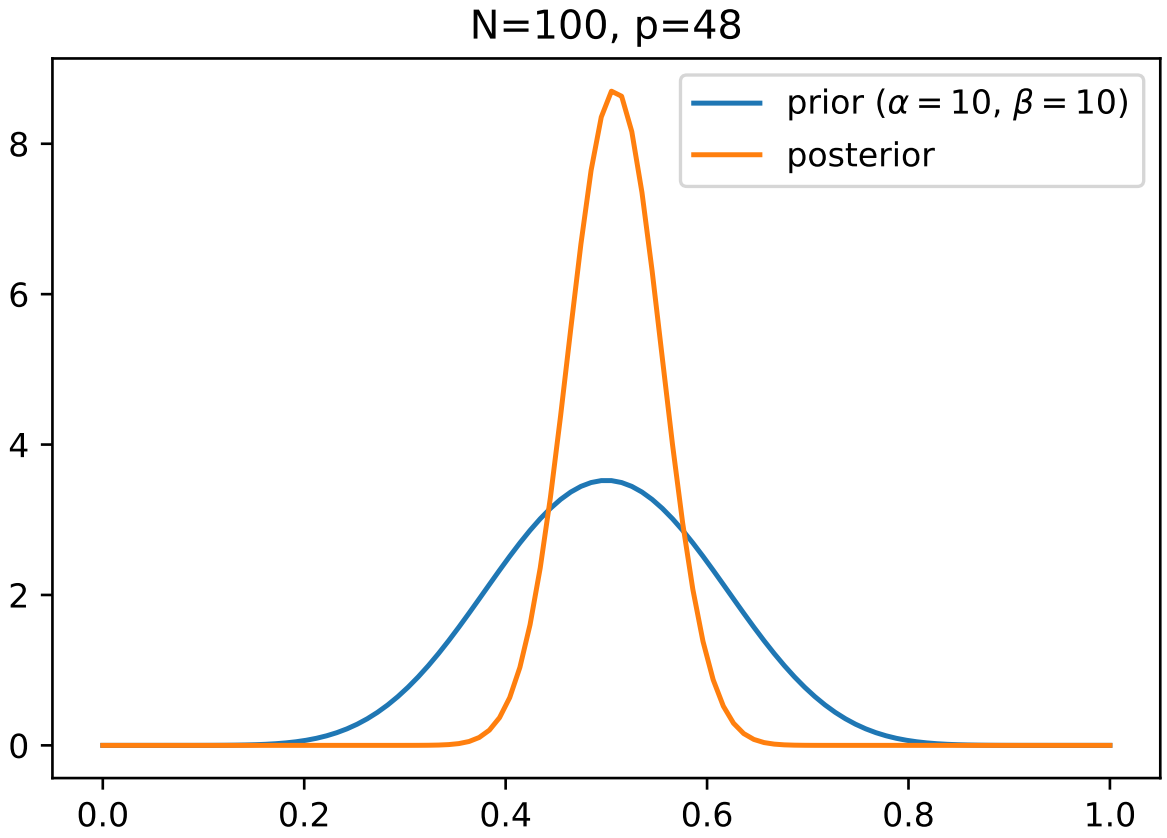
$$\begin{aligned} f_{X=x}(\theta) &\propto \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)+\Gamma(\beta)}\theta^{\alpha+x-1}(1-\theta)^{\beta-x} \\ &\propto \theta^{\alpha+x-1}(1-\theta)^{\beta-x} \end{aligned}$$

- on observe facilement que $f_X(\theta)$ est proportionnelle à une loi beta de paramètres $\alpha + x$ et $\beta + 1 - x$: La loi beta est conjuguée à la loi bernoulli.
- Exercice : calculer la loi *a posteriori* lorsque N lancés sont effectués et où p faces sont obtenues...

Exemple


N=100, p=48
prior ($\alpha = 1$, $\beta = 1$)
posterior
0
1
2
3
4
5
6
7
8
0.0
0.2
0.4
0.6
0.8
1.0

Exemple


N=100, p=48
prior ($\alpha = 10$, $\beta = 10$)
posterior
0
2
4
6
8
0.0
0.2
0.4
0.6
0.8
1.0

Conjugaison et famille exponentielle (remarques)

- Les lois de Bernoulli, beta, normal, etc. ont en commun d'appartenir à la même grande famille :

Conjugaison et famille exponentielle (remarques)

- Les lois de Bernoulli, beta, normal, etc. ont en commun d'appartenir à la même grande famille : la famille exponentielle,

Conjugaison et famille exponentielle (remarques)

- Les lois de Bernoulli, beta, normal, etc. ont en commun d'appartenir à la même grande famille : la famille exponentielle,
- elles peuvent, en particulier, toutes s'exprimer de la manière suivante (avec des variantes) :

$$f_\theta(x) = b(x)\exp\big({}^t\eta(\theta)\,T(x) - \mathcal{A}(\theta)\big),$$

  - où η est le paramètre naturel, $T(x)$, une statistique suffisante, $\mathcal{A}(\theta)$ la fonction de log-partition et $b(x)$ une mesure de base.

Exemple

Conjugaison et famille exponentielle (remarques)

Théorème

Toute loi de la famille exponentielle possède une loi a priori conjuguée (pas nécessairement en forme fermée).
Dans le cas où la densité s'exprime de la manière suivante (variante du slide précédent) :

$$f_\eta(x) = a(\eta)b(x)exp\big({}^t\eta T(x)\big),$$

La loi a priori admet la densité (par rapport à la mesure ...) suivante :

$$\pi(\theta;\chi,\nu) = f(\chi,\nu)exp\big({}^t\eta\chi - \nu\mathcal{A}(\eta)\big) = f(\chi,\nu)a(\eta)^\nu exp\big({}^t\eta\chi\big)$$

où η, $\mathcal{A}(\eta)$ proviennent de la vraisemblance.

Conjugaison et famille exponentielle (remarques)

- Calculons la loi *a posteriori* :

Conjugaison et famille exponentielle (remarques)

- Calculons la loi *a posteriori* :

$$\pi(\theta|X;\chi,\nu) \propto \mathcal{L}_\theta(X)\pi(\theta;\chi,\nu)$$

Conjugaison et famille exponentielle (remarques)

- Calculons la loi *a posteriori* :

$$\begin{aligned}\pi(\theta|X;\chi,\nu) &\propto \mathcal{L}_\theta(X)\pi(\theta;\chi,\nu)\\ &\propto f(\chi,\nu)a(\eta)^\nu \exp\left({}^t\eta\chi\right)a(\eta)b(x)\exp\left({}^t\eta T(x)\right)\end{aligned}$$

Conjugaison et famille exponentielle (remarques)

- Calculons la loi *a posteriori* :

$$\begin{aligned}\pi(\theta|X;\chi,\nu) &\propto \mathcal{L}_\theta(X)\pi(\theta;\chi,\nu)\\ &\propto f(\chi,\nu)a(\eta)^\nu \exp\big({}^t\eta\chi\big)a(\eta)b(x)\exp\big({}^t\eta T(x)\big)\\ &\propto a(\eta)^{\nu+1}\exp\big({}^t\eta(T(x)+\chi)\big)\end{aligned}$$

Conjugaison et famille exponentielle (remarques)

- Calculons la loi *a posteriori* :

$$\begin{aligned}
\pi(\theta|X;\chi,\nu) &\propto \mathcal{L}_\theta(X)\pi(\theta;\chi,\nu) \\
&\propto f(\chi,\nu)a(\eta)^\nu \exp\left({}^t\eta\chi\right)a(\eta)b(x)\exp\left({}^t\eta T(x)\right) \\
&\propto a(\eta)^{\nu+1}\exp\left({}^t\eta(T(x)+\chi)\right) \\
\Leftrightarrow \pi(\theta|X;\chi,\nu) &= \pi(\theta;\chi+T(x),\nu+1)
\end{aligned}$$

Exemple

- La loi de Bernoulli s'exprime comme suit :

$$f_\eta(x) = a(\eta)b(x)\exp\big({}^t\eta T(x)\big),$$

Exemple

- La loi de Bernoulli s'exprime comme suit :

$$f_\eta(x) = a(\eta)b(x)\exp\big({}^t\eta T(x)\big),$$

  - avec $\eta = \ln\frac{p}{1-p}$, $b(x) = 1$, $T(x) = x$ et $a(\eta) = \exp\big(-\ln(1+e^\eta)\big)$ d'où on obtient $f_\eta(x) = p^x(1-p)^{1-x}$.

Exemple

- La loi de Bernoulli s'exprime comme suit :

$$f_\eta(x) = a(\eta)b(x)\exp\big({}^t\eta T(x)\big),$$

  - avec $\eta = \ln\frac{p}{1-p}$, $b(x) = 1$, $T(x) = x$ et $a(\eta) = \exp\big(-\ln(1+e^\eta)\big)$ d'où on obtient $f_\eta(x) = p^x(1-p)^{1-x}$.
- son prior conjugué doit donc être de la forme :

$$\pi(\eta|\chi,\nu) = f(\chi,\nu)a(\eta)^\nu\exp\big({}^t\eta\chi\big)$$

Exemple

- La loi de Bernoulli s'exprime comme suit :

$$f_\eta(x) = a(\eta)b(x)\exp\big({}^t\eta T(x)\big),$$

  - avec $\eta = \ln\frac{p}{1-p}$, $b(x) = 1$, $T(x) = x$ et $a(\eta) = \exp\big(-\ln(1+e^\eta)\big)$ d'où on obtient $f_\eta(x) = p^x(1-p)^{1-x}$.
- son prior conjugué doit donc être de la forme :

$$\pi(\eta|\chi,\nu) = f(\chi,\nu)a(\eta)^\nu\exp\big({}^t\eta\chi\big) \propto a(\eta)^\nu\exp\big({}^t\eta\chi\big)$$

Exemple

- La loi de Bernoulli s'exprime comme suit :

$$f_\eta(x) = a(\eta)b(x)\exp\big({}^t\eta T(x)\big),$$

  - avec $\eta = \ln\frac{p}{1-p}$, $b(x) = 1$, $T(x) = x$ et $a(\eta) = \exp\big(-\ln(1+e^\eta)\big)$ d'où on obtient $f_\eta(x) = p^x(1-p)^{1-x}$.
- son prior conjugué doit donc être de la forme :

$$\begin{aligned}\pi(\eta|\chi,\nu) &= f(\chi,\nu)a(\eta)^\nu\exp\big({}^t\eta\chi\big) \propto a(\eta)^\nu\exp\big({}^t\eta\chi\big)\\ &\propto (1-p)^\nu p^\chi(1-p)^{-\chi} = p^\chi(1-p)^{\nu-\chi}\end{aligned}$$

Exemple

- La loi de Bernoulli s'exprime comme suit :

$$f_\eta(x) = a(\eta)b(x)\exp\big({}^t\eta T(x)\big),$$

  - avec $\eta = \ln\frac{p}{1-p}$, $b(x) = 1$, $T(x) = x$ et $a(\eta) = \exp\big(-\ln(1+e^\eta)\big)$ d'où on obtient $f_\eta(x) = p^x(1-p)^{1-x}$.
- son prior conjugué doit donc être de la forme :

$$\begin{aligned}\pi(\eta|\chi,\nu) &= f(\chi,\nu)a(\eta)^\nu\exp\big({}^t\eta\chi\big) \propto a(\eta)^\nu\exp\big({}^t\eta\chi\big)\\ &\propto (1-p)^\nu p^\chi(1-p)^{-\chi} = p^\chi(1-p)^{\nu-\chi}\end{aligned}$$

- On retrouve la loi beta avec $\alpha = \chi + 1$ et $\beta = \nu - \chi + 1$.

Priors non-informatifs

- L'objectif des *a priori* non informatifs est de limiter l'effet du choix de ce dernier sur l'inférence.
- Un premier choix est le prior uniforme qui échouera souvent à valider $\int \pi = 1$. On parle ici de prior impropre.
- Un autre choix possible est *l'a priori* de Jeffreys qui s'appuie sur la matrice d'information de Fisher. Le prior de Jeffreys est lui-même un *a priori* impropre.

A priori de Jeffreys : l'information de Fisher

- L'information de Fisher quantifie l'information associée à un (resp. des) paramètre θ.
- Dans le cas univarié, soit la fonction de score suivante :

$$\frac{\partial}{\partial\theta}\log\big(f(X;\theta)\big) = \frac{1}{f(X;\theta)}\frac{\partial f(X;\theta)}{\partial\theta}.$$

- L'information de Fisher est le second moment de la fonction de score : $I(\theta) = \mathbb{E}\Big[\Big(\frac{\partial}{\partial\theta}\log\big(f(X;\theta)\big)\Big)^2\Big|\theta\Big]$.
- *L'a priori* de Jeffreys se construit comme : $\pi(\theta) = |I(\theta)|^{1/2}$.
- Dans le cas multivarié, l'information de Fisher est une matrice et $|I(\theta)|$ représente la valeur absolue du déterminant.

Intervalle de confiance

- Un avantage du formalisme Bayésien est son cadre entièrement probabiliste.
- Soit α tel que l'intervalle représente une probabilité de $1-\alpha$, un intervalle de confiance est un ensemble $C_\alpha(\mathcal{D})$ tel que

$$\mathbb{P}\big(\theta \in C_\alpha(\mathcal{D})\big) = 1-\alpha.$$

- Cet ensemble n'est pas nécessairement unique, ce qui nous amène à la définition de "région de plus forte densité" (HDR) :

$$C_k(\mathcal{D}) = \{\theta | \pi(\theta|\mathcal{D}) \geq k\}.$$

- L'intervalle de confiance devient défini par le plus grand k tel que la probabilité de l'ensemble est $1-\alpha$.

Sélection de modèles

- Le cadre fréquentiste (Pearson, Fisher, etc.) permet de tester la validité d'hypothèses sur un paramètre θ.
- Le cadre Bayésien permet une autre philosophie.
- Soit $\mathcal{M} = \{M\}_{i \leq m}$ un ensemble de modèles et un jeu de données $\mathcal{D}$. On suppose que le jeu de données a été généré par l'un des modèle d'indice T_m. Chaque modèle a pour vraisemblance $\mathcal{L}(M_i|\mathcal{D})$.
- On souhaite déterminer $\mathbb{P}(T_m = k|\mathcal{D})$. Soit $\pi(T_m = k)$ un *a priori* (potentiellement uniforme) sur les modèles. La formule de Bayes nous donne :

$$\mathbb{P}(T_m = k|\mathcal{D}) \propto \pi(T_m = k)\mathcal{L}(\mathcal{D}|\mathcal{M}_k) = \pi(T_m = k) \int \mathcal{L}(\mathcal{D}|\theta_k)\pi(\theta_k)d\theta_k$$

- D'autres éléments comme le facteur de Bayes (équivalent du ratio de vraisemblance) peuvent entrer en jeu.

DAG (Directed Acyclic Graph)

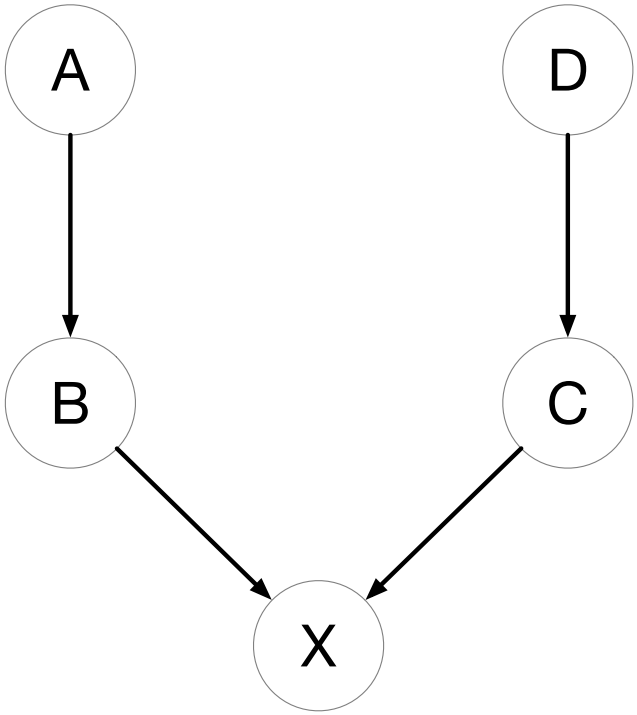
- Soit un modèle composé d'un vecteur de paramètres θ. Sans information sur le modèle, la loi de celui-ci se décomposerait comme suit :

$$\pi(\theta_1, \theta_2, ..., \theta_p|\mathcal{D}) = \pi(\theta_1|\theta_2, ..., \theta_p, \mathcal{D})\pi(\theta_2|\theta_3, ..., \theta_p, \mathcal{D})...\pi(\theta_p|\mathcal{D})$$

- En réalité, construire un modèle Bayésien consiste à définir les dépendances (et indépendances) entre variables aléatoires et à définir les lois associées.

Exemple

$$
\begin{aligned}
&\pi(B, C|X, A, D) \\
&\qquad = \pi(B|X, A, D, C)\pi(C|X, A, D) \\
&\qquad = \pi(B|X, A)\pi(C|X, D)
\end{aligned}
$$



Posterior prédictive

- Notons $\mathcal{D}$ le jeu de données (e.g. $\{(x_1, y_1), ...\}$),
- notre objectif est, étant donné un nouveau x, de prédire la variable à expliquer $\hat{y}$,

Posterior prédictive

- Notons $\mathcal{D}$ le jeu de données (e.g. $\{(x_1, y_1), ...\}$),
- notre objectif est, étant donné un nouveau x, de prédire la variable à expliquer $\hat{y}$,
- dans le cadre fréquentiste, $\mathcal{D}$ servirait à obtenir un estimateur $\hat{\theta}$ et on obtiendrait $\hat{y}$ à partir de x et $\hat{\theta}$.

Posterior prédictive

- Notons $\mathcal{D}$ le jeu de données (e.g. $\{(x_1, y_1), ...\}$),
- notre objectif est, étant donné un nouveau x, de prédire la variable à expliquer $\hat{y}$,
- dans le cadre fréquentiste, $\mathcal{D}$ servirait à obtenir un estimateur $\hat{\theta}$ et on obtiendrait $\hat{y}$ à partir de x et $\hat{\theta}$.
- dans le cadre Bayésien, on intègre l'espace des paramètres :

Posterior prédictive

- Notons $\mathcal{D}$ le jeu de données (e.g. $\{(x_1, y_1), ...\}$),
- notre objectif est, étant donné un nouveau x, de prédire la variable à expliquer $\hat{y}$,
- dans le cadre fréquentiste, $\mathcal{D}$ servirait à obtenir un estimateur $\hat{\theta}$ et on obtiendrait $\hat{y}$ à partir de x et $\hat{\theta}$.
- dans le cadre Bayésien, on intègre l'espace des paramètres :

$$f(\hat{y}|x, \mathcal{D}) =$$

Posterior prédictive

- Notons $\mathcal{D}$ le jeu de données (e.g. $\{(x_1, y_1), ...\}$),
- notre objectif est, étant donné un nouveau x, de prédire la variable à expliquer $\hat{y}$,
- dans le cadre fréquentiste, $\mathcal{D}$ servirait à obtenir un estimateur $\hat{\theta}$ et on obtiendrait $\hat{y}$ à partir de x et $\hat{\theta}$.
- dans le cadre Bayésien, on intègre l'espace des paramètres :

$$f(\hat{y}|x, \mathcal{D}) = \int_{\theta} f(\hat{y}|x, \theta) d\Pi(\theta|\mathcal{D})$$

Posterior prédictive

- Notons $\mathcal{D}$ le jeu de données (e.g. $\{(x_1, y_1), ...\}$),
- notre objectif est, étant donné un nouveau x, de prédire la variable à expliquer $\hat{y}$,
- dans le cadre fréquentiste, $\mathcal{D}$ servirait à obtenir un estimateur $\hat{\theta}$ et on obtiendrait $\hat{y}$ à partir de x et $\hat{\theta}$.
- dans le cadre Bayésien, on intègre l'espace des paramètres :

$$f(\hat{y}|x, \mathcal{D}) = \int_\theta f(\hat{y}|x, \theta) d\Pi(\theta|\mathcal{D}) = \int_\theta f(\hat{y}|x, \theta)\pi(\theta|\mathcal{D}) d\nu(\theta)$$

Typical set

- Soit une pièce (très légèrement) truquée qui tombe 51% du temps sur pile et 49% du temps sur face,

Typical set

- Soit une pièce (très légèrement) truquée qui tombe 51% du temps sur pile et 49% du temps sur face,
- parmi les deux séquences suivantes, laquelle provient le plus vraisemblablement de cette pièce :
  - $0, 0, 1, 0, 0, 1, 1, 0, 0, 1, 0, 0, 1, 1, 1, 1, 0, 0, 1, 0, 1, 0, 1, 0, 0, 0, 0, 0$
  - $1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1$

Typical set

- Soit une pièce (très légèrement) truquée qui tombe 51% du temps sur pile et 49% du temps sur face,
- parmi les deux séquences suivantes, laquelle provient le plus vraisemblablement de cette pièce :
  - $0, 0, 1, 0, 0, 1, 1, 0, 0, 1, 0, 0, 1, 1, 1, 1, 0, 0, 1, 0, 1, 0, 1, 0, 0, 0, 0, 0$
  - $1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1$
- et laquelle maximise la vraisemblance ?

Typical set

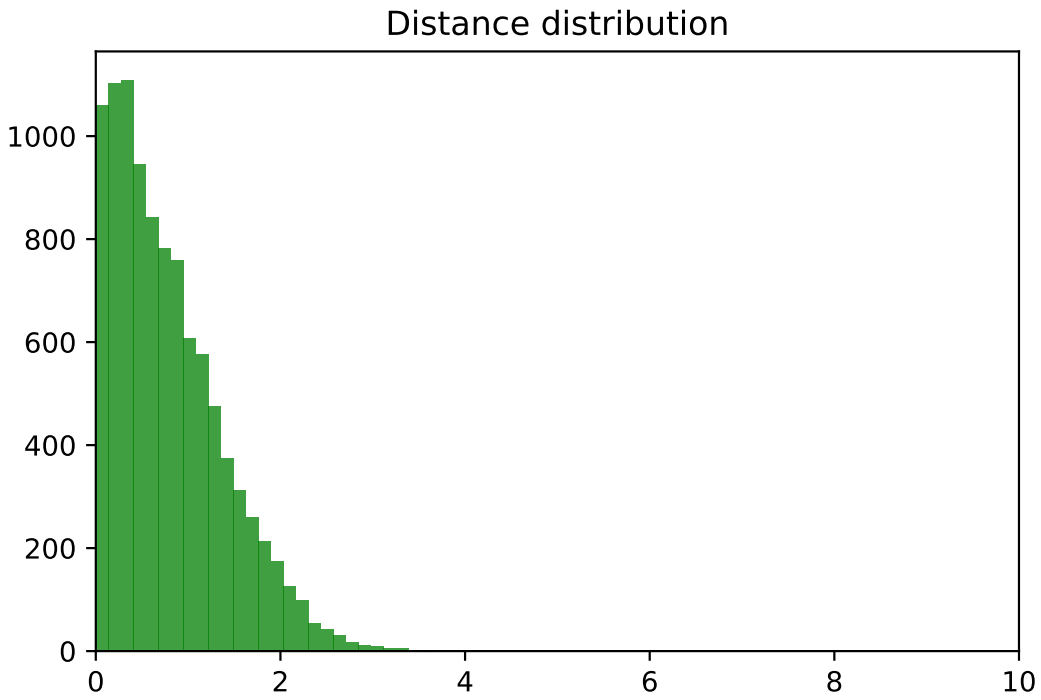
- Soit une pièce (très légèrement) truquée qui tombe 51% du temps sur pile et 49% du temps sur face,
- parmi les deux séquences suivantes, laquelle provient le plus vraisemblablement de cette pièce :
  - $0, 0, 1, 0, 0, 1, 1, 0, 0, 1, 0, 0, 1, 1, 1, 1, 0, 0, 1, 0, 1, 0, 1, 0, 0, 0, 0, 0$
  - $1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1$
- et laquelle maximise la vraisemblance ?
- le typical set décrit "l'endroit" dans l'espace des paramètres où on s'attend à voir les paramètres,
- le(s) point(s) de maximum de vraisemblance n'est pas (nécessairement) dans le typical set.

Malédiction de la dimension

- Soit $X \in \mathbb{R}^d$, $d \in \mathbb{N}^\star$ et $X \sim \mathcal{N}(0, I)$,
- Question : quelle est la forme de la loi de la distance de X à la moyenne selon la norme Euclidienne/ℓ_2 ($\|\boldsymbol{v}\|_2^2 = \sum_i v_i^2$) ?

Malédiction de la dimension

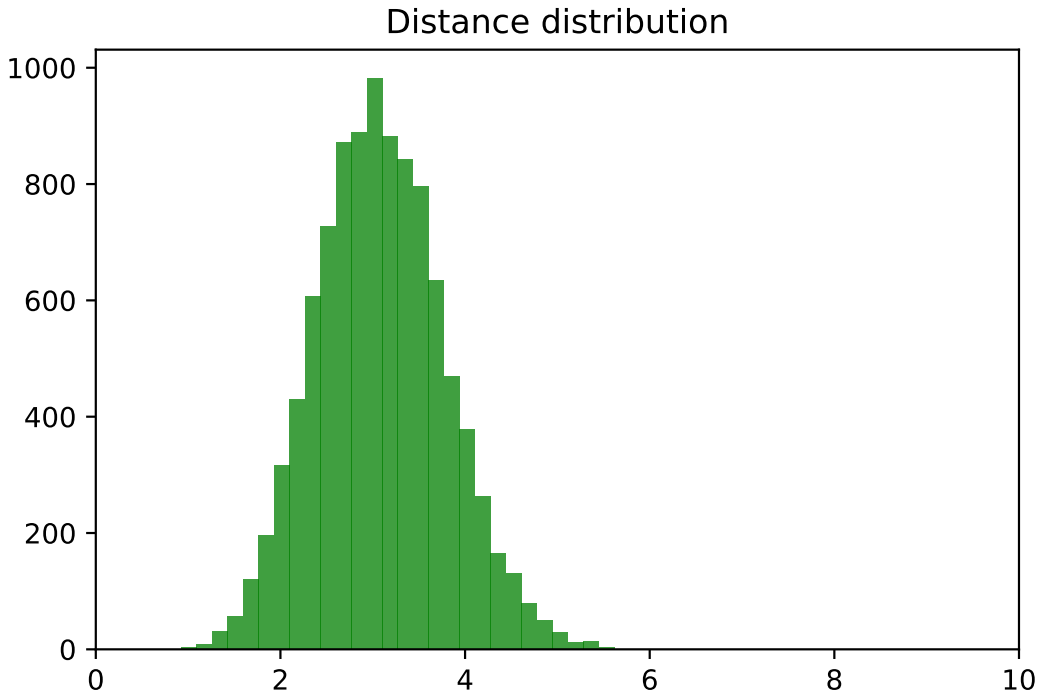
- Soit $X \in \mathbb{R}^d$, $d \in \mathbb{N}^\star$ et $X \sim \mathcal{N}(0, I)$,
- Question : quelle est la forme de la loi de la distance de X à la moyenne selon la norme Euclidienne/ℓ_2 ($\|\boldsymbol{v}\|_2^2 = \sum_i v_i^2$) ?



d=1 (10000 échantillons)

Malédiction de la dimension

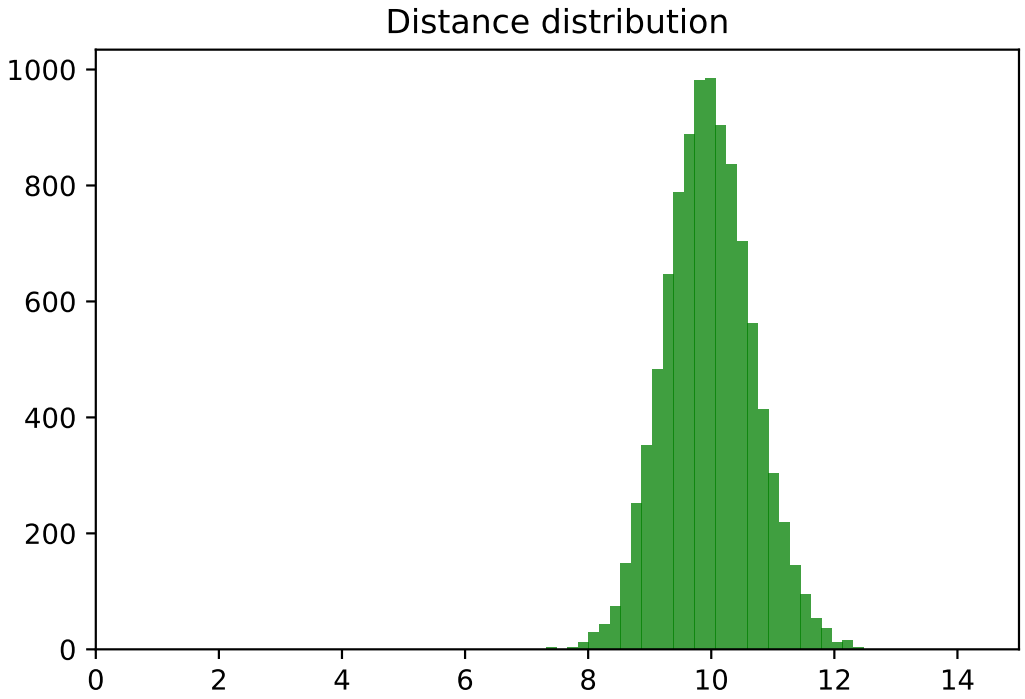
- Soit $X \in \mathbb{R}^d$, $d \in \mathbb{N}^\star$ et $X \sim \mathcal{N}(0, I)$,
- Question : quelle est la forme de la loi de la distance de X à la moyenne ?



d=10 (10000 échantillons)

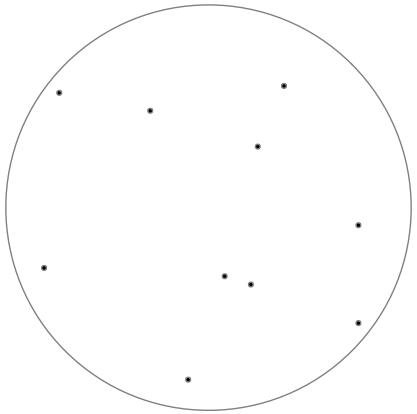
Malédiction de la dimension

- Soit $X \in \mathbb{R}^d$, $d \in \mathbb{N}^\star$ et $X \sim \mathcal{N}(0, I)$,
- Question : quelle est la forme de la loi de la distance de X à la moyenne ?



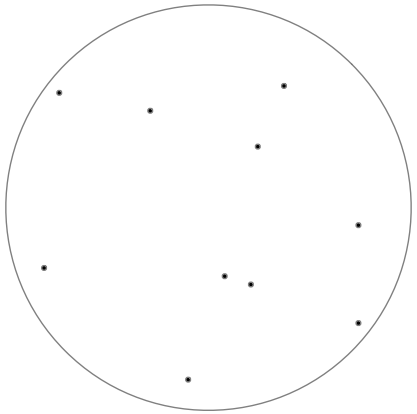
d=15 (10000 échantillons)

Malédiction de la dimension 2



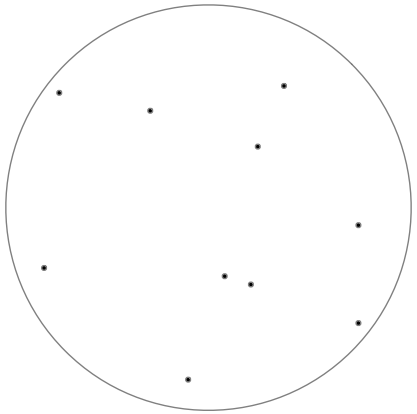
- Si on tire N points uniformément dans une (hyper)sphère s de dimension d et de rayon r, quelle est la distance minimale médiane entre ces derniers et le centre du cercle ?

Malédiction de la dimension 2



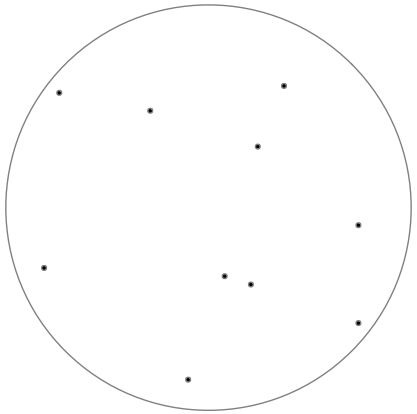
- Si on tire N points uniformément dans une (hyper)sphère s de dimension d et de rayon r, quelle est la distance minimale médiane entre ces derniers et le centre du cercle ?
- $V \propto \pi r^d$,

Malédiction de la dimension 2



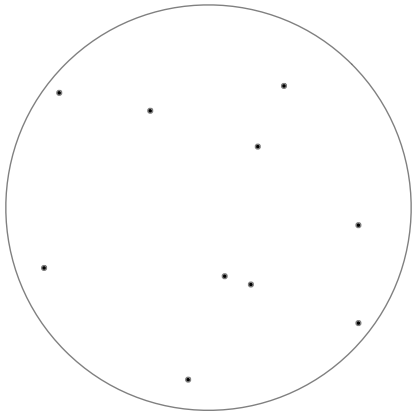
- Si on tire N points uniformément dans une (hyper)sphère s de dimension d et de rayon r, quelle est la distance minimale médiane entre ces derniers et le centre du cercle ?
- $V \propto \pi r^d$,
- considérons une hypersphère s' de rayon r' (valeur que l'on recherche) de même centre que la première,

Malédiction de la dimension 2



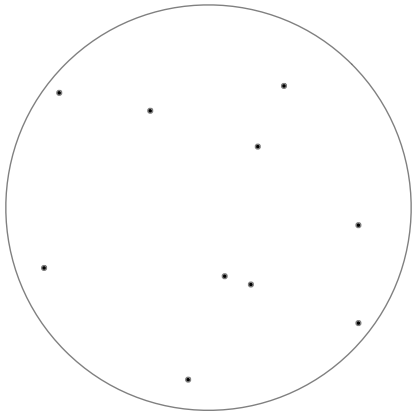
- Si on tire N points uniformément dans une (hyper)sphère s de dimension d et de rayon r, quelle est la distance minimale médiane entre ces derniers et le centre du cercle ?
- $V \propto \pi r^d$,
- considérons une hypersphère s' de rayon r' (valeur que l'on recherche) de même centre que la première,
- $\mathbb{P}(\forall x, x \notin s') = \left(1 - (r')^d\right)^N = \frac{1}{2} \iff r' = (1 - \frac{1}{2}^{\frac{1}{N}})^{\frac{1}{d}}$

Malédiction de la dimension 2



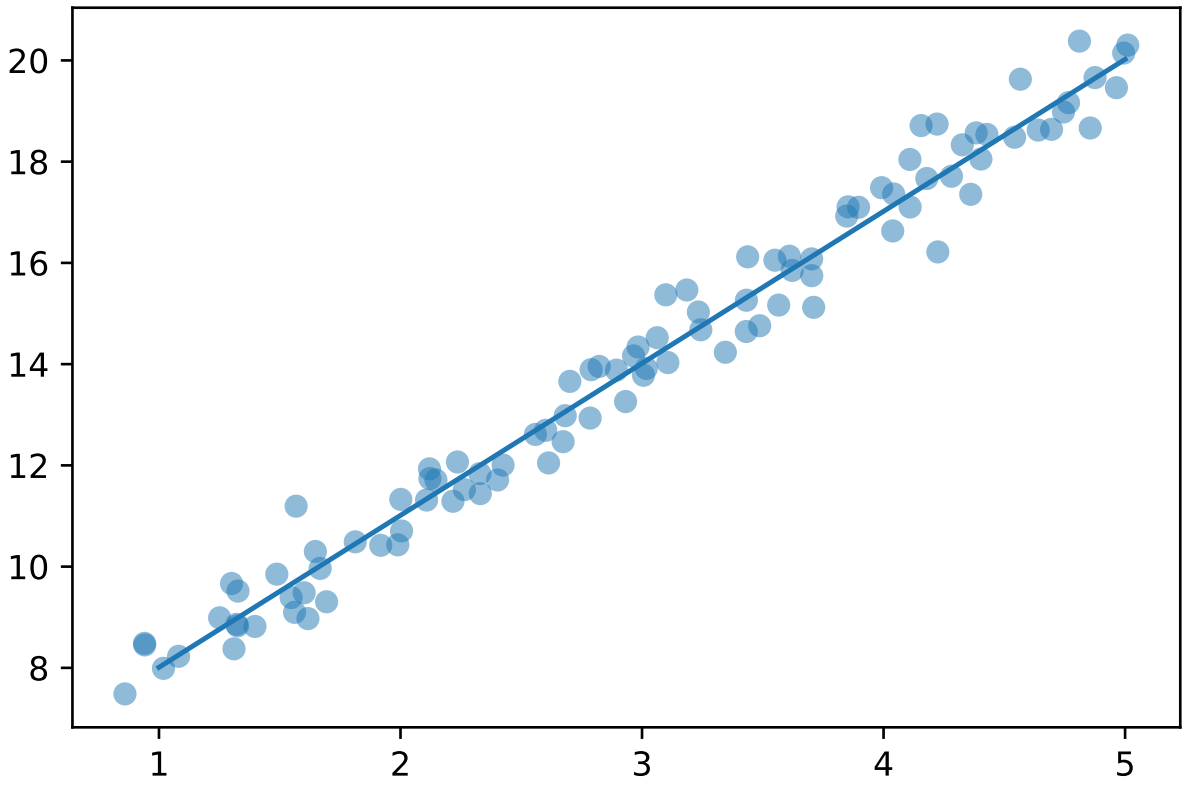
- Si on tire N points uniformément dans une (hyper)sphère s de dimension d et de rayon r, quelle est la distance minimale médiane entre ces derniers et le centre du cercle ?
- $V \propto \pi r^d$,
- considérons une hypersphère s' de rayon r' (valeur que l'on recherche) de même centre que la première,
- $\mathbb{P}(\forall x, x \notin s') = \left(1 - (r')^d\right)^N = \frac{1}{2} \iff r' = (1 - \frac{1}{2}^{\frac{1}{N}})^{\frac{1}{d}}$
  - e.g. $d = 100, N = 100 \iff r' = 0.95$

Malédiction de la dimension 2



- Si on tire N points uniformément dans une (hyper)sphère s de dimension d et de rayon r, quelle est la distance minimale médiane entre ces derniers et le centre du cercle ?
- $V \propto \pi r^d$,
- considérons une hypersphère s' de rayon r' (valeur que l'on recherche) de même centre que la première,
- $\mathbb{P}(\forall x, x \notin s') = \left(1 - (r')^d\right)^N = \frac{1}{2} \iff r' = (1 - \frac{1}{2}^{\frac{1}{N}})^{\frac{1}{d}}$
  - e.g. $d = 100, N = 100 \iff r' = 0.95$
  - e.g. $d = 100 \times 100, N = 100 \iff r' = 0.9995$

La régression linéaire



La régression linéaire (vision fréquentiste)

- On dispose d'un jeu de données i.i.d.
  $\mathcal{D} = \{(x_1, y_1), ..., (x_N, y_N)\}$ avec $x_i \in \mathbb{R}^d$ et $y_i \in \mathbb{R}$,

La régression linéaire (vision fréquentiste)

- On dispose d'un jeu de données i.i.d.
  $\mathcal{D} = \{(x_1, y_1), ..., (x_N, y_N)\}$ avec $x_i \in \mathbb{R}^d$ et $y_i \in \mathbb{R}$,
- on suppose qu'il existe une combinaison linéaire des variables explicatives (les dimensions du vecteur x_i) permettant d'expliquer la variable y_i à un bruit près :

$$y_i = f(x_i) + \epsilon_i = {}^t x_i \theta + \epsilon_i$$

où $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ et $\theta \in \mathbb{R}^d$.

La régression linéaire (vision fréquentiste)

- On dispose d'un jeu de données i.i.d.
  $\mathcal{D} = \{(x_1, y_1), ..., (x_N, y_N)\}$ avec $x_i \in \mathbb{R}^d$ et $y_i \in \mathbb{R}$,
- on suppose qu'il existe une combinaison linéaire des variables explicatives (les dimensions du vecteur x_i) permettant d'expliquer la variable y_i à un bruit près :

$$y_i = f(x_i) + \epsilon_i = {}^t x_i \theta + \epsilon_i$$

  où $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ et $\theta \in \mathbb{R}^d$.
- Ainsi, en reformulant, on obtient :

$$y_i \sim \mathcal{N}({}^t x_i \theta, \sigma^2)$$

La régression linéaire (vision fréquentiste)

- On dispose d'un jeu de données i.i.d.
  $\mathcal{D} = \{(x_1, y_1), ..., (x_N, y_N)\}$ avec $x_i \in \mathbb{R}^d$ et $y_i \in \mathbb{R}$,
- on suppose qu'il existe une combinaison linéaire des variables explicatives (les dimensions du vecteur x_i) permettant d'expliquer la variable y_i à un bruit près :

$$y_i = f(x_i) + \epsilon_i = {}^t x_i \theta + \epsilon_i$$

  où $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ et $\theta \in \mathbb{R}^d$.
- Ainsi, en reformulant, on obtient :

$$y_i \sim \mathcal{N}({}^t x_i \theta, \sigma^2)$$

- Exercice : écrire la vraisemblance du jeu de données $\mathcal{D}$.

La régression linéaire (vision fréquentiste)

- La (log) vraisemblance est le produit (somme) des (log) densités (on remarque le lien direct avec les moindres carrés) :

$$\log \mathcal{L}_\theta(\mathcal{D}) \propto -\frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - {}^t x_i \theta)^2$$

La régression linéaire (vision fréquentiste)

- La (log) vraisemblance est le produit (somme) des (log) densités (on remarque le lien direct avec les moindres carrés) :

$$\log \mathcal{L}_\theta(\mathcal{D}) \propto -\frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - {}^t x_i \theta)^2$$

- Un estimateur $\hat{\theta}$ est donc : $\hat{\theta} = \arg\min_\theta \frac{1}{2} \sum_{i=1}^{N} (y_i - {}^t x_i \theta)^2$

La régression linéaire (vision fréquentiste)

- La (log) vraisemblance est le produit (somme) des (log) densités (on remarque le lien direct avec les moindres carrés) :

$$\log \mathcal{L}_\theta(\mathcal{D}) \propto -\frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - {}^t x_i \theta)^2$$

- Un estimateur $\hat{\theta}$ est donc : $\hat{\theta} = \arg\min_\theta \frac{1}{2} \sum_{i=1}^{N} (y_i - {}^t x_i \theta)^2$
- Soit la matrice $X = \begin{bmatrix} {}^t x_1 \\ .... \\ {}^t x_N \end{bmatrix}$ et $y = \begin{bmatrix} y_1 \\ .... \\ y_N \end{bmatrix}$, la solution est donnée par :

La régression linéaire (vision fréquentiste)

- La (log) vraisemblance est le produit (somme) des (log) densités (on remarque le lien direct avec les moindres carrés) :

$$\log \mathcal{L}_\theta(\mathcal{D}) \propto -\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - {}^t x_i \theta)^2$$

- Un estimateur $\hat{\theta}$ est donc : $\hat{\theta} = \arg\min_\theta \frac{1}{2}\sum_{i=1}^{N}(y_i - {}^t x_i \theta)^2$
- Soit la matrice $X = \begin{bmatrix} {}^t x_1 \\ .... \\ {}^t x_N \end{bmatrix}$ et $y = \begin{bmatrix} y_1 \\ .... \\ y_N \end{bmatrix}$, la solution est donnée par :

$$\hat{\theta} = ({}^t XX)^{-1}{}^t XY$$

La régression linéaire (vision Bayésienne)

- De la même manière que dans le cadre fréquentiste, on a :

$$y_i = f(x_i) + \epsilon_i = {}^t x_i \theta + \epsilon_i$$

où $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ et $\theta \in \mathbb{R}^d$.

La régression linéaire (vision Bayésienne)

- De la même manière que dans le cadre fréquentiste, on a :

$$y_i = f(x_i) + \epsilon_i = {}^t x_i \theta + \epsilon_i$$

  où $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ et $\theta \in \mathbb{R}^d$.
- Ainsi, en reformulant, on obtient :

$$y_i \sim \mathcal{N}({}^t x_i \theta, \sigma^2)$$

La régression linéaire (vision Bayésienne)

- De la même manière que dans le cadre fréquentiste, on a :

$$y_i = f(x_i) + \epsilon_i = {}^t x_i \theta + \epsilon_i$$

où $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ et $\theta \in \mathbb{R}^d$.

- Ainsi, en reformulant, on obtient :

$$y_i \sim \mathcal{N}({}^t x_i \theta, \sigma^2)$$

- Considérons la variance σ^2 connue et munissons θ d'une mesure *a priori* conjuguée :

$$\pi(\theta) = \mathcal{N}(0, \sigma^2 \Lambda^{-1}),$$

où Λ est la matrice de précision (si $\Lambda = \lambda I$, calculer un MAP revient à faire un Ridge).

La régression linéaire (vision Bayésienne)

- L'objectif final reste d'obtenir des prédictions sur $\hat{y}$ sachant un x. Il s'agit donc de calculer la densité a posteriori prédictive :

$$f(\hat{y}|x, \mathcal{D}) =$$

La régression linéaire (vision Bayésienne)

- L'objectif final reste d'obtenir des prédictions sur $\hat{y}$ sachant un x. Il s'agit donc de calculer la densité a posteriori prédictive :

$$f(\hat{y}|x, \mathcal{D}) = \int_{\theta} f(\hat{y}|x, \theta, \sigma^2) d\Pi(\theta|\mathcal{D})$$

La régression linéaire (vision Bayésienne)

- L'objectif final reste d'obtenir des prédictions sur $\hat{y}$ sachant un x. Il s'agit donc de calculer la densité a posteriori prédictive :

$$\begin{aligned} f(\hat{y}|x,\mathcal{D}) &= \int_\theta f(\hat{y}|x,\theta,\sigma^2)d\Pi(\theta|\mathcal{D}) \\ &= \int_\theta f(\hat{y}|x,\theta,\sigma^2)\pi(\theta|\mathcal{D})d\nu(\theta) \end{aligned}$$

La régression linéaire (vision Bayésienne)

- L'objectif final reste d'obtenir des prédictions sur $\hat{y}$ sachant un x. Il s'agit donc de calculer la densité a posteriori prédictive :

$$\begin{aligned} f(\hat{y}|x,\mathcal{D}) &= \int_\theta f(\hat{y}|x,\theta,\sigma^2)d\Pi(\theta|\mathcal{D}) \\ &= \int_\theta f(\hat{y}|x,\theta,\sigma^2)\pi(\theta|\mathcal{D})d\nu(\theta) \end{aligned}$$

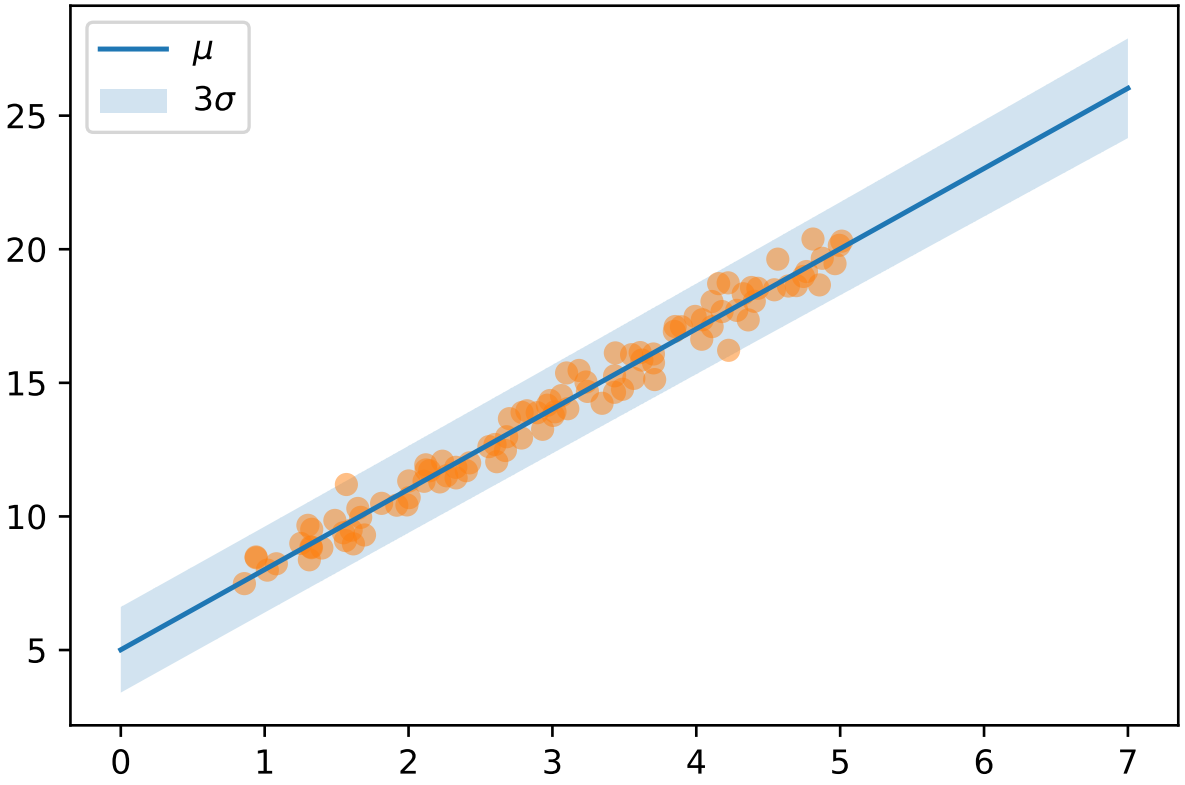
- On obtient la loi suivante (en évitant les très nombreux calculs) :

$$\hat{y}|x,\mathcal{D} \sim \mathcal{N}({}^t x\theta_n, \sigma^2 + {}^t x\Sigma_n x)$$

avec

$$\theta_n = ({}^t XX + \Lambda)^{-1}\,{}^t XY \quad \text{et} \quad \Sigma_n = \sigma^2({}^t XX + \Lambda)^{-1}$$

La régression linéaire (vision Bayésienne)



Plan

1 Rappels

2 L'approche Bayésienne

3 Les méthodes MCMC
- Motivation et intuition
- Metropolis-Hastings
- Échantilloneur de Gibbs
- Hamiltonian Monte Carlo

4 Les méthodes approchées

Motivation

- Faire du "Bayésien" revient à calculer des intégrales

Exemple

Si on souhaite obtenir l'espérance d'un paramètre θ, alors on doit calculer :

$$\mathbb{E}_X[\theta] = \int_\theta \theta f_X(\theta) d\nu(\theta)$$

Et c'est pareil lorsqu'on veut marginaliser, calculer la variance, la posterior prédictive, etc.

- En pratique, calculer ces intégrales peut être (très, très, très) difficile, voire impossible,
- Or on sait que si on simule $X_1, ..., X_n \sim d\nu$ alors $\frac{1}{n}\sum_{i=1}^N g(X_i)$ convergera en probabilité vers $\mathbb{E}[g(X)]$,

Monte Carlo et π

- On souhaite calculer π en utilisant la formule de l'aire du cercle de rayon r :

$$Aire(r) = \pi r^2, r \in \mathbb{R}^+$$

- Un cercle est donné par l'ensemble des points équidistants à son centre. Ainsi, un cercle de rayon 1 centré en $(0, 0)$ est défini par :

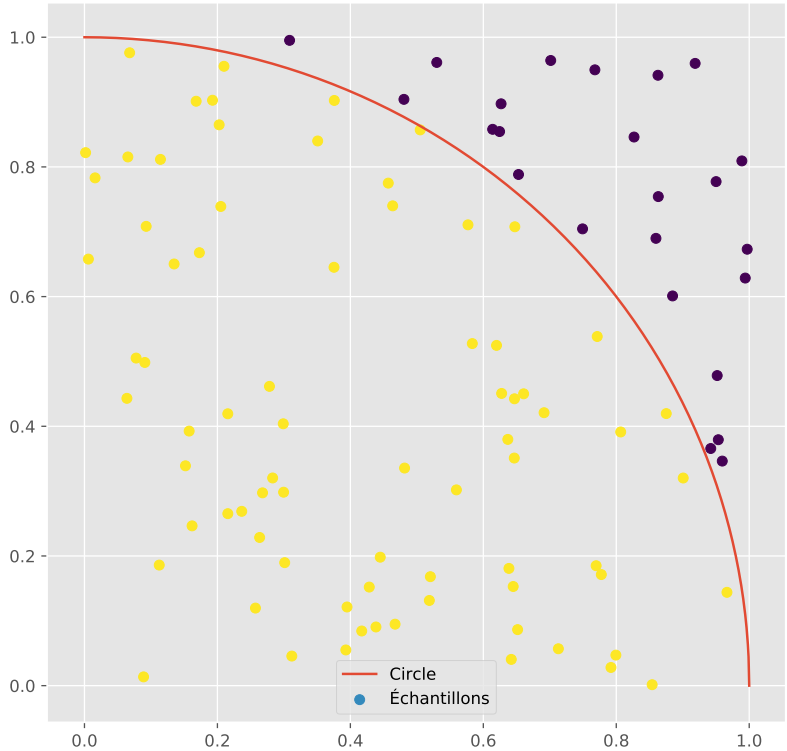
$$\{(x, y) : \sqrt{x^2 + y^2} = 1\}$$

- Si on se restreint à $0 \leq x, y \leq 1$, alors on obtient un quart de cercle. La formule suivante donne ainsi π :

$$\pi = 4 \int_0^1 \sqrt{1 - x^2} dx$$

Monte Carlo et π

- Supposons qu'on tire des points uniformément dans le carré $[0,1]^2$. Le ratio du nombre de points satisfaisants $\sqrt{x^2+y^2} \leq 1$ sera l'aire du quart de cercle relativement à celle du carré (d'aire 1).
- On fait ainsi une approximation de π.



Importance sampling

- Nous disposons maintenant d'une densité de probabilité g ne permettant ni un calcul analytique ni de simuler.
- On constate cependant l'égalité suivante :

$$\mathbb{E}[h(X)] = \int h(x)g(x)dx = \int h(x)\frac{g(x)}{\gamma(x)}\gamma(x)dx = \mathbb{E}_{X\sim\gamma}\left[h(X)\frac{g(X)}{\gamma(X)}\right]$$

  où on sait simuler selon γ.
- Ainsi, pour calculer l'espérance, il suffit de simuler selon γ et d'appliquer l'a formule $h(X)\frac{g(X)}{\gamma(X)}$.

L'approche MCMC

- On se situe dans l'espace des paramètres $\{\Theta, \mathcal{F}, \Pi_{\theta|\mathcal{D}}\}$ où $\Pi_{\theta|\mathcal{D}}$ est la loi *a posteriori*,

L'approche MCMC

- On se situe dans l'espace des paramètres $\{\Theta, \mathcal{F}, \Pi_{\theta|\mathcal{D}}\}$ où $\Pi_{\theta|\mathcal{D}}$ est la loi *a posteriori*,
- l'approche MCMC consiste à générer une suite de vecteurs $\{\theta_1, ..., \theta_N\}$ telle que chaque $\theta_i \sim \Pi_{\theta|\mathcal{D}}$,

L'approche MCMC

- On se situe dans l'espace des paramètres $\{\Theta, \mathcal{F}, \Pi_{\theta|\mathcal{D}}\}$ où $\Pi_{\theta|\mathcal{D}}$ est la loi *a posteriori*,
- l'approche MCMC consiste à générer une suite de vecteurs $\{\theta_1, ..., \theta_N\}$ telle que chaque $\theta_i \sim \Pi_{\theta|\mathcal{D}}$,
- la procédure est itérative, pour chaque θ_i, seule la connaissance de θ_{i-1} est nécessaire (Markov Chain) : on simule selon $q(\theta_i|\theta_{i-1})$,

L'approche MCMC

- On se situe dans l'espace des paramètres $\{\Theta, \mathcal{F}, \Pi_{\theta|\mathcal{D}}\}$ où $\Pi_{\theta|\mathcal{D}}$ est la loi *a posteriori*,
- l'approche MCMC consiste à générer une suite de vecteurs $\{\theta_1, ..., \theta_N\}$ telle que chaque $\theta_i \sim \Pi_{\theta|\mathcal{D}}$,
- la procédure est itérative, pour chaque θ_i, seule la connaissance de θ_{i-1} est nécessaire (Markov Chain) : on simule selon $q(\theta_i|\theta_{i-1})$,
- enfin, supposons que l'on soit intéressé par $\mathbb{E}[\theta|\mathcal{D}] = \int \theta d\Pi(\theta|\mathcal{D})$,

L'approche MCMC

- On se situe dans l'espace des paramètres $\{\Theta, \mathcal{F}, \Pi_{\theta|\mathcal{D}}\}$ où $\Pi_{\theta|\mathcal{D}}$ est la loi *a posteriori*,
- l'approche MCMC consiste à générer une suite de vecteurs $\{\theta_1, ..., \theta_N\}$ telle que chaque $\theta_i \sim \Pi_{\theta|\mathcal{D}}$,
- la procédure est itérative, pour chaque θ_i, seule la connaissance de θ_{i-1} est nécessaire (Markov Chain) : on simule selon $q(\theta_i|\theta_{i-1})$,
- enfin, supposons que l'on soit intéressé par $\mathbb{E}[\theta|\mathcal{D}] = \int \theta d\Pi(\theta|\mathcal{D})$, la loi des grands nombre nous dit que $\frac{1}{N}\sum_{i=1}^{N} \theta_i$ converge en probabilité vers $\mathbb{E}[\theta|\mathcal{D}]$,

L'approche MCMC

- On se situe dans l'espace des paramètres $\{\Theta, \mathcal{F}, \Pi_{\theta|\mathcal{D}}\}$ où $\Pi_{\theta|\mathcal{D}}$ est la loi *a posteriori*,
- l'approche MCMC consiste à générer une suite de vecteurs $\{\theta_1, ..., \theta_N\}$ telle que chaque $\theta_i \sim \Pi_{\theta|\mathcal{D}}$,
- la procédure est itérative, pour chaque θ_i, seule la connaissance de θ_{i-1} est nécessaire (Markov Chain) : on simule selon $q(\theta_i|\theta_{i-1})$,
- enfin, supposons que l'on soit intéressé par $\mathbb{E}[\theta|\mathcal{D}] = \int \theta d\Pi(\theta|\mathcal{D})$, la loi des grands nombre nous dit que $\frac{1}{N}\sum_{i=1}^{N} \theta_i$ converge en probabilité vers $\mathbb{E}[\theta|\mathcal{D}]$,
- et c'est la même chose pour les autres statistiques que l'on souhaiterait calculer (Attention aux biais).

La variance empirique

- Soit des variables $X_1, ..., X_n$, i.i.d. de variance σ^2 et moyenne μ. Soit $\sigma_e^2 = \frac{1}{N}\sum_i (X_i - \overline{X})^2$. Calculons $\mathbb{E}\left[\sigma_e^2\right]$.

$$\begin{aligned}
\mathbb{E}\left[\sigma_e^2\right] &= \mathbb{E}\left[\frac{1}{N}\sum_i (X_i - \overline{X})^2\right] = \mathbb{E}\left[\frac{1}{N}\left(\sum_i X_i^2 - 2\overline{X}\sum_i X_i + \sum_i \overline{X}^2\right)\right] \\
&= \mathbb{E}\left[\frac{1}{N}\left(\sum_i X_i^2 - n\overline{X}^2\right)\right] = \frac{1}{N} N\mathbb{E}\left[X_i^2\right] - \mathbb{E}\left[\overline{X}^2\right] \\
&= \mu^2 + \sigma^2 - \left(\mu^2 - \frac{\sigma^2}{N}\right) = \frac{N-1}{N}\sigma^2
\end{aligned}$$

- Afin d'obtenir un estimateur non biaisé de la variance, il faut multiplier par $N/(N-1)$.

Exemple

- On dispose d'un jeu de données i.i.d.
  $\mathcal{D} = \{(x_1, y_1), ..., (x_N, y_N)\}$ où $x_i \in \mathbb{R}^2$ et $y_i \in \mathbb{R}$,

Exemple

- On dispose d'un jeu de données i.i.d. $\mathcal{D} = \{(x_1, y_1), ..., (x_N, y_N)\}$ où $x_i \in \mathbb{R}^2$ et $y_i \in \mathbb{R}$,
- on sait qu'il existe une combinaison linéaire des variables explicatives $x_{i,1}$ et $x_{i,2}$ permettant d'obtenir la variable à expliquer y_i à un bruit près :

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \epsilon_i$$

Exemple

- On dispose d'un jeu de données i.i.d. $\mathcal{D} = \{(x_1, y_1), ..., (x_N, y_N)\}$ où $x_i \in \mathbb{R}^2$ et $y_i \in \mathbb{R}$,
- on sait qu'il existe une combinaison linéaire des variables explicatives $x_{i,1}$ et $x_{i,2}$ permettant d'obtenir la variable à expliquer y_i à un bruit près :

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \epsilon_i$$

- où $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ et on munit les paramètres d'une loi *a priori* $\sigma \sim \mathcal{IG}(1, 1)$ et $\beta_0, \beta_1, \beta_2 \sim \mathcal{N}(0, 10)$.

Exemple

- On dispose d'un jeu de données i.i.d. $\mathcal{D} = \{(x_1, y_1), ..., (x_N, y_N)\}$ où $x_i \in \mathbb{R}^2$ et $y_i \in \mathbb{R}$,
- on sait qu'il existe une combinaison linéaire des variables explicatives $x_{i,1}$ et $x_{i,2}$ permettant d'obtenir la variable à expliquer y_i à un bruit près :

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \epsilon_i$$

- où $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ et on munit les paramètres d'une loi *a priori* $\sigma \sim \mathcal{IG}(1, 1)$ et $\beta_0, \beta_1, \beta_2 \sim \mathcal{N}(0, 10)$.
  - Notez que des *a priori* conjugués par paramètres ont été choisis...
- Notez que *l'a priori* normale-inverse-gamma ("produit des deux priors") est directement conjugué avec les deux paramètres de la loi normale.

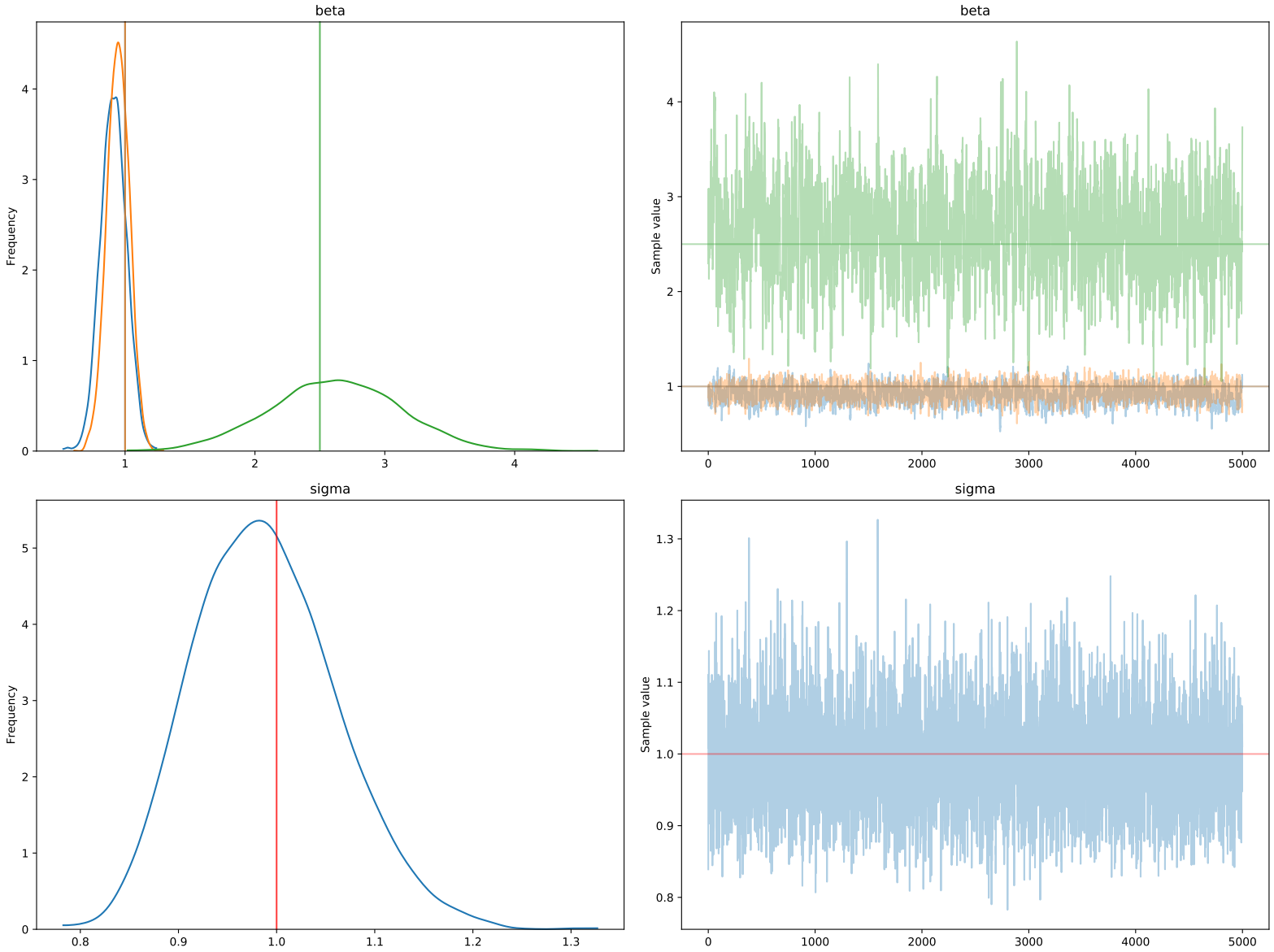
Exemple

- On dispose d'un jeu de données i.i.d. $\mathcal{D} = \{(x_1, y_1), ..., (x_N, y_N)\}$ où $x_i \in \mathbb{R}^2$ et $y_i \in \mathbb{R}$,
- on sait qu'il existe une combinaison linéaire des variables explicatives $x_{i,1}$ et $x_{i,2}$ permettant d'obtenir la variable à expliquer y_i à un bruit près :

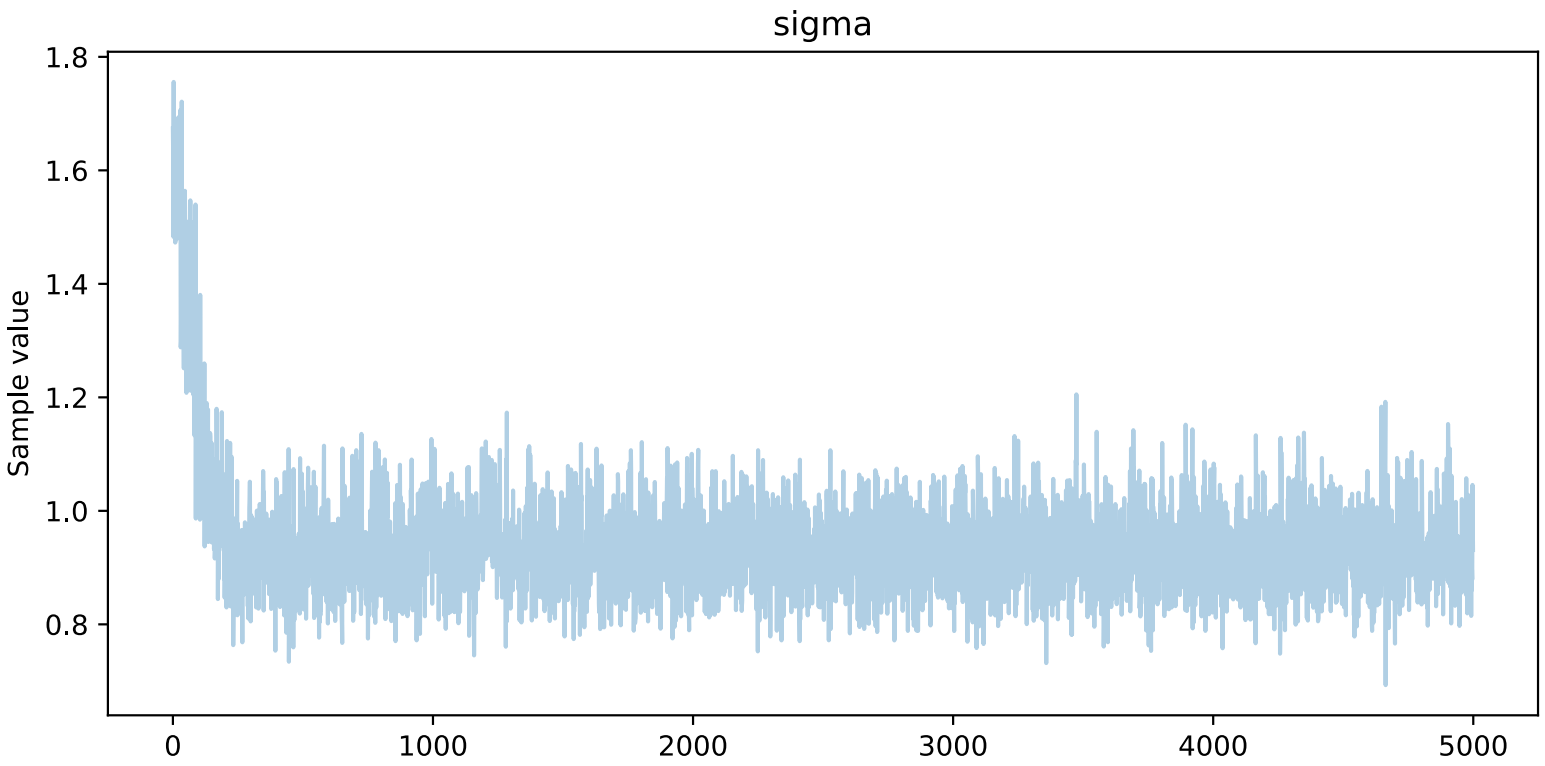
$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \epsilon_i$$

- où $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ et on munit les paramètres d'une loi *a priori* $\sigma \sim \mathcal{IG}(1, 1)$ et $\beta_0, \beta_1, \beta_2 \sim \mathcal{N}(0, 10)$.
  - Notez que des *a priori* conjugués par paramètres ont été choisis...
- Notez que *l'a priori* normale-inverse-gamma ("produit des deux priors") est directement conjugué avec les deux paramètres de la loi normale.
- Les algorithmes d'inférence permettent aussi de résoudre des cas où la solution analytique serait "trop" difficile à trouver...

Exemple (suite)

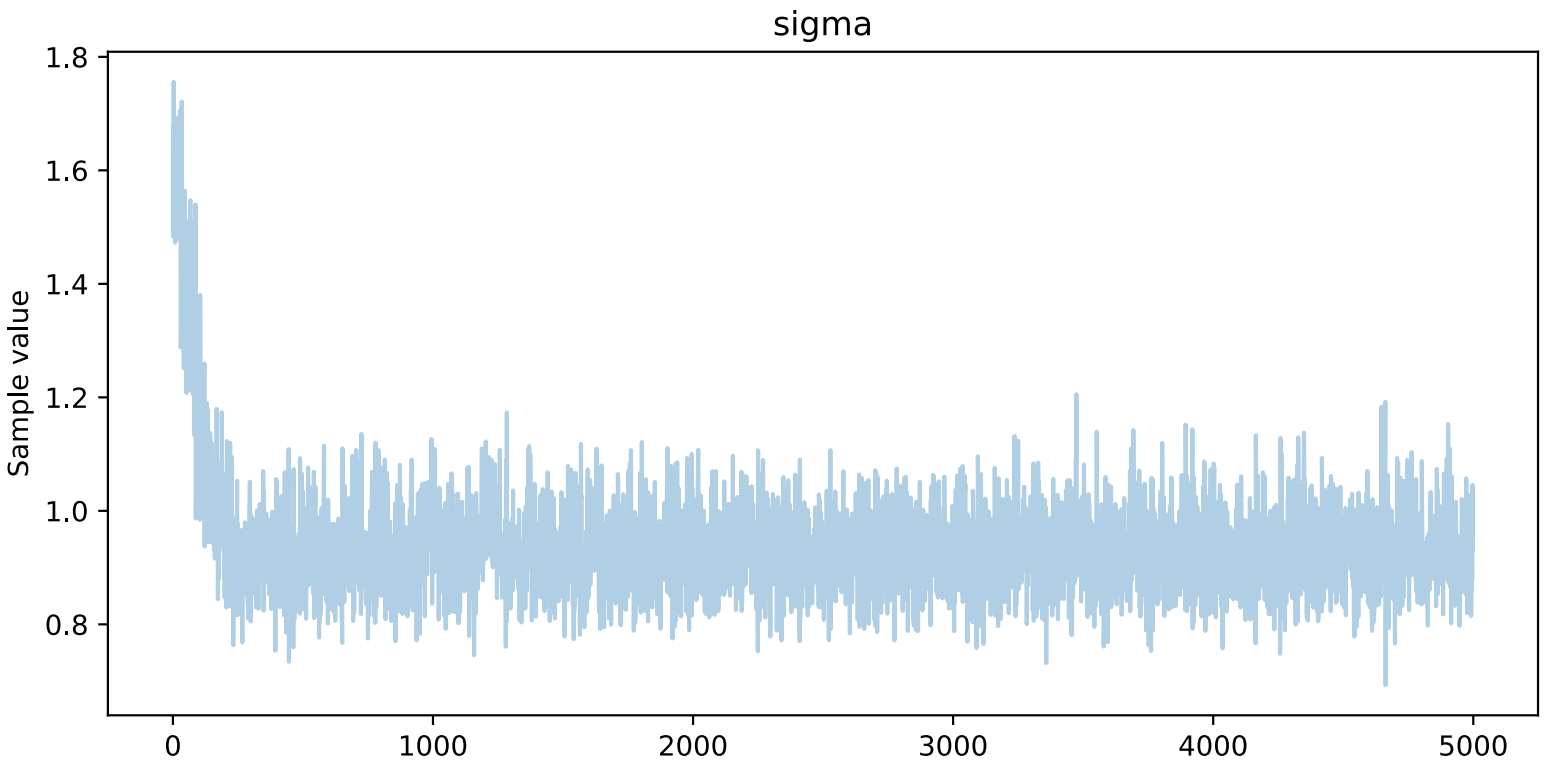


Typical set



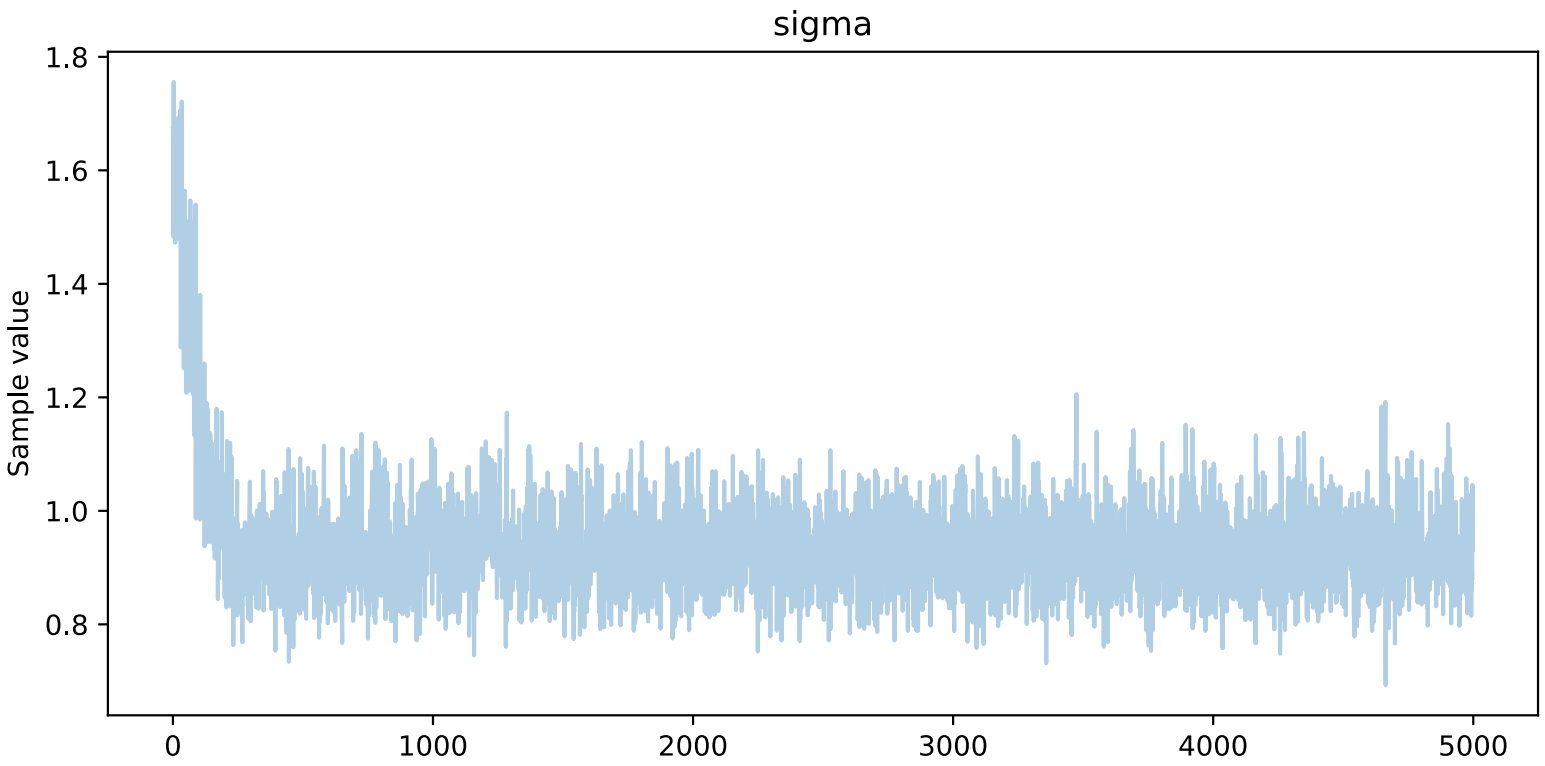
- Généralement, lorsque l'inférence commence, les paramètres ne sont pas situés dans les zones de fortes densités (typical set),

Typical set



- Généralement, lorsque l'inférence commence, les paramètres ne sont pas situés dans les zones de fortes densités (typical set),
- la première étape consiste à retrouver le typical set (e.g. MCMC ou optimisation MAP),

Typical set



- Généralement, lorsque l'inférence commence, les paramètres ne sont pas situés dans les zones de fortes densités (typical set),
- la première étape consiste à retrouver le typical set (e.g. MCMC ou optimisation MAP),
- tous les échantillons créés à ce moment ne servent à rien (burn-in).

Chaîne de Markov (rappels)

Chaîne de Markov (rappels)

Définition

Toute suite $\{X_i\}_{i>0}$ *de variables aléatoires sur un espace probabilisé* $\{A, \mathcal{T}, P_\theta\}$ *satisfont la propriété de Markov faible si :*

$$\forall B \in \mathcal{T}, P_\theta(X_n \in B | X_1, ..., X_i, ..., X_{n-1}) = P_\theta(X_n \in B | X_{n-1})$$

Définition

Toute suite $\{X_i\}_{i>0}$ *de variables aléatoires satisfaisant la propriété de Markov est une chaîne de Markov.*

Exemple

- soit une particule associée à une variable aléatoire symbolisant sa position à chaque itération X_i,

Exemple

- soit une particule associée à une variable aléatoire symbolisant sa position à chaque itération X_i,
- à chaque itération, la particule a 50% de chance d'aller 1 mètre à droite et autant d'aller 1 mètre à gauche,

Exemple

- soit une particule associée à une variable aléatoire symbolisant sa position à chaque itération X_i,
- à chaque itération, la particule a 50% de chance d'aller 1 mètre à droite et autant d'aller 1 mètre à gauche,
- ici, clairement, la position à l'itération d'après dépend uniquement de la position actuelle et non des positions précédentes :

$$P(X_n = x | X_1, ..., X_{n-1}) = P(X_n = x | X_{n-1})$$

État stationnaire

Définition

La probabilité P est dite stationnaire pour la chaîne de Markov $\{X_i\}_{i>0}$ décrivant des états discrets si :

$$P(X = i) = \lim_{x \to \infty} \frac{\mathbb{1}[X_1 = i] + ... + \mathbb{1}[X_n = i]}{n}$$

ou, de manière équivalente (similaire pour le cas continu),

$$P(X_n = i) = \sum_j P(X_{n-1} = j) \times P(X_n = j | X_{n-1} = j)$$

$$f(x) = \int f(y) p(x|y) dy$$

Équilibre ponctuel

Définition

Une chaîne de Markov $\{X_i\}_{i>0}$ *vérifie un équilibre ponctuel si*

$$f(x)p(y|x) = f(y)p(x|y)$$

Équilibre ponctuel

Définition

Une chaîne de Markov $\{X_i\}_{i>0}$ *vérifie un équilibre ponctuel si*

$$f(x)p(y|x) = f(y)p(x|y)$$

Théorème

De plus, si une chaîne de Markov $\{X_i\}_{i>0}$ *vérifie un équilibre ponctuel, alors la densité* f *est dans un état stationnaire.*

Équilibre ponctuel

Définition

Une chaîne de Markov $\{X_i\}_{i>0}$ vérifie un équilibre ponctuel si

$$f(x)p(y|x) = f(y)p(x|y)$$

Théorème

De plus, si une chaîne de Markov $\{X_i\}_{i>0}$ vérifie un équilibre ponctuel, alors la densité f est dans un état stationnaire.

- Preuve :

$$\int f(x)p(y|x)dy =$$

Équilibre ponctuel

Définition

Une chaîne de Markov $\{X_i\}_{i>0}$ vérifie un équilibre ponctuel si

$$f(x)p(y|x) = f(y)p(x|y)$$

Théorème

De plus, si une chaîne de Markov $\{X_i\}_{i>0}$ vérifie un équilibre ponctuel, alors la densité f est dans un état stationnaire.

- Preuve :

$$\int f(x)p(y|x)dy = f(x)\int p(y|x)dy =$$

Équilibre ponctuel

Définition

Une chaîne de Markov $\{X_i\}_{i>0}$ vérifie un équilibre ponctuel si

$$f(x)p(y|x) = f(y)p(x|y)$$

Théorème

De plus, si une chaîne de Markov $\{X_i\}_{i>0}$ vérifie un équilibre ponctuel, alors la densité f est dans un état stationnaire.

- Preuve :

$$\int f(x)p(y|x)dy = f(x)\int p(y|x)dy = f(x)$$

Équilibre ponctuel

Définition

Une chaîne de Markov $\{X_i\}_{i>0}$ vérifie un équilibre ponctuel si

$$f(x)p(y|x) = f(y)p(x|y)$$

Théorème

De plus, si une chaîne de Markov $\{X_i\}_{i>0}$ vérifie un équilibre ponctuel, alors la densité f est dans un état stationnaire.

- Preuve :

$$\int f(x)p(y|x)dy = f(x)\int p(y|x)dy = f(x) = \int f(y)p(x|y)dy$$

MCMC

- L'objectif d'un algorithme de Monte Carlo Markov Chain et de construire une séquence d'échantillons telle qu'elle est une chaîne de Markov dont la densité stationnaire est celle de notre loi *a posteriori*.

L'algorithme de Metropolis-Hastings

Input: $\theta_0 \in \Theta$ et $n \in \mathbb{N}*$ le nombre d'échantillons souhaités

L'algorithme de Metropolis-Hastings

Input: $\theta_0 \in \Theta$ et $n \in \mathbb{N}*$ le nombre d'échantillons souhaités
Output: $\{\theta_i\}_{i>0}^n$

L'algorithme de Metropolis-Hastings

Input: $\theta_0 \in \Theta$ et $n \in \mathbb{N}*$ le nombre d'échantillons souhaités
Output: $\{\theta_i\}_{i>0}^{n}$
for i ***in*** $1, ..., n$ **do**

L'algorithme de Metropolis-Hastings

Input: $\theta_0 \in \Theta$ et $n \in \mathbb{N}*$ le nombre d'échantillons souhaités
Output: $\{\theta_i\}_{i>0}^n$
for i ***in*** $1, ..., n$ **do**
  simuler $y \sim q(\cdot|\theta_{i-1})$;

L'algorithme de Metropolis-Hastings

Input: $\theta_0 \in \Theta$ et $n \in \mathbb{N}*$ le nombre d'échantillons souhaités
Output: $\{\theta_i\}_{i>0}^n$
for i ***in*** $1, ..., n$ **do**

simuler $y \sim q(\cdot|\theta_{i-1})$;

soit $r(x, y) = \min\left(\frac{f(y)q(x|y)}{f(x)q(y|x)}, 1\right)$;

L'algorithme de Metropolis-Hastings

Input: $\theta_0 \in \Theta$ et $n \in \mathbb{N}*$ le nombre d'échantillons souhaités
Output: $\{\theta_i\}_{i>0}^n$
for i ***in*** $1, ..., n$ **do**

simuler $y \sim q(\cdot|\theta_{i-1})$;

soit $r(x, y) = \min\left(\frac{f(y)q(x|y)}{f(x)q(y|x)}, 1\right)$;

$$\theta_i = \begin{cases} y & \text{avec probabilité } r(\theta_i, y) \\ \theta_{i-1} & \text{avec probabilité } 1 - r(\theta_i, y) \end{cases}$$

L'algorithme de Metropolis-Hastings

Input: $\theta_0 \in \Theta$ et $n \in \mathbb{N}*$ le nombre d'échantillons souhaités
Output: $\{\theta_i\}_{i>0}^n$
for i ***in*** $1, ..., n$ **do**

simuler $y \sim q(\cdot|\theta_{i-1})$;

soit $r(x, y) = \min\left(\frac{f(y)q(x|y)}{f(x)q(y|x)}, 1\right)$;

$$\theta_i = \begin{cases} y & \text{avec probabilité } r(\theta_i, y) \\ \theta_{i-1} & \text{avec probabilité } 1 - r(\theta_i, y) \end{cases}$$

end

Algorithm 1: Metropolis-Hasting

L'algorithme de Metropolis-Hastings

Input: $\theta_0 \in \Theta$ et $n \in \mathbb{N}*$ le nombre d'échantillons souhaités
Output: $\{\theta_i\}_{i>0}^n$
for i ***in*** $1, ..., n$ **do**

simuler $y \sim q(\cdot|\theta_{i-1})$;

soit $r(x, y) = \min\left(\frac{f(y)q(x|y)}{f(x)q(y|x)}, 1\right)$;

$$\theta_i = \begin{cases} y & \text{avec probabilité } r(\theta_i, y) \\ \theta_{i-1} & \text{avec probabilité } 1 - r(\theta_i, y) \end{cases}$$

end

Algorithm 1: Metropolis-Hasting

⚠Il suffit de connaître la densité f à un <u>facteur proportionnel</u> près !

Preuve de l'état stationnaire

Dans le cas Bayésien, f est la loi *a posteriori*. Montrons que l'état stationnaire est f :

Preuve de l'état stationnaire

Dans le cas Bayésien, f est la loi *a posteriori*. Montrons que l'état stationnaire est f :

- Soit $\theta_i \neq \theta_{i-1}$ tels que $f(\theta_i)q(\theta_{i-1}|\theta_i) \leq f(\theta_{i-1})q(\theta_i|\theta_{i-1})$

Preuve de l'état stationnaire

Dans le cas Bayésien, f est la loi *a posteriori*. Montrons que l'état stationnaire est f :

- Soit $\theta_i \neq \theta_{i-1}$ tels que $f(\theta_i)q(\theta_{i-1}|\theta_i) \leq f(\theta_{i-1})q(\theta_i|\theta_{i-1})$

$$p(\theta_i|\theta_{i-1}) = q(\theta_i|\theta_{i-1})r(\theta_{i-1}, \theta_i) = \frac{f(\theta_i)q(\theta_{i-1}|\theta_i)}{f(\theta_{i-1})}$$

Preuve de l'état stationnaire

Dans le cas Bayésien, f est la loi *a posteriori*. Montrons que l'état stationnaire est f :

- Soit $\theta_i \neq \theta_{i-1}$ tels que $f(\theta_i)q(\theta_{i-1}|\theta_i) \leq f(\theta_{i-1})q(\theta_i|\theta_{i-1})$

$$p(\theta_i|\theta_{i-1}) = q(\theta_i|\theta_{i-1})r(\theta_{i-1}, \theta_i) = \frac{f(\theta_i)q(\theta_{i-1}|\theta_i)}{f(\theta_{i-1})}$$

$$\Leftrightarrow p(\theta_i|\theta_{i-1})f(\theta_{i-1}) = f(\theta_i)q(\theta_{i-1}|\theta_i)$$

Preuve de l'état stationnaire

Dans le cas Bayésien, f est la loi *a posteriori*. Montrons que l'état stationnaire est f :

- Soit $\theta_i \neq \theta_{i-1}$ tels que $f(\theta_i)q(\theta_{i-1}|\theta_i) \leq f(\theta_{i-1})q(\theta_i|\theta_{i-1})$

$$p(\theta_i|\theta_{i-1}) = q(\theta_i|\theta_{i-1})r(\theta_{i-1}, \theta_i) = \frac{f(\theta_i)q(\theta_{i-1}|\theta_i)}{f(\theta_{i-1})}$$

$$\Leftrightarrow p(\theta_i|\theta_{i-1})f(\theta_{i-1}) = f(\theta_i)q(\theta_{i-1}|\theta_i)$$

- de manière symétrique, on obtient $p(\theta_{i-1}|\theta_i) = q(\theta_{i-1}|\theta_i)$ ce qui nous donne un équilibre ponctuel pour $\theta_i \neq \theta_{i-1}$:

$$p(\theta_i|\theta_{i-1})f(\theta_{i-1}) = p(\theta_{i-1}|\theta_i)f(\theta_i)$$

Preuve de l'état stationnaire

Dans le cas Bayésien, f est la loi *a posteriori*. Montrons que l'état stationnaire est f :

- Soit $\theta_i \neq \theta_{i-1}$ tels que $f(\theta_i)q(\theta_{i-1}|\theta_i) \leq f(\theta_{i-1})q(\theta_i|\theta_{i-1})$

$$p(\theta_i|\theta_{i-1}) = q(\theta_i|\theta_{i-1})r(\theta_{i-1}, \theta_i) = \frac{f(\theta_i)q(\theta_{i-1}|\theta_i)}{f(\theta_{i-1})}$$

$$\Leftrightarrow p(\theta_i|\theta_{i-1})f(\theta_{i-1}) = f(\theta_i)q(\theta_{i-1}|\theta_i)$$

- de manière symétrique, on obtient $p(\theta_{i-1}|\theta_i) = q(\theta_{i-1}|\theta_i)$ ce qui nous donne un équilibre ponctuel pour $\theta_i \neq \theta_{i-1}$:

$$p(\theta_i|\theta_{i-1})f(\theta_{i-1}) = p(\theta_{i-1}|\theta_i)f(\theta_i)$$

- on retrouve de la même manière l'équilibre ponctuel pour $\theta_i = \theta_{i-1}$.

L'échantilloneur de Gibbs

- L'échantilloneur de Gibbs peut, en quelque sorte, être considéré comme un cas très particulier de Metropolis-Hasting,
- l'algorithme s'appuie sur les distributions marginales *a posteriori* des paramètres qui doivent donc être disponibles,

L'échantilloneur de Gibbs

- L'échantilloneur de Gibbs peut, en quelque sorte, être considéré comme un cas très particulier de Metropolis-Hasting,
- l'algorithme s'appuie sur les distributions marginales *a posteriori* des paramètres qui doivent donc être disponibles,

Input: $\theta^{(0)} \in \Theta$ et $n \in \mathbb{N}*$ le nombre d'échantillons souhaités
Output: $\{\theta_i\}_{i>0}^{n}$
for t ***in*** $1, ..., n$ **do**
  simuler $\theta_1^{(t)} \sim \pi(\theta_1^{(t)} | \theta_2^{(t-1)}, ..., \theta_P^{(t-1)})$;
  ...;
  simuler $\theta_i^{(t)} \sim \pi(\theta_i^{(t)} | \theta_1^{(t)}, ..., \theta_{i-1}^{(t)}, \theta_{i+1}^{(t-1)}, .., \theta_P^{(t-1)})$;
  ...;
  simuler $\theta_P^{(t)} \sim \pi(\theta_P^{(t)} | \theta_1^{(t)}, ..., \theta_{P-1}^{(t)})$;
end

Algorithm 2: Échantillonneur de Gibbs

L'échantillonneur de Gibbs

- Il est souvent très facile d'obtenir les loi marginales *a posteriori* dans les modèles Bayésiens hiérarchiques,
- ⚠attention, si les *a priori* ne sont pas conjugués, simuler selon la loi *a posteriori* peut être compliqué et d'autres stratégies deviennent nécessaires,
- contrairement à la version naïve de Metropolis-Hastings, l'échantillonneur de Gibbs a une idée de la "direction à prendre" dans l'espace des paramètres.

Exemple

- Nous disposons d'un ensemble de tâches T pour lesquelles la réponse (inconnue) est 1 ou 0,

Exemple

- Nous disposons d'un ensemble de tâches T pour lesquelles la réponse (inconnue) est 1 ou 0,
- pour déterminer cette réponse, nous allons demander à un certain nombre d'individus de voter,

Exemple

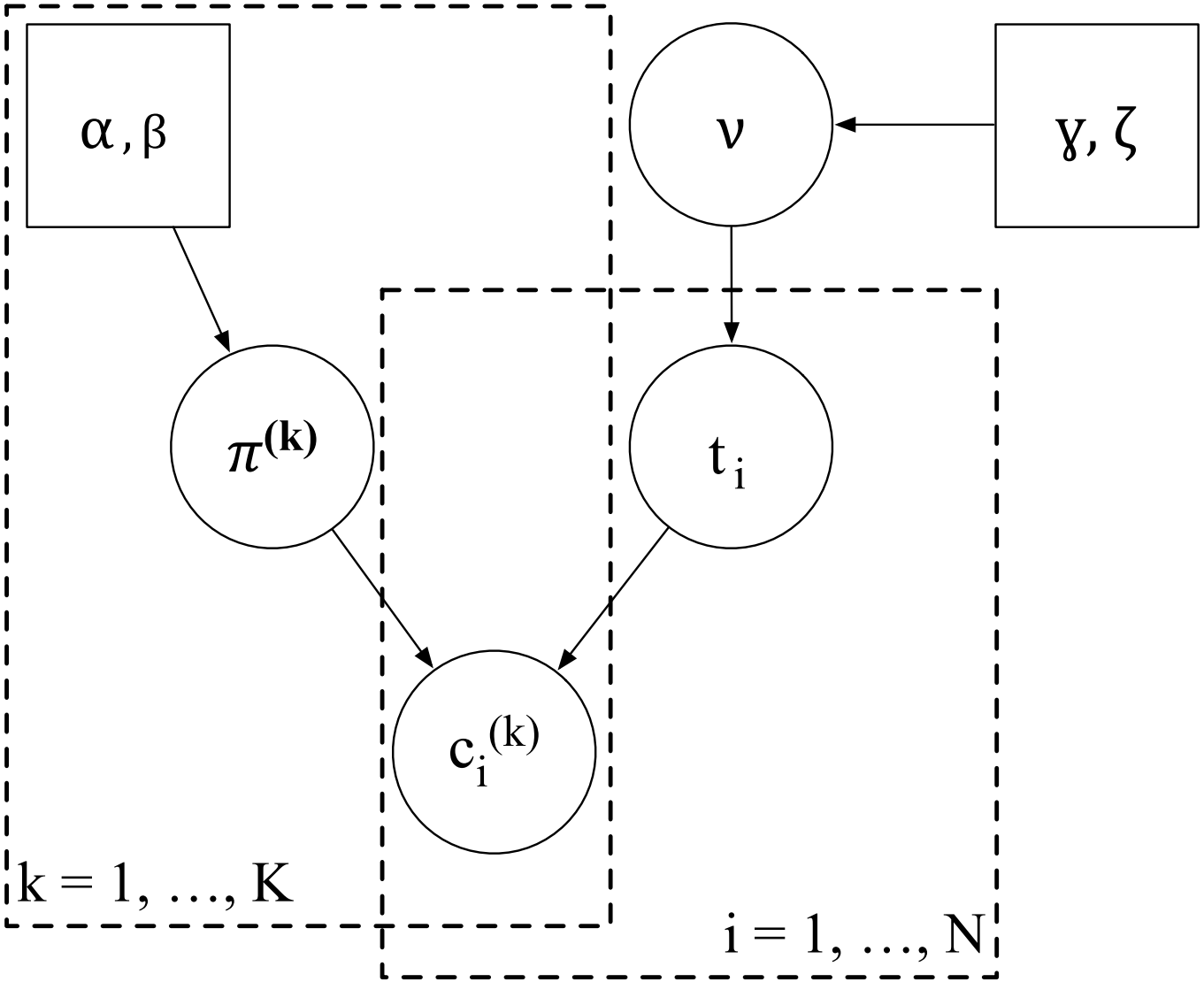
- Nous disposons d'un ensemble de tâches T pour lesquelles la réponse (inconnue) est 1 ou 0,
- pour déterminer cette réponse, nous allons demander à un certain nombre d'individus de voter,
- chaque tâches est indexée par i et est associée à une vraie réponse inconnue notée t_i,

$$t_i \sim Be(\nu),\ \nu \sim \text{Beta}(\gamma, \zeta)$$

- chaque individu fournit pour chaque tâche une réponse $c_i^{(k)} \in \{0, 1\}$:

$$\mathbb{P}(c_i^{(k)} = t_i | t_i) = \pi^{(k)},\ \pi^{(k)} \sim \text{Beta}(\alpha, \beta)$$

Exemple (DAG)



Exemple

- Une fois le modèle construit et les données collectées, nous sommes intéressés par les lois marginales *a posteriori*,

Exemple

- Une fois le modèle construit et les données collectées, nous sommes intéressés par les lois marginales *a posteriori*,
  - des individus (quelles sont leurs performances moyennes)
  - et des tâches (quelle étiquette associer à chaque tâche),

Exemple

- Une fois le modèle construit et les données collectées, nous sommes intéressés par les lois marginales *a posteriori*,
  - des individus (quelles sont leurs performances moyennes)
  - et des tâches (quelle étiquette associer à chaque tâche),
- L'échantillonneur de Gibbs a besoin que nous dérivions (analytiquement) les loi marginales *a posteriori* des paramètres.
- Le plus simple est de commencer par formuler la loi jointe :

$$p(\mathbf{\Pi}, \boldsymbol{T}, \nu|\alpha, \beta, \gamma, \zeta, \mathcal{D}) \propto \prod_i \Big[\prod_k \big(\pi^{(k)\,\delta[c_i^{(k)}=t_i]}(\mathbf{1}-\pi^{(k)})^{\delta[c_i^{(k)}\neq t_i]}\big)\Big] p(\mathbf{\Pi}|\alpha, \beta) p(\nu|\gamma, \zeta)$$

Exemple

- Commençons par $\pi^{(k)}$:

$$\pi^{(k)}|\text{rest} \propto (\pi^{(k)})^{\alpha+N_+^{(k)}-1}(1-\pi^{(k)})^{\beta+N-N_+^{(k)}-1}$$

avec $N_+^{(k)} = \sum_{i=1}^{N} \delta_{c_i^{(k)}-t_i}$ où δ_x vaut 0 partout sauf quand $x = 1$ (mesure de Dirac). On retrouve la loi Beta de paramètres $\alpha + N_+^{(k)}$ et $\beta + N - N_+^{(k)}$.

Exemple

- Commençons par $\pi^{(k)}$:

$$\pi^{(k)}|\text{rest} \propto (\pi^{(k)})^{\alpha+N_+^{(k)}-1}(1-\pi^{(k)})^{\beta+N-N_+^{(k)}-1}$$

  avec $N_+^{(k)} = \sum_{i=1}^{N} \delta_{c_i^{(k)}-t_i}$ où δ_x vaut 0 partout sauf quand $x = 1$ (mesure de Dirac). On retrouve la loi Beta de paramètres $\alpha + N_+^{(k)}$ et $\beta + N - N_+^{(k)}$.
- Pour ν, de manière très similaire, nous obtenons :

Exemple

- Commençons par $\pi^{(k)}$:

$$\pi^{(k)}|\text{rest} \propto (\pi^{(k)})^{\alpha+N_+^{(k)}-1}(1-\pi^{(k)})^{\beta+N-N_+^{(k)}-1}$$

avec $N_+^{(k)} = \sum_{i=1}^{N} \delta_{c_i^{(k)}-t_i}$ où δ_x vaut 0 partout sauf quand $x = 1$ (mesure de Dirac). On retrouve la loi Beta de paramètres $\alpha + N_+^{(k)}$ et $\beta + N - N_+^{(k)}$.

- Pour ν, de manière très similaire, nous obtenons :

$$\nu|\text{rest} \propto Beta(\gamma + N_+, \zeta + N - N_+)$$

où $N_+ = \sum_{i=1}^{N} \delta_{t_i-1}$.

Exemple

- Commençons par $\pi^{(k)}$:

$$\pi^{(k)}|\text{rest} \propto (\pi^{(k)})^{\alpha+N_+^{(k)}-1}(1-\pi^{(k)})^{\beta+N-N_+^{(k)}-1}$$

  avec $N_+^{(k)} = \sum_{i=1}^{N} \delta_{c_i^{(k)}-t_i}$ où δ_x vaut 0 partout sauf quand $x = 1$ (mesure de Dirac). On retrouve la loi Beta de paramètres $\alpha + N_+^{(k)}$ et $\beta + N - N_+^{(k)}$.
- Pour ν, de manière très similaire, nous obtenons :

$$\nu|\text{rest} \propto Beta(\gamma + N_+, \zeta + N - N_+)$$

  où $N_+ = \sum_{i=1}^{N} \delta_{t_i-1}$.
- Enfin, t_i est une catégorielle :

Exemple

- Commençons par $\pi^{(k)}$:

$$\pi^{(k)}|\text{rest} \propto (\pi^{(k)})^{\alpha+N_+^{(k)}-1}(1-\pi^{(k)})^{\beta+N-N_+^{(k)}-1}$$

avec $N_+^{(k)} = \sum_{i=1}^{N} \delta_{c_i^{(k)}-t_i}$ où δ_x vaut 0 partout sauf quand $x = 1$ (mesure de Dirac). On retrouve la loi Beta de paramètres $\alpha + N_+^{(k)}$ et $\beta + N - N_+^{(k)}$.

- Pour ν, de manière très similaire, nous obtenons :

$$\nu|\text{rest} \propto Beta(\gamma + N_+, \zeta + N - N_+)$$

où $N_+ = \sum_{i=1}^{N} \delta_{t_i-1}$.

- Enfin, t_i est une catégorielle :

$$t_i = j|\text{rest} \propto \nu^j(1-\nu)^{1-j} \prod_{k=1}^{K} (\pi^{(k)})^{\delta_{c_i^{(k)}-j}}(1-\pi^{(k)})^{1-\delta_{c_i^{(k)}-j}}$$

Hamiltonian Monte Carlo (HMC)

Plan

1 Rappels

2 L'approche Bayésienne

3 Les méthodes MCMC

4 Les méthodes approchées
- Motivation
- L'inférence variationnelle

Motivation

- Les méthodes MCMC sont très couteuses en temps de calcul...
- On aimerait jouer avec la forme analytique d'une loi...
- ...

L'inférence variationnelle

- Ici, nous nous appuierons sur la notion de divergence et en particulier la divergence de Kullback et Leibler :

$$D_{KL}(q||p) = \int q(Z) \ln \frac{q(Z)}{p(Z|X)} dZ$$

  où p est notre loi *a posteriori* et q la distribution que l'on manipulera en pratique (p étant intractable).
- L'objectif est de trouver q telle que la divergence est minimisée

L'inférence variationnelle

- Une reformulation de la divergence KL donne :

$$D_{KL}(q||p) = \int q(Z) \ln \frac{q(Z)}{p(Z|X)} dZ$$
$$= \int q(Z) \ln \frac{q(Z)}{p(Z,X)} dZ + \ln p(X) = \mathcal{L}(q) + \ln p(X)$$

où $\mathcal{L}(q)$ est appelé l'énergie variationnelle libre...

L'inférence variationnelle

- Une reformulation de la divergence KL donne :

$$D_{KL}(q||p) = \int q(Z)\ln\frac{q(Z)}{p(Z|X)}dZ$$
$$= \int q(Z)\ln\frac{q(Z)}{p(Z,X)}dZ + \ln p(X) = \mathcal{L}(q) + \ln p(X)$$

  où $\mathcal{L}(q)$ est appelé l'énergie variationnelle libre...

- L'expression précédente est équivalente à

$$\ln p(X) = D_{KL}(q||p) - \mathcal{L}(q)$$

L'inférence variationnelle

$$\ln p(X) = D_{KL}(q||p) - \mathcal{L}(q)$$

- $\ln p(X)$ est une constante... Minimiser $\mathcal{L}(q)$ minimisera $D_{KL}(q||p)$ par symétrie...
- Solution :

L'inférence variationnelle

$$\ln p(X) = D_{KL}(q||p) - \mathcal{L}(q)$$

- $\ln p(X)$ est une constante... Minimiser $\mathcal{L}(q)$ minimisera $D_{KL}(q||p)$ par symétrie...
- Solution :
  - $q = p$!

L'inférence variationnelle

$$\ln p(X) = D_{KL}(q||p) - \mathcal{L}(q)$$

- $\ln p(X)$ est une constante... Minimiser $\mathcal{L}(q)$ minimisera $D_{KL}(q||p)$ par symétrie...
- Solution :
  - $q = p$!
  - q doit être restreint à une classe de distributions plus simples que p !

L'inférence variationnelle

$$\ln p(X) = D_{KL}(q||p) - \mathcal{L}(q)$$

- $\ln p(X)$ est une constante... Minimiser $\mathcal{L}(q)$ minimisera $D_{KL}(q||p)$ par symétrie...
- Solution :
  - $q = p$!
  - q doit être restreint à une classe de distributions plus simples que p !
- Il n'existe pas qu'une manière de restreindre la classe de q (e.g. fixer une distribution à q),

L'inférence variationnelle

$$\ln p(X) = D_{KL}(q||p) - \mathcal{L}(q)$$

- $\ln p(X)$ est une constante... Minimiser $\mathcal{L}(q)$ minimisera $D_{KL}(q||p)$ par symétrie...
- Solution :
  - $q = p$!
  - q doit être restreint à une classe de distributions plus simples que p !
- Il n'existe pas qu'une manière de restreindre la classe de q (e.g. fixer une distribution à q),
- nous allons chercher q parmi les distributions dont les paramètres se factorisent (mean field theory) :

$$q(Z) = \prod_{i=1}^{M} q(Z_i)$$

L'inférence variationnelle

- La solution sera alors de la forme :

$$\ln q(Z_i) = \mathbb{E}_{j \neq i}\left[\ln p(Z, X)\right] + \text{const}$$

où de manière similaire :

$$q(Z_i) = \frac{\exp(\mathbb{E}_{j \neq i}\left[\ln p(Z, X)\right])}{\int \exp(\mathbb{E}_{j \neq k}\left[\ln p(Z, X)\right] dZ_k)}$$

- Si les paramètres de p sont déjà factorisables, alors q retombera sur p,
- chaque $q(Z_i)$ dépendra des autres Z_i. Cette dépendance circulaire définit l'algorithme d'inférence (de type EM).

Exemple

- Construisons un système de vote binaire,
- chaque individu peut voter sur un ensemble de N tâches. Le vote d'un individu k sur la tâche i est noté : $c_i^{(k)}$,
- on suppose que la loi de $c_i^{(k)}$ est une Bernoulli dont les paramètres dépendent de l'individu qui a voté et on choisit un prior conjugué :

$$\mathbb{P}(c_i^{(k)} = t_i | t_i) = \pi^{(k)},\ \pi^{(k)} \sim \text{Beta}(\alpha, \beta)$$

- on suppose que chaque tâche est associée à une "vérité" que l'on cherche à découvrir. Celle-ci est notée t_i pour la i^e tâche :

$$t_i \sim Be(\nu),\ \nu \sim \text{Beta}(\gamma, \zeta)$$

Exemple

- La loi jointe ainsi obtenue est :

$$\mathcal{L}((\mathcal{D}, \Pi, \nu, T)) = \prod_{i=1}^{N} \left\{ \nu^{t_i} (1-\nu)^{1-t_i} \prod_{k=1}^{K} \left[\left(\pi^{(k)}\right)^{\delta_{t_i - c_i^{(k)}}} \left(1-\pi^{(k)}\right)^{1-\delta_{t_i - c_i^{(k)}}} \right] \right\}$$
$$\prod_{k=1}^{K} \left[\left(\pi^{(k)}\right)^{\alpha-1} \left(1-\pi^{(k)}\right)^{\beta-1} \right] \nu^{\gamma-1} (1-\nu)^{\zeta-1}$$

Exemple

- Commençons par t_i. Soit une variable temporaire ρ_{ij} :

$$\begin{aligned}\ln\rho_{ij} =& \mathbb{E}_{\Pi,\nu}\left[\mathcal{L}((\mathcal{D},\Pi,\nu,T))\right]\\ =& \sum_{i=1}^{N} j \times \mathbb{E}_{\nu}\left[\ln\nu\right] + (1-j)\times\mathbb{E}_{\nu}\left[\ln(1-\nu)\right] + \sum_{k=1}^{K}\mathbb{1}[j=c_i^{k)}]\\ &\times \mathbb{E}_{\pi^{(k)}}\left[\ln\pi^{(k)}\right] + \mathbb{1}[j\neq c_i^{(k)}]\times\mathbb{E}_{\pi^{(k)}}\left[\ln(1-\pi^{(k)})\right]\end{aligned}$$

On obtient donc :

$$q(t_i=j) = \mathbb{E}\left[\mathbb{1}[t_i=j]\right] = \frac{\rho_{ij}}{\sum_{\tau=1}^{N}\rho_{i\tau}}, j\in\{0,1\}$$

Exemple

- Définissons des variables qui faciliteront les notations par la suite :

$$N_j = \sum_{i=1}^{N} \mathbb{E}_t \left[\mathbb{1}[t_i = j] \right], j \in \{0, 1\}$$

$$N_k^+ = \sum_{i=1}^{N} \mathbb{E}_t \left[\mathbb{1}[t_i = c_i^{(k)}] \right], j, l \in \{0, 1\}$$

$$N_k^- = \sum_{i=1}^{N} \mathbb{E}_t \left[\mathbb{1}[t_i \neq c_i^{(k)}] \right], j, l \in \{0, 1\}$$

Exemple

- Occupons-nous maintenant de ν :

$$\begin{aligned}\ln q(\nu) &= \sum_{i=1}^{N} \mathbb{E}_t\Big[\mathbb{1}[t_i = 1]\Big]\ln\nu + \mathbb{E}_t\Big[\mathbb{1}[t_i \neq 1]\Big]\ln(1-\nu) \\ &+ (\gamma - 1)\ln\nu + (\zeta - 1)\ln(1-\nu) + \text{const} \\ &= (\gamma + N_1 - 1)\ln\nu + (N_0 + \zeta - 1)\ln(1-\nu) + c\end{aligned}$$

- soit une loi Beta de paramètres $\gamma + N_1$ et $\zeta + N_0$,
- de plus, la loi Beta donne :

$$\begin{aligned}\mathbb{E}[\ln\nu] &= \psi(\gamma + N_1) - \psi(\gamma + \zeta + N) \\ \mathbb{E}[\ln(1-\nu)] &= \psi(\zeta + N_0) - \psi(\gamma + \zeta + N)\end{aligned}$$

où ψ est la fonction digamma.

Exemple

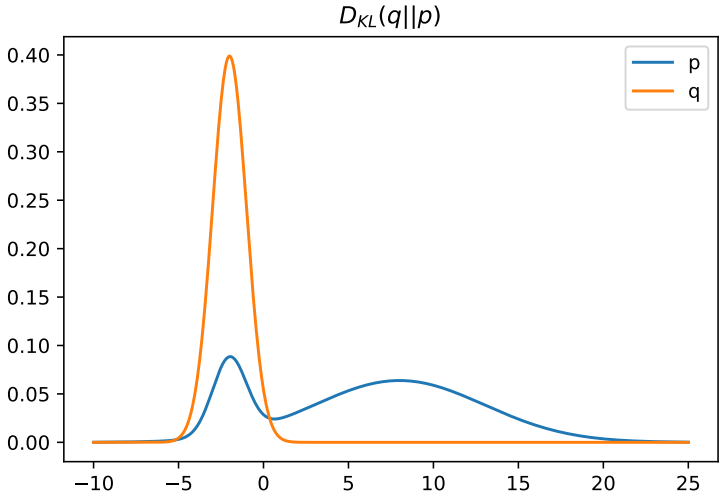
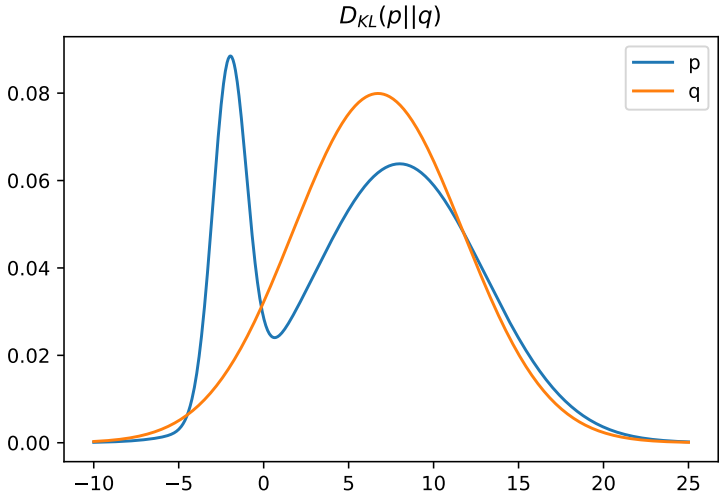
- De la même manière, on obtient que $\pi^{(k)}$ est une loi Beta de paramètres $\alpha + N_k^+$ et $\beta + N_k^-$ avec :

$$\begin{aligned}\mathbb{E}\left[\ln\pi^{(k)}\right] =& \psi(\alpha + N_k^+) - \psi(\alpha + N_k^+ + \beta + N_k^-)\\ \mathbb{E}\left[\ln(1-\pi^{(k)})\right] =& \psi(\beta + N_k^-) - \psi(\alpha + N_k^+ + \beta + N_k^-)\end{aligned}$$

Exemple

random initialization of Π, ν and T;
for *it* ***in*** *nb_iterations* **do**
  evaluate $q(T)$;
  evaluate $q(\Pi)$;
  evaluate $q(\nu)$;
end

Algorithm 3: Variationnal EM

$D_{KL}(q||p)$ versus $D_{KL}(p||q)$

- q est une loi normale, p est un mélange Gaussien, et on choisit q telle que la divergence est minimisée...