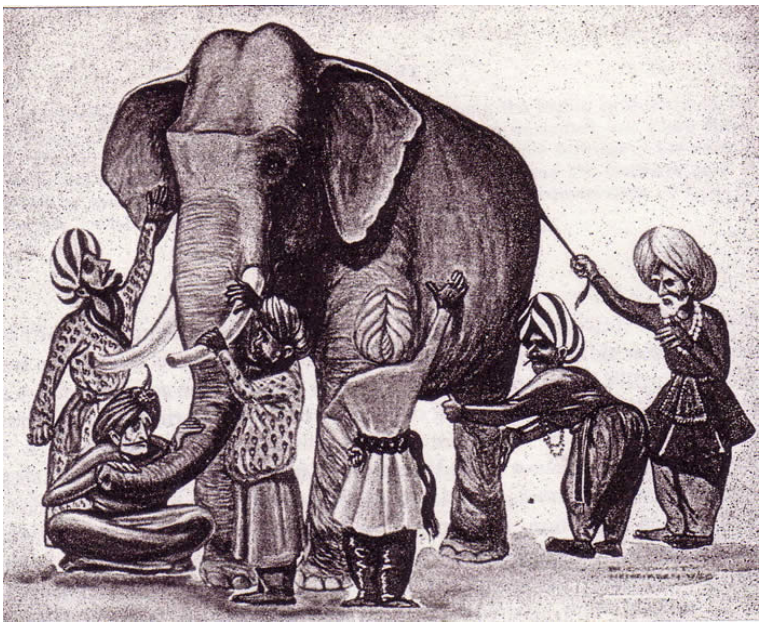
Modèles Probabilistes pour l'Apprentissage

Olivier François

Ensimag – Grenoble INP

olivier.francois@grenoble-inp.fr



The blind men and the elephant

It was six men of Indostan

To learning much inclined

Who went to see the Elephant

(Though all of them were blind)

That each by observation

Might satisfy his mind

...

And so these men of Indostan

Disputed loud and long,

Each in his own opinion

Exceeding stiff and strong,

Though each was partly in the right,

And all were in the wrong!

J.G. Saxe (1816 – 1887)

Brève présentation du cours de MPA

Créer des machines capables d'autonomie, capables d'apprendre à réaliser des tâches par elles-mêmes sans le contrôle de l'homme, est un objectif de longue haleine des sciences numériques. Descartes qui voyait les animaux comme des assemblages de mécanismes, prétendait que l'on pourra un jour créer une machine indistinguable d'un animal, capable de remplacer l'homme pour certaines de ses activités.

La vision de Descartes a probablement conduit au développement des sciences de l'automatisme. Les automates, tels que le fameux *canard de Vaucanson* (figure 1), ne sont toutefois pas dotés de mécanismes adaptatifs. Une fois programmé pour une tâche donnée, ils sont limités à répéter cette tâche. Ce que l'on entend dans ce cours par le mot *apprentissage*, représente un ensemble de méthodes et d'algorithmes qui permettent à une machine d'évoluer dans un monde incertain grâce à un processus adaptatif.

Le principe fondateur de ce cours est que la notion d'apprentissage peut être conceptualisée à l'aide d'un modèle probabiliste. Nous cherchons à modéliser la réponse d'un système à partir de variables d'entrées et des données empiriques aléatoires recueillies pour ce système (nous considérons des versions souvent très simplifiées de tels systèmes dans le cours). Les connaissances accumulées grâce à l'expérimentation permettent de réviser et de mettre à jour un modèle probabiliste conçu *a priori* afin d'en réduire l'incertitude initiale et d'en augmenter le pouvoir prédictif. Les algorithmes permettant la mise à jour du modèle sont adaptatifs et flexibles. Ils sont sensés prendre en compte l'évolution de la base d'information des comportements enregistrés. En pratique, la phase de révision

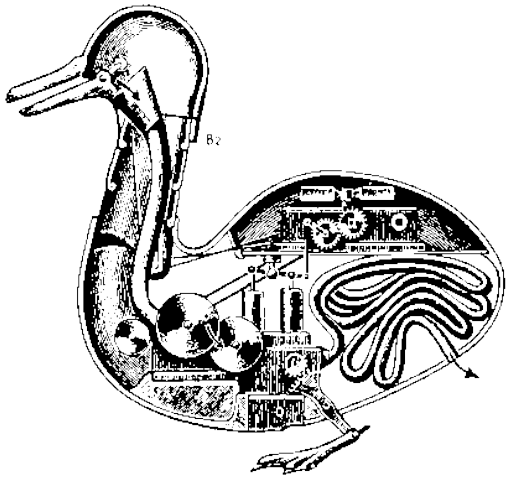


FIGURE 1 – *Le canard de Vaucanson, célèbre automate dont une copie est visible au musée dauphinois à Grenoble.*

du modèle utilise la célèbre formule d'inversion des causes et des effets – la **formule de Bayes** – un outil central de ce cours.

Les applications concrètes du formalisme présenté dans ce cours sont très nombreuses, en reconnaissance des formes, en reconnaissance vocale, en vision par ordinateur, pour l'aide au diagnostic, pour la détection de fraude, d'anomalies, de spams, pour l' analyse financière, pour la bio-informatique. On retrouve ces concepts au cœur de la conception de sites web adaptatifs qui pratiquent la recommandation individualisée de produits, ou encore dans les moteurs de recherche.

Quelques mots sont aussi nécessaires pour expliquer l'allégorie présentée dans l'image de couverture de ce cours. La figure de couverture illustre une légende hindoue, traduite en anglais par le poète John Saxe sous le titre *The blind men and the elephant*. Dans cette légende, six sages, tous aveugles, décident

d'apprendre ce qu'est un éléphant – un gros animal, servant par ailleurs de mascotte à l'Ensimag. Chacun des sages ne peut accéder qu'à une seule partie de l'animal et l'évalue avec son propre système de représentation (un modèle donc). Ainsi, le sage qui tient la queue de l'éléphant prétend qu'il tient une *corde*, celui qui tient la jambe de l'éléphant dit qu'il tient un *tronc*, etc. Les six hommes débattent très fort sur la nature de l'éléphant, mais ne peuvent pas se mettre d'accord. La morale de cette fable dit que chacun des six sages détient une part de vérité, mais tous sont dans l'erreur. Aucun des six sages n'a pu construire une bonne représentation de ce qu'est un éléphant à partir de ses connaissances a priori.

Les élèves intéressés par *aller plus loin* pourront approfondir le cours de MPA – et peut être la philosophie sous-jacente – en étudiant l'ouvrage de référence écrit par Gelman et ses collègues (2004). Chacun aura grand bénéfice à consulter le guide d'initiation au langage `R` écrit par E. Paradis (2005). Pour suivre ce cours, il est en effet nécessaire d'avoir un ordinateur sur lequel on aura installé préalablement le logiciel `R`. Ce logiciel libre est disponible sur le site `http://cran.r-project.org/`. Les scripts `R` illustrant le cours de MPA sont disponibles sur la page personnelle de Olivier François (pages destinées à l'enseignement).

Références citées dans l'introduction :

Gelman A, Carlin JB, Stern HS, Rubin DB (2004) Bayesian Data Analysis 2nd ed. Chapman & Hall, New-York.

E. Paradis (2005) R pour les débutants. Univ. Montpellier II.

1 Séance 1 : Rappels de probabilités – Formule de Bayes

1.1 Rappels et notations

Dans ce cours, toutes les variables, **paramètres ou données** sont considérées comme étant des variables **aléatoires**. Pour alléger les notations, nous ne faisons pas de distinction entre une variable aléatoire et sa réalisation. Ainsi, lorsque x est une variable de Bernoulli susceptible de prendre la valeur 1 avec la probabilité π ou la valeur 0 avec la probabilité $(1-\pi)$, nous écrivons

$$p(x) = \pi^x (1-\pi)^{(1-x)}\,, \quad x = 0, 1\,.$$

Nous venons de décrire une loi de probabilité discrète prenant deux valeurs. La notation se généralise à une loi discrète prenant plus de 2 valeurs sans difficulté. La grandeur $p(x)$ est alors comprise entre 0 et 1 et la somme totale des valeurs est égale à 1. Pour mémoire, la notation traditionnelle distinguant la variable aléatoire X de sa réalisation x consiste à écrire

$$\mathbf{P}(X = x) = \pi^x (1-\pi)^{(1-x)}\,.$$

On choisit donc de ne pas faire mention explicite à X. Lorsque x est une variable continue, prenant par exemple ses valeurs dans $\mathbb{R}$, on parle alors de densité de probabilité. Une densité de probabilité est une fonction positive dont l'intégrale est égale à 1. Par exemple, si x est une variable réelle de loi normale, on écrit alors

$$p(x) = \exp(-x^2/2)/\sqrt{2\pi}\,, \quad x \in \mathbb{R}.$$

Supposons que y est une seconde variable aléatoire, et qu'elle est positive de loi exponentielle de paramètre 1. Nous notons alors sa loi de probabilité de la

manière suivante

$$p(y) = \exp(-y)\,, \quad y > 0.$$

Pour mémoire, la notation traditionnelle distinguant la variable aléatoire Y de sa réalisation y consiste à écrire

$$f_Y(y) = \exp(-y)\,, \quad y > 0.$$

Il est important de remarquer que la notation simplifiée, supprimant la référence la variable Y, peut être ambigüe. C'est le contexte et la notation en lettre minuscule associée à une variable qui permet de comprendre de quelle loi il s'agit. Il est clair que $p(x)$ (égale à $f_X(x)$) et $p(y)$ (égale à $f_Y(y)$) ne désignent pas la même densité, bien que la notation mathématique, $p(.)$, semble l'indiquer.

La loi exponentielle de paramètre $\theta > 0$ se note de la manière suivante

$$p(y|\theta) = \theta \exp(-\theta y)\,, \quad y > 0.$$

La barre | introduite dans la notation précédente n'est pas anodine. Puisque θ est un paramètre, nous le considérons comme étant une variable aléatoire. Ainsi, y peut être vue comme une réalisation d'une loi conditionnelle sachant le paramètre θ. Nous y reviendrons dans la section suivante.

En définitive, nous considérons de manière unifiée des variables aléatoires continues ou discrètes en notant les lois de la même manière. Cela signifie que l'on utilise la même notation pour désigner des intégrales et des sommes (c'est tout à fait rigoureux !). Par exemple, pour une variable aléatoire continue, nous avons

$$\int p(y)dy = 1\,.$$

Pour une variable y discrète, prenant un nombre fini de valeurs, il faut comprendre (ou réécrire) cette intégrale comme étant égale à la somme

$$\sum p(y) = 1\,.$$

Dans le cas discret ou dans le cas continu, nous appelons **espérance** de la variable y, la grandeur

$$\mathrm{E}[y] = \int yp(y)dy\,.$$

Lorsque cette intégrale est définie, on dit que la variable aléatoire y est intégrable. Cela correspond à la situation où l'intégrale de la valeur absolue de la variable y converge, c'est à dire, $\mathrm{E}[|y|] < \infty$. Lorsque y est de carré intégrable, nous appelons **variance** la grandeur définie par

$$\mathrm{Var}[y] = \mathrm{E}[(y - \mathrm{E}[y])^2] = \mathrm{E}[y^2] - \mathrm{E}[y]^2\,.$$

1.2 Loi de couple et probabilité conditionnelle

Nous considérons maintenant deux variables aléatoires, y et θ, discrètes ou continues. Groupées, ces deux variables forment le couple (y, θ). Il est commode d'imaginer que y représente une donnée issue d'un tirage aléatoire et que θ représente un paramètre déterminant ce tirage. Dans nos notations, la loi jointe du couple (y, θ) s'écrit $p(y, \theta)$. Il s'agit en général d'une fonction positive et nous avons

$$\int p(y, \theta)dyd\theta = 1.$$

On parle alors de modèle probabiliste pour le couple (y, θ). Dans ce cours, la loi de la donnée y s'appelle la **loi marginale**. Elle est définie à l'aide de l'intégrale suivante

$$p(y) = \int p(y, \theta)d\theta\,,$$

dont la valeur exacte est souvent difficile à expliciter. La loi $p(\theta)$ est théoriquement aussi une loi marginale. Nous l'appelons plus communément la **loi a priori**. Elle est en général donnée directement lors de la description du modèle.

La loi conditionnelle de y sachant θ est notée $p(y|\theta)$

$$p(y|\theta) = \frac{p(y,\theta)}{p(\theta)}\,.$$

Cette loi est appelée la **loi d'échantillonnage** ou la **loi générative**. Elle décrit la manière avec laquelle, sachant la valeur du paramètre θ, on peut échantillonner une variable y (on dit aussi générer y).

Nous voyons avec la loi conditionnelle, un intérêt évident à utiliser les notations allégées. En effet, une notation rigoureuse nous demanderait d'écrire

$$\mathbf{P}(Y = y|\Theta = \theta) = p(y|\theta)$$

dans le cas ou y est une variable discrète et

$$f_Y^{\Theta=\theta}(y) = p(y|\theta)$$

dans le cas d'une variable continue. L'utilisation d'indices et d'exposants peut être source de confusion et nous l'abandonnons donc par la suite pour utiliser la notation allégée.

Important : la donnée de la loi a priori, $p(\theta)$ et de la loi générative $p(y|\theta)$ permet de définir un modèle probabiliste. En effet, nous avons

$$p(y,\theta) = p(y|\theta)p(\theta)\,.$$

Par exemple, considérons une probabilité inconnue, θ, répartie de manière uniforme sur l'intervalle $(0,1)$, et supposons que l'on tire un nombre au hasard

selon une loi binomiale de paramètre $n = 20$ et de probabilité θ. Dans ce cas, nous avons

$$p(y, \theta) = \frac{n!}{y!(n-y)!} \theta^y (1-\theta)^{n-y} .$$

pour tout $0 \leq y \leq n$ et $\theta \in (0, 1)$.

1.3 Formule de Bayes

La **formule de Bayes** est une **formule essentielle** du cours de MPA. Elle permet de calculer la loi de probabilité conditionnelle de θ sachant y à partir de la donnée de la loi générative. Pour la distinguer de la loi a priori, cette loi s'appelle la **loi a posteriori**

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} .$$

Dans cette formule, la loi marginale $p(y)$ peut être difficile à exprimer de manière explicite. Nous notons qu'elle correspond en fait à la constante de normalisation de l'expression du numérateur

$$p(y) = \int p(y|\theta)p(\theta)d\theta .$$

On se contentera donc, sauf mention contraire, du terme général de la loi a posteriori

$$p(\theta|y) \propto p(y|\theta)p(\theta) ,$$

où le symbole $\propto$ signifie *proportionnel à*. La formule ci-dessus fait clairement apparaître le lien existant entre la loi a priori et la loi a posteriori. La loi a posteriori peut être vue comme une **mise à jour** de la loi a priori, une fois que la variable y est générée et observée.

Par exemple, considérons une proportion inconnue, θ, répartie de manière uniforme sur l'intervalle $(0, 1)$, et supposons que l'on tire une valeur au hasard, 0 ou 1, telle que la probabilité d'observer 1 est égale à θ. A l'issue de cette expérience, supposons que le résultat soit égal à 1. Nous avons

$$p(\theta|y = 1) \propto p(y = 1|\theta)p(\theta) = \theta p(\theta) = \theta$$

Dans ce cas, la constante de normalisation est en fait très facile à calculer. Nous obtenons

$$p(\theta|y = 1) = 2\theta, \quad \theta \in (0, 1)\,.$$

La loi **a priori** décrit l'incertitude sur le paramètre θ avant l'expérience, et la loi **a posteriori** décrit l'incertitude sur ce paramètre à l'issue de l'expérience. La formule de Bayes formalise le processus d'**apprentissage** correspondant à la mise à jour de l'information en fonction du résultat de l'expérience.

1.4 Principes de base de simulation

Une difficulté rencontrée en calcul des probabilités est que de nombreuses lois ne sont pas calculables explicitement. En revanche, il est souvent possible de les simuler à l'aide d'un générateur aléatoire afin de calculer les probabilités ou les espérances (valeurs moyennes de certaines fonctions) numériquement. Cette méthode est communément appelée la **méthode de Monte Carlo**, en référence aux tirages aléatoires qu'elle effectue.

Simulation par inversion. Supposons que l'on cherche à simuler une variable aléatoire θ, de loi donnée $p(\theta)$ (non uniforme). Lorsque cela est possible,

on calcule l'inverse, G, de la fonction de répartition, F, définie par

$$F(t) = p(\theta \leq t), \quad t \in \mathbb{R}\,.$$

La méthode de simulation par inversion consiste à tirer un nombre au hasard

$$\theta = G(u)\,,$$

où u est un tirage de loi uniforme sur (0,1). Formellement séduisante, cette méthode reste toutefois difficilement applicable en pratique, sauf pour des cas simples. Pour utiliser la méthode d'inversion, on peut chercher à décomposer la loi cible comme un mélange de lois très simples, afin d'appliquer la méthode à chaque composante du mélange.

Simulation par rejet. La méthode de simulation par rejet ne demande que très peu de calcul analytique. Il en existe de nombreuses variantes fort utiles dans ce cours. Dans ce paragraphe, nous rappelons l'algorithme vu en première année de l'Ensimag lors du cours de probabilités appliquées. Soit $p(\theta)$ la loi de probabilité d'une variable définie sur l'intervalle $(0, 1)$ supposée continue sur cet intervalle, et c une constante supérieure au maximum de $p(\theta)$. L'algorithme de rejet peut s'écrire de la manière suivante.

```
Repeat
theta <- unif(0,1)
x <- unif(0,1)
Until (c * x < p(theta) )
return(theta)
```

Simuler un couple de variables aléatoires. Pour simuler un couple de variables aléatoires (y, θ) de loi $p(y, \theta)$, une méthode élémentaire consiste à simuler la loi a priori, $p(\theta)$, puis à simuler la loi générative $p(y|\theta)$. Dans ce cas, on se ramène donc à la simulation de variables uni-dimensionnelles. Nous verrons dans la suite de ce cours de nouvelles méthodes de simulation de lois multivariées.

1.5 Résumé

Résumer les points à retenir et donner quelques exemples illustrant les concepts principaux de la séance.

1.6 Exercices

Pour les exercices suivants, il est nécessaire d'avoir un ordinateur sur lequel on aura installé préalablement le logiciel `R`. Ce logiciel libre est disponible sur le site `http://cran.r-project.org/`.

Exercice 1. On considère la loi de densité

$$p(\theta) \propto \theta + 2\theta^4\,, \quad \theta \in (0,1).$$

1. Calculer la constante de proportionnalité.
2. Proposer un algorithme de simulation par rejet par rapport à la loi uniforme pour la loi $p(\theta)$.
3. Evaluer le nombre moyen de rejets nécessaires pour obtenir un seul tirage.
4. Proposer un algorithme de simulation de mélange pour la loi $p(\theta)$.
5. Ecrire les algorithmes précédents dans le langage de programmation `R`.
6. Simuler un échantillon de taille 10000 à l'aide de chacun des algorithmes.
7. Vérifier que les histogrammes estiment bien la densité $p(\theta)$, et pour l' algorithme de rejet, estimer la probabilité de rejet numériquement.

Exercice 2. Couples de variables aléatoires. On considère la loi de densité

$$\forall y, \theta \in \mathbb{R}, \quad p(y,\theta) = \frac{1}{2\sqrt{y\theta}}\, \mathbb{1}_D(y,\theta)\,,$$

où $D = \{(x,\theta) \in \mathbb{R}^2,\, 0 < y < \theta < 1\}$.

1. Rappeler le principe de simulation d'une loi de densité $p(y,\theta)$ définie sur $\mathbb{R}^2$.
2. Calculer les lois marginale, $p(y)$, et a priori, $p(\theta)$, du couple de densité $p(y,\theta)$.

3. Lorsqu'elles sont définies, calculer les lois conditionnelles $p(y|\theta)$, $p(\theta|y)$ du couple (y, θ).
4. Proposer un algorithme de simulation pour la loi de densité $p(y, \theta)$ et prouver sa validité en appliquant le théorème de changement de variables.
5. Programmer cet algorithme dans le langage `R`.
6. Effectuer 10000 tirages selon cet algorithme. Afficher les résultats.
7. Proposer des approximations de Monte Carlo des grandeurs $\mathrm{E}[y]$ et ρ (le coefficient de corrélation). Calculer les valeurs théoriques de ces grandeurs.

Exercice 3. On considère la loi définie de la manière suivante

$$p(\theta) \propto \theta^3(1-\theta)^5, \quad \theta \in (0,1),$$

dont un échantillon de taille m peut être créé en utilisant le générateur aléatoire `rbeta(m, 4, 6)`.

1. Déterminer le terme général de la loi conditionnelle de la variable θ sachant que θ est supérieur à $1/2$.
2. Ecrire en langage `R` un algorithme de simulation de cette loi et en représenter un histogramme.
3. On considère la variable $\varphi = 2\theta - 1$, où θ résulte de la simulation précédente ($\theta \geq 1/2$). Représenter un histogramme de la variable φ et montrer que

$$p(\varphi) \propto (1+\varphi)^3(1-\varphi)^5, \quad \varphi \in (0,1).$$

Exercice 4. Soit $\theta \in (0,1)$, $n \in \mathbb{N}^*$ et (y,z) un couple de variables aléatoires telles que la loi marginale $p(z|\theta)$ est égale à

$$p(z|\theta) = \binom{n}{z}\theta^z(1-\theta)^{n-z}, \quad z = 0,\dots,n.$$

On suppose de plus que la loi conditionnelle $p(y|z,\theta)$ est égale à

$$p(y|z,\theta) = \binom{n-z}{y-z}\theta^{y-z}(1-\theta)^{n-y}, \quad y = z,\dots,n.$$

1. Montrer que la loi conditionnelle $p(z|\theta,y)$ est égale à

$$p(z|\theta,y) = \binom{y}{z}\frac{(1-\theta)^{y-z}}{(2-\theta)^y}, \quad z = 0,\dots,y.$$

2. Application. Ecrire en langage `R` un algorithme de simulation de la loi décrite ci-dessus pour $n = 100$, $y = 62$ et $\theta = 0.3$. Produire un histogramme des données simulées et vérifier l'adéquation de la loi théorique à cet histogramme.

2 Séance 2 : Quantifier l'incertitude et effectuer des prédictions à partir de résultats d'expériences

2.1 Introduction

Dans ce cours, nous appelons **processus d'apprentissage** le processus consistant à mettre à jour l'information dont on dispose sur un phénomène à partir de l'observation répétée ou non de ce phénomène. Mathématiquement, l'information disponible sur un phénomène est généralement résumée par un ensemble de paramètres permettant d'expliquer le phénomène.

Afin de quantifier et mettre à jour l'incertitude sur les paramètres des modèles, les modèles considérés dans ce cours sont probabilistes (au sens large, car les modèles déterministes peuvent être un cas particulier intéressant). L'approche probabiliste permet aussi de prédire les nouvelles données, d'évaluer l'incertitude des prédictions et d'évaluer la pertinence des modèles.

L'idée d'apprentissage probabiliste se traduit par l'utilisation systématique de la formule de Bayes. Cette formule traduit comment l'incertitude a priori sur les paramètres d'un modèle se réduit lorsque des données supplémentaires alimentent le modèle. L'utilisation de la formule de Bayes pour le calcul de la loi a posteriori justifie que l'on parle traditionnellement d'**inférence bayésienne**. Le terme *inférence* est utilisé ici comme un terme générique mettant en exergue la nature probabiliste des résultats obtenus et la possibilité d'utiliser les lois a posteriori dans un but prédictif.

2.2 Définitions

Nous reprenons et généralisons les notations introduites dans la séance précédente. Dans un modèle, il y a des **variables observées** et des **variables non-observées**. Les variables observées sont les **données** disponibles sur le phénomène à expliquer ou à prédire

$$y = (y_1, \ldots, y_n), \quad n \geq 1.$$

Les données forment un vecteur de longueur n, que l'on appelle l'**échantillon**. Pour cette raison la loi générative des données s'appelle la loi d'échantillonnage. Plus généralement, échantillonner, c'est tirer ou recueillir des données au hasard.

Les variables non-observées sont appelées **variables cachées**, **variables latentes** ou **paramètres**

$$\theta = (\theta_1, \ldots, \theta_J), \quad J \geq 1.$$

L'ensemble des variables non-observées constitue un vecteur de dimension J (J peut être plus grand que n). Les composantes de ce vecteur peuvent être des paramètres probabilistes tels que la moyenne ou la variance, mais aussi des données manquantes dont l'introduction "augmente" le modèle.

Un modèle est donc décrit par une loi de probabilité multivariée, autrement dit, multi-dimensionnelle :

$$p(y, \theta) = p(y|\theta)p(\theta)\,,$$

Nous avons déjà parlé de la loi du paramètre, notée $p(\theta)$. Cette loi s'appelle la loi *a priori*. Elle peut être choisie afin de modéliser l'incertitude initiale sur un phénomène, ou bien les connaissances d'un expert n'ayant pas accès aux données y.

Important : La loi a priori peut parfois être **dégénérée**. Dans ce cas la fonction $p(\theta)$ n'est pas intégrable

$$\int p(\theta) d\theta = \infty \,.$$

En pratique, cela peut être acceptable si la loi *a posteriori*

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}$$

est bien définie et vérifie

$$\int p(\theta|y) d\theta = 1 \,.$$

2.3 Apprentissage d'une fréquence d'émission

Nous considérons un exemple très élémentaire d'apprentissage probabiliste : l'apprentissage de la fréquence d'émission d'une source à partir d'observations émises par la source. Une source émet donc des signaux binaires, $x_i = 0$ ou 1, et on cherche à estimer la probabilité pour cette source d'émettre un 1. On note cette probabilité θ. Sans information préalable sur la source, nous supposons que la loi a priori est uniforme

$$p(\theta) = 1 \,.$$

La théorie de Shannon nous indique que la loi uniforme maximise l'incertitude sur le paramètre θ. Le modèle est entièrement spécifié par la loi générative

$$p(x_i = 1|\theta) = \theta \,, \quad p(x_i = 0|\theta) = 1 - \theta \,,$$

Il s'agit d'une loi de Bernoulli de paramètre θ. Nous supposons que l'on observe une suite de signaux indépendants émis par la source. Par exemple, cette suite est donnée par $x_1 = 1, x_2 = 0, x_3 = 1, \ldots$

Afin de bien comprendre comment l'incertitude sur le paramètre θ se modifie au fur et à mesure que les données sont acquises, nous cherchons à calculer la loi conditionnelle de θ sachant les données. Suite à la première observation, nous avons

$$p(\theta|x_1 = 1) \propto p(x_1 = 1|\theta)p(\theta) = \theta.$$

Suite à la seconde observation, nous avons

$$p(\theta|x_2 = 0, x_1 = 1) \propto p(x_2 = 0|\theta)p(\theta|x_1 = 1) = (1 - \theta)\theta\,.$$

On remarque que l'on peut calculer la loi conditionnelle $p(\theta|x_2 = 0, x_1 = 1)$ directement ou bien, comme ci-dessus, en utilisant la loi $p(\theta|x_1 = 1)$. De la même manière, les lois suivantes peuvent être calculées en utilisant un argument de récurrence simple. Suite à la troisième observation, nous avons

$$p(\theta|x_3 = 1, x_2 = 0, x_1 = 1) \propto p(x_3 = 1|\theta)(1 - \theta)\theta = (1 - \theta)\theta^2\,.$$

Plus généralement, après n observations, nous obtenons

$$p(\theta|x_n, \ldots, x_1) \propto p(x_n|\theta)p(\theta|x_{n-1}, \ldots, x_1) = (1 - \theta)^{n-y}\theta^y$$

où l'on note $y = \sum_{i=1}^{n} x_i$. La figure 2 illustre graphiquement l'évolution de la connaissance acquise ou l'évolution de l'incertitude résiduelle sur la probabilité d'émission θ en fonction de l'observation successive des données suivantes : $x_1 = 1, x_2 = 0$ et $x_3 = 1$.

2.4 Un exemple de sondage : modèle bêta-binomial

L'exemple précédent met en évidence une loi particulière jouant un rôle important lorsque l'on cherche à modéliser une fréquence. Il s'agit de la **loi**

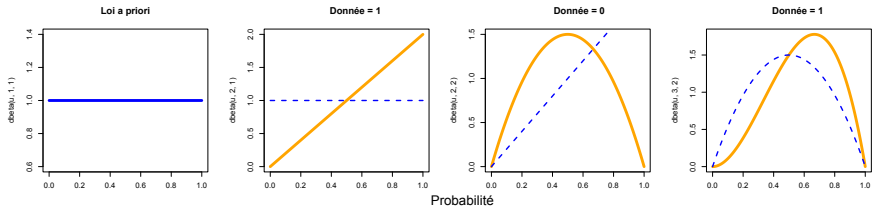


FIGURE 2 – *Mise à jour de l'information a priori sur la fréquence d'émission d'une source binaire en fonction de l'observation successive des données* $x_1 = 1, x_2 = 0$ *et* $x_3 = 1$.

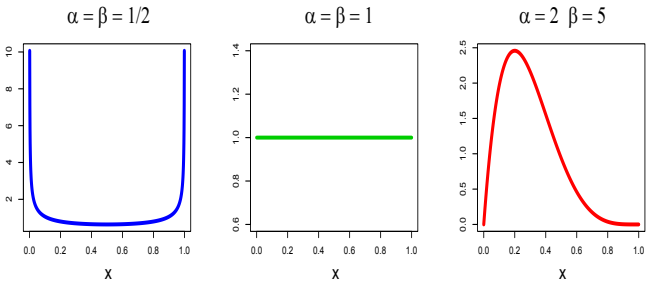


FIGURE 3 – *Densité de probabilité de la loi bêta pour différentes valeurs des paramètres* α *et* β.

bêta. La loi bêta modélise une variable aléatoire comprise entre 0 et 1. Cette loi comporte 2 paramètres positifs, notés α et β. Elle est définie de la manière suivante

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

La figure 3 représente les formes typiques que peut prendre la courbe de la densité de la loi bêta. Pour des valeurs de α et β inférieures à 1, nous obtenons une forme en U. Pour α et β supérieurs à 1, nous obtenons une forme en cloche. Notons que pour $\alpha = \beta = 1$, la loi correspond à la loi uniforme sur $(0, 1)$. La valeur moyenne de la loi bêta se calcule grâce aux propriétés mathématiques de

la fonction $\Gamma(\alpha)$. Nous avons

$$\mathrm{E}[\theta] = \int_0^1 \theta p(\theta) d\theta = \frac{\alpha}{\alpha + \beta} \, .$$

Voyons maintenant comment la loi bêta intervient lorsque l'on cherche à modéliser une fréquence. Pour cela, considérons le problème de sondage – tout à fait fictif – suivant.

Avant le second tour d'une élection présidentielle, il reste deux candidats que nous appelons A et B. Nous interrogeons au hasard $n = 20$ *électeurs supposés représentatifs et indépendants les uns des autres. Le résultat est* $y = 9$ *voix pour A et* $n - y = 11$ *voix pour B. 1) Comment évaluer la probabilité pour que A soit élu ? 2) Si on interroge un* $(n+1)^{\mathrm{e}}$ *électeur, quelle est la probabilité que ce nouvel électeur vote pour A ?*

Commençons par éliminer les raisonnements faux consistant à dire que la probabilité cherchée dans la question 1) est égale à la fréquence empirique, soit 45% (= 9/20), ou nulle car $9/20 < 1/2$. Pour résoudre ce problème et répondre aux deux questions posées, nous introduisons un modèle probabiliste. Nous considérons la proportion totale inconnue, θ, d'électeurs qui voteront pour A lors du second tour de l'élection. La réponse à la première question demande de calculer la probabilité $p(\theta > .5|y)$, qui n'est pas égale à 9/20. Nous supposons pour cela que nous n'avons pas d'information disponible a priori sur la proportion θ. Cette hypothèse n'est certainement pas tout à fait vérifiée, et elle peut conduire à des simplifications du modèle qui pourraient être discutées si nécessaire (six men required). Nous supposons donc que la loi *a priori* est uniforme sur $(0, 1)$. En considérant les n voix de manière simultanée, la loi

générative est la loi binomiale de paramètres n et θ

$$p(y|\theta) = \text{binom}(n, \theta)(y) \propto \theta^y (1-\theta)^{n-y}\,.$$

Le calcul de la loi *a posteriori* est immédiat. Par identification avec la formule donnée pour la loi bêta, nous avons

$$p(\theta|y) = \text{beta}(y+1, n+1-y)(\theta)\,, \quad \theta \in (0,1).$$

Notons que les constantes de normalisation ont été inutiles pour trouver ce résultat. Les termes généraux de la loi générative et de la loi a posteriori permettent seuls l'identification de la loi a posteriori. À l'aide de la fonction de répartition de la loi bêta, calculable en `R` grâce à la commande `pbeta`, nous pouvons donner une valeur numérique précise de la probabilité que le candidat A soit élu au second tour

$$p(\theta > .5|y) \approx 0.331811\,.$$

Cela répond donc à la première question. Suivant les mêmes calculs, nous pouvons donner un intervalle de crédibilité, I, tel que $p(\theta \in I|y) = .95$. Dans ce cas précis, cet intervalle est égal à $I = (0.25, 0.65)$.

Afin de répondre à la seconde question, nous devons introduire un concept appelé loi prédictive a posteriori, que nous décrirons avec plus de détails dans la séance suivante. La loi prédictive correspond à la loi d'une nouvelle donnée connaissant les n données précédemment recueillies. Pour la calculer, on calcule la probabilité pour qu'un $(n+1)^{\text{e}}$ électeur vote pour A sachant y en intégrant sur toutes les valeurs possibles de θ (sachant y)

$$p(\text{vote pour A}|y) = \int_0^1 p(\text{vote pour A}|\theta)p(\theta|y)d\theta = \int_0^1 \theta p(\theta|y)d\theta$$

On reconnaît l'espérance de la loi conditionnelle de θ sachant y

$$p(\text{vote pour A}|y) = \frac{y+1}{n+2} \approx 0.4545\,.$$

Sans surprise, ce résultat est très proche de la fréquence empirique égale à 45%. Nous verrons dans la suite que l'espérance de la loi conditionnelle de θ sachant y représente la meilleure prédiction que l'on peut faire de θ sachant y. Dans ce cas, la prédiction donne A battu. Toutefois, ce résultat est incertain et nous avons pu calculer son incertitude : il y a environ 33% de chance que A remporte l'élection.

2.5 Approche par simulation numérique

Les calculs présentés dans le paragraphe précédent sont séduisants car ils donnent une valeur exacte des probabilités cherchées. En réalité, ils cachent des calculs d'intégrale qui ne sont pas toujours possibles dans les modèles probabilistes. La philosophie de ce cours est de se passer le plus possible du calcul intégral en le remplaçant par des estimations obtenues par simulation numérique. Les méthodes considérées sont appelées des méthodes de Monte Carlo.

Revenons un instant sur le modèle considéré dans le paragraphe précédent. Ce modèle consiste à créer une variable aléatoire θ uniformément répartie sur l'intervalle $(0, 1)$, puis, sachant θ, à tirer une variable aléatoire y selon la loi binomiale de paramètre $n = 20$ et θ. Ces deux opérations sont très faciles à réaliser à l'aide des générateurs aléatoires `runif` et `rbinom` du langage `R`. Les commandes ci-dessous génèrent 10000 simulations de ce modèle.

```
theta <- runif(10000)
y.s <- rbinom(10000, 20, theta)
```

Un résultat correspondant à cette simulation est représenté dans la figure 4. La colonne en vert correspond aux valeurs de la proportion θ qui ont permis de générer exactement $y.s = 9$ voix dans un échantillon de taille $n = 20$. Les valeurs en ordonnée représentent les points simulés selon la loi conditionnelle de θ sachant $y = 9$. En conservant les valeurs de θ correspondant à la colonne de points verts, nous obtenons des tirages de θ effectués selon la loi a posteriori. A partir de ces tirages nous pouvons facilement donner des réponses numériques aux questions posées précédemment. La précision de ces réponses dépend alors du nombres de points présents dans la colonne $y = 9$.

Avec ce procédé de simulation très simple, nous venons en fait de décrire un algorithme de rejet pour simuler la loi a posteriori. En effet, l'algorithme consiste à rejeter les valeurs du couple (y, θ) telles que la valeur y_s simulée est différente de la valeur observée $y = 9$. En langage `R`, nous pouvons écrire la ligne suivante

```
theta.post <- theta[ y.s == y ]
```

De manière générale, nous pouvons prouver que cet algorithme produit bien des simulations selon la loi *a posteriori*. En effet, notons $p_y(\theta)$ la loi de θ à l'issue de la procédure de rejet, nous avons

$$p_y(\theta) = \sum_{s=1}^{\infty} (1 - p(y))^{s-1} p(y, \theta) = p(\theta|y).$$

Cette formule exprime le fait que le temps d'attente, s, nécessaire à l'observation d'une donnée simulée égale à y est de loi géométrique de paramètre $p(y)$. Le résultat découle des propriétés de la série géométrique.

De retour au problème initial, la probabilité pour que le candidat A soit élu

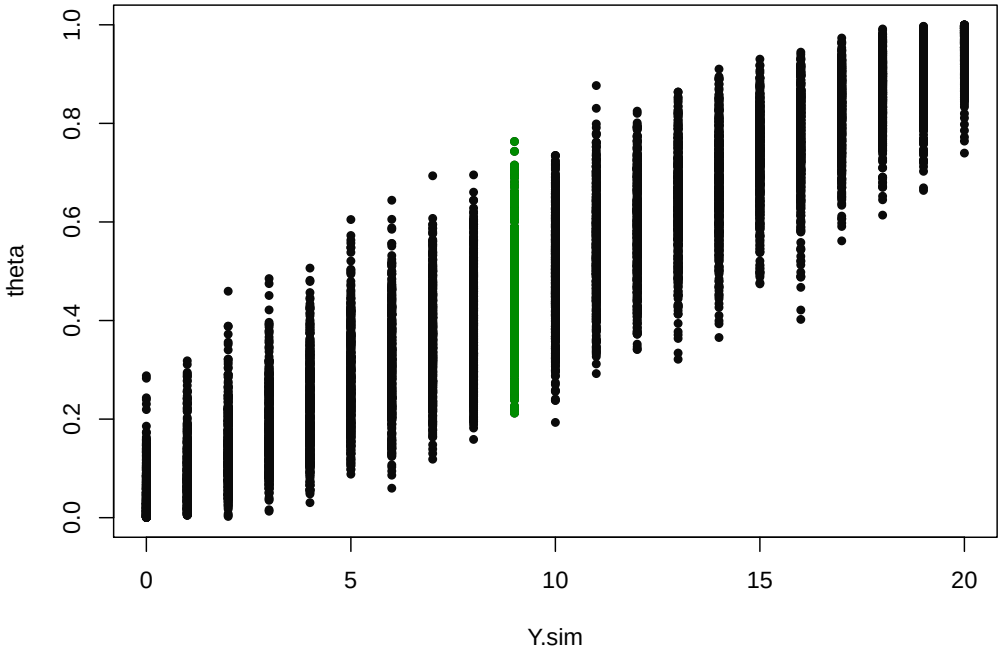


FIGURE 4 – *Simulation de la loi jointe $p(y, \theta)$ pour le modèle binomial ($n = 20$). La colonne en vert représente les points simulés selon la loi conditionnelle de θ sachant $y = 9$. En conservant les valeurs de θ correspondant à ces points, nous obtenons des tirages effectués selon la loi a posteriori.*

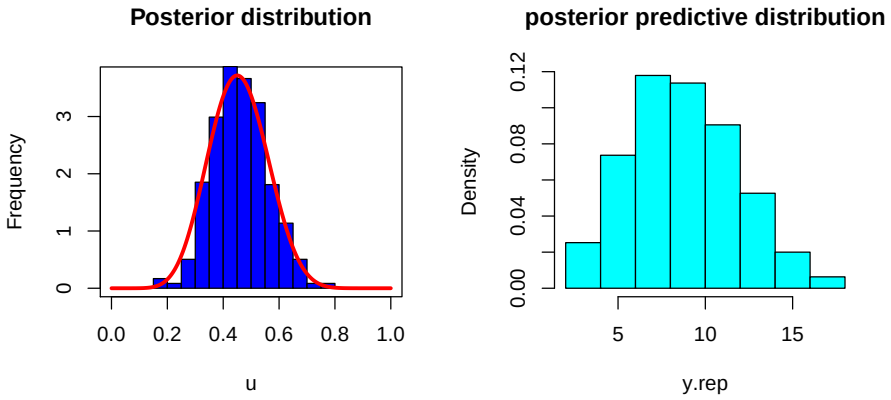


FIGURE 5 – *A gauche : Loi a posteriori obtenue par simulation (en rouge la loi théorique). A droite : Loi prédictive a posteriori pour de nouveaux échantillons.*

au second tour se calcule numériquement à l'aide de la commande `R` suivante :

```
mean(theta.post > .5)
```

Dans notre simulation, la valeur obtenue est autour de 31%. Cette valeur peut être rendue plus précise en augmentant le nombre de simulations. La figure 5 montre l'histogramme de la loi a posteriori obtenu par l'algorithme de rejet. La courbe superposée à cet histogramme correspond à la courbe de la densité théorique, dans ce cas égale à la loi bêta(10,12). Nous voyons que l'approximation numérique est tout à fait satisfaisante.

2.6 Résumé

Résumer les points à retenir et donner quelques exemples illustrant les concepts principaux de la séance.

2.7 Exercices

Pour les exercices suivants, il est nécessaire d'avoir un ordinateur sur lequel on aura installé préalablement le logiciel `R`. Ce logiciel libre est disponible sur le site `http://cran.r-project.org/`.

Exercice 1. Chez les mammifères, les males possèdent un chromosome X et un chromosome Y, tandis que les femelles possèdent deux copies du chromosome X. Chez un individu, chacun des 2 chromosomes est hérité d'un seul des 2 parents.

L'hémophilie est une maladie génétique humaine liée à un allèle récessif localisé sur le chromosome X. Cela signifie qu'une femme portant l'allèle responsable de la maladie sur l'un de ses deux chromosomes n'est pas affectée par la maladie. La maladie est, en revanche, généralement fatale pour une femme qui porterait les deux copies mutées (un événement très rare de toutes façons).

1. Une femme sait que son frère est affecté, mais que son père ne l'est pas. A priori, quelle est la loi de son propre état, θ, que l'on considère comme un paramètre binaire (porteuse ou non de l'allèle récessif) ?
2. On prend maintenant en compte les données suivantes. Cette femme a deux fils (non-jumeaux) et aucun des deux n'est affecté. Quelle est la loi a posteriori du paramètre θ ?

Exercice 2. La loi de l'ouest. A Las Vegas, on rencontre une proportion θ de tricheurs, $0 < \theta < 1$, supposée connue. Lorsque l'on joue contre un tricheur, la probabilité de gagner une partie est nulle, tandis que, lorsque l'on joue contre une personne honnête, cette probabilité est $1/2$. On joue et on perd une partie

à Las Vegas. Quelle est la probabilité d'avoir joué contre un tricheur ?

Exercice 3. La loi de l'ouest (suite). À Las Vegas, il y a une proportion inconnue, θ, de tricheurs. Lorsque l'on joue à *pile* ou *face* contre un tricheur, il est impossible de gagner. Dans le cas contraire, on gagne avec la probabilité 1/2. On joue contre n personnes choisies de manière indépendante, et on perd y parties ($0 < y \leq n$). On suppose que $p(\theta) = 1$.

1. Calculer la probabilité conditionnelle de perdre une partie sachant θ.
2. Montrer que la loi générative est égale, à une constante multiplicative près, à
$$p(y|\theta) \propto \frac{n!}{y!(n-y)!}(1+\theta)^y(1-\theta)^{n-y}.$$
3. On suppose que $y = n$. Soit $t \in (0,1)$. Montrer que l'expression exacte de la probabilité $p(\theta > t|y = n)$ est donnée par la formule suivante
$$p(\theta > t|y = n) = \frac{2^{n+1} - (1+t)^{n+1}}{2^{n+1} - 1}\,.$$
Indication : Déterminer tout d'abord la loi a posteriori $p(\theta|y = n)$.
4. On suppose désormais que y est quelconque. Déterminer à une constante près la densité la loi de la variable $2\varphi - 1$ où φ est une variable aléatoire de loi Bêta$(y+1, n-y+1)$ conditionnée à être supérieure à 1/2. Montrer que cette loi correspond à la loi a posteriori de θ.
5. Ecrire un algorithme de simulation de la loi a posteriori en langage R.

Exercice 4. Tarzan est-il un bon biologiste de terrain ? La loi du célèbre naturaliste Ramon Homstein-Grunberger indique que la taille des girafes pigmées

est uniformément répartie entre 0.5 et 1.5m. Adepte du scepticisme, Tarzan fait une sortie dans la jungle avec son mètre mesureur. Au retour il prétend que les girafes pigmées ne dépassent pas 1.25m. Le problème consiste à estimer le nombre de girafes mesurées par Tarzan, et à donner la probabilité qu'il ait mesuré moins de 10 girafes. Pour cela, on considère que les observations sont des réalisations indépendantes, $(u_i)_{i\geq 1}$, de loi uniforme sur $(0,1)$ (correspondant aux tailles des girafes auxquelles on soustrait 0.5m). Soit $t \in (0,1)$.

1. Soit $k \geq 1$. On pose $y_k = \max_{i=1,\ldots,k} u_i$. Montrer que
$$p(y_k \leq t) = t^k \,.$$
2. Soit $n \geq 1$. On observe le résultat du maximum d'un nombre inconnu et aléatoire, θ, de variables u_i. On suppose que θ est indépendant des u_i et, a priori, de loi uniforme sur $\{1,\ldots,n\}$. Déterminer la fonction de répartition de la variable y_θ. (Rappel : $\sum_{\ell=0}^{n} t^\ell = (1-t^{n+1})/(1-t)$.)
3. Tarzan observe la réalisation de l'événement $(y_\theta \leq t)$ pour $t = 3/4$. Montrer que la loi conditionnelle de θ sachant $(y_\theta \leq t)$ est donnée par
$$p(\theta = k | y_\theta \leq t) = \frac{t^{k-1}(1-t)}{1-t^n}\,, \quad k = 1,\ldots,n.$$
4. Démontrer que, pour tout $k = 1,\ldots,n$, nous avons
$$p(\theta = k | y_\theta \leq t) = p(\theta_\star = k | \theta_\star \leq n).$$
où $\theta_\star$ est une variable de loi géométrique de paramètre $1-t$.
5. Déterminer la limite de la loi a posteriori de θ lorsque n tend vers l'infini. Quelle est l'espérance mathématique de la loi limite ?
6. Donner une estimation du nombre de girafes mesurées par Tarzan s'appuyant sur les questions précédentes. Calculer la probabilité que Tarzan

ait mesuré moins de 10 girafes.

Exercice 5. Estimation de la fréquence revisitée par Laplace. Pierre-Simon observe qu'un événement de probabilité inconnue, θ, se produit $y = 11$ fois lors de la répétition de $n = 25$ épreuves indépendantes. Il souhaite estimer la probabilité que cet événement se réalise à nouveau lors de la 26^{e} épreuve. Il considère θ la probabilité inconnue comme s'il s'agissait du tirage d'une variable aléatoire de loi uniforme sur $(0, 1)$. On note $y_i = 1$ si l' événement se réalise lors de l'épreuve i et $y_i = 0$ sinon. L'observation y correspond donc à la somme

$$y = \sum_{i=1}^{n} y_i \,.$$

On rappelle la définition de la fonction $\text{beta}(a, b)$

$$\text{beta}(\alpha, \beta) = \int_0^1 \theta^{a-1}(1-\theta)^{b-1} = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}\,, \quad a, b > 0.$$

1. Calculer la probabilité conditionnelle $p(y|\theta)$.
2. Déterminer la loi conditionnelle $p(\theta|y)$ (loi a posteriori).
3. Calculer l'espérance conditionnelle de θ sachant y. Que se passe-t-il lorsque n (et y) $\to \infty$? Interprétation ?
4. Programmer un algorithme en langage `R` permettant de simuler la loi conditionnelle $p(\theta|y)$ en utilisant les générateurs aléatoires `rbinom()` et `runif()`.
5. Montrer que la probabilité que l'événement se réalise lors d'une épreuve future sachant y est égale à

$$p(y_{n+1} = 1|y) = \frac{y+1}{n+2}\,.$$

Vérifier cette équation numériquement à l'aide de l'algorithme de simulation.

6. Calculer la probabilité de l'événement $(y_{n+1} = 1, y_{n+2} = 0, y_{n+3} = 1)$ sachant y.
7. À l'aide de l'algorithme de simulation, estimer la probabilité pour que l'on observe plus de 5 fois l'événement lors des 10 épreuves suivant n.

Exercice 6. Recherche d'un objet. Un objet a disparu dans une zone déterminée divisée en trois régions de même surface pour les recherches. La recherche commence par la fouille de la région 1. Soit q_1 la probabilité de ne pas trouver l'objet dans cette région conditionnellement à l'événement qu'il s'y trouve effectivement. Quelle est la probabilité que l'objet se trouve dans les régions 2 ou 3 sachant que les recherches dans la région 1 ont été infructueuses ?

Exercice 7. Nombre de taxis londoniens. En arrivant à Londres, Jeff voit un taxi numéroté $n = 65536$. Soit θ le nombre (inconnu) de taxis circulant à Londres. En supposant une loi dégénérée pour θ, de la forme $p(\theta) \propto 1/\theta$, donner une estimation de la médiane de la loi a posteriori $p(\theta|n)$.

Exercice 8. Casino (de Martin Scorcese). Gogoland est un lieu avec des machines à sous. En jouant une partie sur une machine, on peut y gagner un jackpot avec la probabilité p, et ne rien gagner avec la probabilité $q = 1 - p$. Les machines sont indépendantes les unes des autres. Le patron du lieu reçoit

un visiteur. Il ne dit pas au visiteur combien il possède de machines à sous, mais il lui dit que **chacune de ses machines a joué plus de k parties sans qu'aucun jackpot n'ait encore été gagné**. Pour le visiteur, le nombre inconnu de machines à sous est modélisé comme la réalisation d'une variable aléatoire notée N.

1) Sous quelles hypothèses la loi de probabilité décrivant le nombre de parties jouées sur une machine à sous avant un jackpot est la loi géométrique de paramètre p ? On supposera ces hypothèses vérifiées.

2) Soit T le plus petit nombre de parties jouées quelle que soit la machine à sous. Montrer que

$$\mathrm{P}(T > k | N = n) = q^{nk}\,, \quad n \geq 1.$$

3) On suppose que N suit la loi géométrique de paramètre $(1-\alpha)$, où α vérifie $0 < \alpha < 1$. Calculer la probabilité de l'événement $(T > k)$.

4) À l'aide des questions précédentes, montrer que la probabilité conditionnelle de l'événement $(N = n)$ sachant $(T > k)$ est égale à

$$\mathrm{P}(N = n | T > k) = (1 - \alpha q^k)(\alpha q^k)^{n-1}\,, \quad n \geq 1.$$

Identifier la loi de probabilité conditionnelle définie ci-dessus.

5) Vérifier que l'expression précédente admet une limite pour tout $n \geq 1$ lorsque α tend vers 1. Vérifier que le passage à limite définit une loi de probabilité que l'on peut reconnaître.

6) Calculer l'espérance de la loi identifiée à la question 5). On suppose que α tend vers 1, que $p = 10^{-6}$ et que $k = 50000$. En déduire que l'espérance du nombre de machines à sous tend vers une valeur proche de $1/(1-e^{-0.05}) \approx$

21.

3 Séance 3 : Espérance conditionnelle – Loi prédictive

3.1 Introduction

Nous discutons dans cette séance de la valeur prédictive d'un modèle probabiliste et de la manière d'évaluer les prédictions d'un modèle.

3.2 Définitions

Nous considérons le couple (y, θ), où y est un vecteur de dimension n et θ un vecteur de dimension J, de loi jointe $p(y, \theta)$. L'**espérance conditionnelle** de θ sachant y est définie comme l'espérance de la loi conditionnelle $p(\theta|y)$

$$\mathrm{E}[\theta|y] = \int \theta p(\theta|y) d\theta \,.$$

Lorsque la dimension du paramètre θ est supérieure à 1, l'intégrale est calculée composante par composante, et l'espérance est définie comme un vecteur de taille J. Dans un sens à préciser, l'espérance conditionnelle représente la meilleure estimation que l'on peut faire du paramètre θ à l'issue de la phase d'apprentissage. Nous illustrons le calcul de l'espérance conditionnelle dans une section séparée.

Un concept particulièrement utile en matière de prédiction est la notion de **loi prédictive** a posteriori. Afin de définir cette loi, considérons un vecteur de données, y, et un modèle probabiliste pour ces données, décrit par la loi a priori, $p(\theta)$, et par la loi générative, $p(y|\theta)$. La loi prédictive est la loi d'une nouvelle donnée, y_{rep}, prédite par la loi générative suite à l'observation du vecteur y

$$p(y_{\mathrm{rep}}|y) = \int p(y_{\mathrm{rep}}|\theta) p(\theta|y) d\theta.$$

Dans ce calcul, nous avons supposé que la réplique y_{rep} est indépendante de y sachant θ. En condensé, cette intégrale se réécrit de la manière suivante

$$p(y_{\text{rep}}|y) = \mathrm{E}[p(y_{\text{rep}}|\theta)|y],$$

où le calcul de la moyenne porte sur la variable aléatoire θ. L'intégrale précédente traduit le mécanisme de génération de données a posteriori par le modèle. Dans une première étape, les données, y, permettent de mettre à jour l'incertitude sur le paramètre θ. Cela est possible grâce à la formule de Bayes et au calcul de la loi a posteriori, $p(\theta|y)$. Dans une seconde étape, nous utilisons la loi a posteriori pour tirer de nouveaux paramètres, et nous créons des répliques des données selon la loi générative.

Notons que l'intégrale définissant la loi prédictive peut être très difficile à évaluer mathématiquement. Nous aurons le plus souvent recours à un algorithme de Monte Carlo pour échantillonner cette loi. Cet algorithme résume l'interprétation que nous venons de faire de la définition de la loi prédictive.

```
1. Simuler theta selon la loi p(theta|y)
2. Simuler y_rep selon la loi p(y_rep|theta)
```

L'information disponible sur la variable cachée θ est apprise des données, y. Dans la seconde étape de cet algorithme, nous avons

$$p(y_{\text{rep}}|\theta) = p(y_{\text{rep}}|\theta, y).$$

L'algorithme décrit ci-dessus n'est autre que l'algorithme de simulation du couple (θ, y_{rep}) sachant y. La loi prédictive est donc **la loi marginale** dans la simulation de ce couple de variable aléatoires

$$p(y_{\text{rep}}|y) = \int p(y_{\text{rep}}|\theta)p(\theta|y)d\theta = \int p(y_{\text{rep}}, \theta|y)d\theta.$$

Cet argument justifie l'algorithme proposé pour simuler la loi prédictive. Elle clarifie la manière d'obtenir les prédictions en pratique. Elles seront obtenues en **utilisant la loi générative pour les paramètres du modèle appris des données**.

La loi prédictive peut être utilisée pour évaluer la capacité du modèle à capturer les caractéristiques essentielles des données. Ces caractéristiques sont bien souvent spécifiques au problème considéré, et nous entrons ici dans le contexte de la vérification du modèle (model checking). Dans ce contexte, il faut définir avec précaution les aspects d'un modèle que l'on souhaite évaluer. Pour reprendre une citation du célèbre statisticien G. Box, *tous les modèles sont faux, mais certains modèles faux peuvent être utiles*. En effet un modèle de la météo de Grenoble peut mal prévoir la météo à Nice. Nous devons donc définir avec soin les critères (ou les statistiques) permettant l'évaluation d'un modèle.

3.3 Propriétés de l'espérance conditionnelle

Dans cette section, nous présentons deux propriétés importantes de l'espérance conditionnelle. La première propriété généralise la formule des probabilités totales, vue en première année. La deuxième propriété établit l'optimalité prédictive de l'espérance conditionnelle au sens des moindres carrés.

Propriété 1 : Formule de conditionnement. Nous avons

$$\mathrm{E}[\mathrm{E}[\theta|y]] = \mathrm{E}[\theta]$$

En d'autres termes, lorsque l'espérance conditionnelle de θ est moyennée sur l'ensemble des données, nous retrouvons l'espérance tout simplement

$$\mathrm{E}[\theta] = \int \mathrm{E}[\theta|y]p(y)dy.$$

La preuve de ce résultat résulte d'une inversion de l'ordre des intégrales intervenant dans le calcul de l'espérance

$$\int \mathrm{E}[\theta|y]p(y)dy = \int \left(\int \theta p(\theta|y)d\theta \right) p(y)dy$$

Sous réserve d'existence, inversons les symboles d'intégration

$$\int \mathrm{E}[\theta|y]p(y)dy = \int \theta \left(\int p(\theta|y)p(y)dy \right) d\theta \,.$$

Le résultat découle du fait que

$$p(\theta) = \int p(\theta|y)p(y)dy \,.$$

Propriété 2 : Espérance conditionnelle et prédiction. Dans ce paragraphe, nous établissons une propriété d'optimalité de l'espérance conditionnelle. Vue comme une fonction des données, y, **l'espérance conditionnelle $\mathbf{E}[\theta|y]$ est le prédicteur optimal du paramètre** θ au sens des moindres carrés.

Supposons donc que l'on observe y, et que l'on souhaite prédire la valeur de θ, en s'appuyant sur l'observation de y. On note $g(y)$ la fonction servant de *prédicteur* pour θ. On suppose que $\mathrm{E}[g^2(y)] < \infty$. On dit q'un prédicteur, $g^\star(y)$ est optimal s'il est solution du problème de minimisation suivant

$$g^\star(y) = \arg\min_{g \in L^2} \{\mathrm{E}[(\theta - g(y))^2]\} \,,$$

où L^2 représente l'ensemble des prédicteurs tels que $\mathrm{E}[g^2(y)] < \infty$.

Voyons pourquoi la solution de ce problème de minimisation est donnée par l'espérance conditionnelle. Tout d'abord, grâce à la propriété 1, nous avons

$$\mathrm{E}[(\theta - g(y))^2] = \mathrm{E}[\mathrm{E}[(\theta - g(y))^2|y]] \,.$$

Il suffit donc de minimiser l'expression $\mathrm{E}[(\theta - g(y))^2|y]$. En développant le terme quadratique à l'intérieur de l'espérance, nous obtenons

$$\mathrm{E}[(\theta - g(y))^2|y] = \mathrm{E}[(\theta - \mathrm{E}[\theta|y])^2|y] + \mathrm{E}[(g(y) - \mathrm{E}[\theta|y])^2|y]\,.$$

En remarquant que le terme $\mathrm{E}[(g(y) - \mathrm{E}[\theta|y])^2|y]$ est toujours positif ou nul, nous obtenons le résultat annoncé

$$\mathrm{E}[(\theta - g(y))^2|y] \geq \mathrm{E}[(\theta - \mathrm{E}[\theta|y])^2|y].$$

Puisque une variable aléatoire positive d'espérance nulle est nécessairement p.s. nulle, cela implique que $g^\star(y) = \mathrm{E}[\theta|y]$.

3.4 Exemple : le modèle d'erreur

Pour illustrer le concept d'apprentissage et le calcul de la loi prédictive, considérons un exemple très simple. Nous observons la valeur d'une mesure unidimensionnelle continue, y, effectuée avec une erreur. Nous ne connaissons donc pas la véritable valeur de cette mesure, θ, et nous souhaitons l'estimer. Pour cela, nous disposons d'un modèle de l'erreur. En particulier, l'écart-type de l'erreur est connu exactement, et il est égal à σ.

Nous considérons le modèle défini par

$$\theta \sim p(\theta) = \mathrm{N}(0, \sigma_0^2)$$

et

$$y|\theta \sim N(\theta, \sigma^2)\,.$$

La constante σ_0^2 représente la **variance** de la loi a priori. Nous notons $\beta_0 = 1/\sigma_0^2$. La constante β_0 est appelée la **précision** de la loi a priori. Nous notons $\beta = 1/\sigma^2$

la précision de la loi générative. La loi $\mathrm{N}(\theta, \sigma^2)$ est la loi normale de moyenne θ et de variance σ^2. Elle admet pour densité

$$p(y|\theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-(y-\theta)^2/2\sigma^2)\,, \quad y \in \mathbb{R}.$$

Pour modéliser une incertitude a priori absolue sur la valeur de θ, il est tentant de choisir $\sigma_0^2 = \infty$. Cela correspond à une loi a priori dégénérée. Malgré le caractère non-défini de cette loi, nous verrons que les calculs ont tout de même du sens, la loi a posteriori étant bien définie. Le modèle se résume donc de la manière suivante. La valeur de la mesure θ est inconnue, et supposée tirée d'une loi non-informative (ou plate). On observe une version bruitée de θ, notée y, obtenue par l'ajout à θ d'une perturbation gaussienne d'écart-type σ.

La loi a posteriori s'obtient directement en effectuant le produit des densités $p(\theta)$ et $p(y|\theta)$

$$p(\theta|y) \propto \exp\left(-\frac{1}{2}(\beta_0\theta^2 + \beta(y-\theta)^2)\right)$$

En réorganisant les termes à l'intérieur de l'exponentielle, nous obtenons

$$p(\theta|y) \propto \exp\left(-\frac{1}{2}\beta_1(\theta - m_1)^2\right)$$

où

$$\left\{ \begin{array}{ll} \beta_1 & = \beta_0 + \beta \\ m_1 & = \beta y/(\beta + \beta_0)\,. \end{array} \right.$$

La loi a posteriori est donc la loi normale de moyenne m_1 et $\sigma_1^2 = 1/\beta_1$. Dans le cas où l'incertitude a priori est totale ($\beta_0 = 0$), nous obtenons

$$p(\theta|y) = \mathrm{N}(\theta|y, \sigma^2)\,.$$

En d'autres termes, le modèle d'erreur s'applique directement, et θ s'estime à l'aide de y en ajoutant un bruit gaussien centré d'écart-type σ. Dans le cas où

la **loi a priori est informative** ($\beta_0 > 0$), nous obtenons une loi a posteriori plus précise que les lois générative et a priori. Lorsque $\beta_0 > 0$ et y est positif, nous obtenons que

$$m_1 = \mathrm{E}[\theta|y] < y\,.$$

Dans le cas $y < 0$, l'inégalité est renversée. Nous observons donc un phénomène de rétrécissement (*shrinkage*), aussi appelé régularisation. L'introduction d'une loi a priori informative induit une modification de l'estimation de la valeur réelle qui la tire vers la valeur 0. La régularisation peut être un aspect important des modèles afin d'obtenir de meilleures qualités prédictives, par exemple en réduisant la variance des estimations. Elle peut aussi jouer un rôle numérique lors de l'exécution des algorithmes de Monte Carlo que nous verrons plus loin dans ce cours.

Nous déterminons maintenant la loi prédictive en supposant que $\beta_0 = 0$. Le calcul de la loi prédictive pour $\beta_0 > 0$ est un exercice à faire en travaux dirigés. Soit y^* (plutôt que y_{rep}) une réplique des données. Nous souhaitons calculer la loi

$$p(y^*|y) \propto \int \exp\left(-\frac{1}{2}(\beta(y^* - \theta)^2 + \beta(y - \theta)^2)\right) d\theta\,.$$

En séparant les termes de l'exponentielle dépendant de θ, nous avons

$$(y^* - \theta)^2 + (y - \theta)^2 = \frac{1}{2}(y - y^*)^2 + 2\left(\theta - \left(\frac{1}{2}(y + y^*)\right)\right)^2$$

Ainsi, nous avons,

$$p(y^*|y) \propto \exp\left(-\frac{1}{4}\beta(y - y^*)^2\right) \int \exp\left(-\beta\left(\theta - \left(\frac{1}{2}(y + y^*)\right)\right)^2\right) d\theta\,.$$

L'intégrale apparaissant dans le membre de droite est une constante que l'on peut calculer en effectuant un changement de variables affine. La valeur exacte

est connue grâce à l'expression de la loi gaussienne (exercice). Il est important de remarquer que cette valeur est indépendante de y^*. Cela n'est pas utile de faire le calcul explicite, nous obtenons directement la loi prédictive

$$p(y^*|y) = \mathrm{N}(y^*|y, 2\sigma^2)\,.$$

La loi prédictive s'interprète donc comme la somme de deux variables indépendantes de loi normale. Sachant y, la première variable correspond à θ, de loi $\mathrm{N}(y, \sigma^2)$ et la seconde variable correspond au modèle d'erreur $\mathrm{N}(0, \sigma^2)$ que l'on peut ajouter à θ.

Les calculs des intégrales des lois prédictives peuvent être longs et fastidieux ou alors ils reposent sur des techniques peu intuitives. Voyons comment, dans le cas précédent, nous pouvons nous passer de tels calculs en utilisant une approche à la fois intuitive et rigoureuse. Cette approche repose directement sur la modélisation du couple (θ, y^*) sachant y, évoquée lors de l'algorithme de simulation de la loi prédictive. Dans le cas précédent, la loi prédictive se construit en deux étapes. Lors de la première étape, nous modélisons la loi a posteriori de la manière suivante

$$\theta = y + \epsilon'$$

où ϵ' est une variable de loi $N(0, \sigma^2)$. Cette représentation est bien équivalente à celle de la loi a posteriori, $N(y, \sigma^2)$. Lors de la seconde étape, nous utilisons la loi générative

$$y^* = \theta + \epsilon^*$$

où ϵ^* est à nouveau une variable de loi $N(0, \sigma^2)$, indépendante de ϵ'. Dans ce cas, nous obtenons que y^* est une variable de loi normale caractérisée par sa

moyenne

$$\mathrm{E}[y^*] = \mathrm{E}[y + \epsilon' + \epsilon^*] = y\,,$$

et sa variance

$$\mathrm{Var}(y^*) = \mathrm{Var}(\epsilon' + \epsilon^*) = 2\sigma^2\,.$$

Ce raisonnement se généralise très bien à des lois a priori informatives et nous obtenons le résultat suivant. Pour $\beta_0 > 0$, la loi prédictive est une loi normale de moyenne m_1 et de précision β^*

$$y^*|y \sim N(m_1, \beta^*)$$

où m_1 est l'espérance conditionnelle $\mathrm{E}[\theta|y]$ et la précision β^* est décrite par la formule suivante

$$\beta^* = \frac{\beta(\beta_0 + \beta)}{2\beta + \beta_0}\,.$$

3.5 Approche par simulation numérique

Nous montrons dans ce paragraphe comment il possible de générer un histogramme de la loi prédictive sans effectuer le calcul de l'intégrale. La simulation s'appuie sur le générateur aléatoire `rnorm` du langage `R`. Les commandes ci-dessous génèrent 10000 simulations de la loi prédictive pour la valeur observée $y = 2.0$ et pour la variance connue $\sigma^2 = 1$ et nous posons $\beta_0 = 0$.

```
theta <- rnorm(10000, mean = y, sd = 1) #loi a posteriori
y.etoile <- rnorm(10000, mean = theta, sd= 1) #loi générative
```

Un résultat correspondant à cette simulation est représenté dans la figure 6. Nous observons une bonne adéquation à la courbe obtenue par le calcul

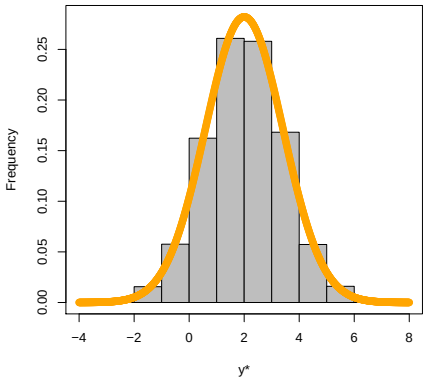


FIGURE 6 – *Loi prédictive a posteriori pour le modèle d'erreur gaussien. La donnée observée est $y = 2$ et la variance du bruit, supposée connue, est égale à 1. La courbe théorique obtenue grâce au calcul intégral est superposée à l'histogramme.*

théorique. Nous pouvons alors sans difficulté particulière faire des calculs concernant des prédictions. Par exemple, la probabilité d'observer une réplique y^* de valeur supérieure à 3 sera égale à

$$p(y^* > 3|y = 2) = 1 - F(3)$$

où F est la fonction de répartition de la loi $N(2, 2\sigma^2 = 2)$. Numériquement, cette probabilité est égale 0.24 (théoriquement 0.239). La valeur numérique est obtenue simplement par

```
mean( y.etoile > 3 )
```

et la valeur théorique est donnée par

```
pnorm(3, mean = 2, sd = sqrt(2), lower = F)
```

3.6 Données séquentielles et apprentissage

Nous supposons maintenant que nous pouvons répéter la mesure de la valeur inconnue θ. Nous n'observons plus une unique valeur continue, y, mais une suite de valeurs effectuées avec erreur. Nous notons cette suite d'observations $y_1, \ldots, y_n$, et nous cherchons à prédire de manière optimale la valeur de θ au fur et à mesure que les observations arrivent. Cela revient à calculer l'espérance conditionnelle de θ sachant y_1, puis y_1, y_2, etc.

Les calculs s'effectuent en remarquant que les lois a posteriori calculées séquentiellement sont toujours des lois normales. On identifie l'espérance conditionnelle dans le terme quadratique en reconnaissant la moyenne de la loi normale.

Suite à la première observation, y_1, nous avons, d'après les calculs effectués précédemment, le résultat suivant

$$\mathrm{E}[\theta|y_1] = m_1 = \frac{\beta y_1}{\beta_0 + \beta}\,.$$

Pour calculer $\mathrm{E}[\theta|y_1, y_2]$, notons que

$$p(\theta|y_1, y_2) \propto p(y_2|\theta)p(\theta|y_1)$$

A l'aide d'un calcul similaire à celui effectué dans la section précédente (*cf* Exercice 2), nous obtenons

$$\mathrm{E}[\theta|y_1, y_2] = m_2 = \frac{\beta_1 \mathrm{E}[\theta|y_1] + \beta y_2}{\beta_1 + \beta}\,,$$

où $\beta_1 = \beta_0 + \beta$. Suite à n observations, $y_1, \ldots, y_n$, nous obtenons

$$\mathrm{E}[\theta|y_1, \ldots, y_n] = m_n = \frac{\beta_{n-1} \mathrm{E}[\theta|y_1, \ldots, y_{n-1}] + \beta y_n}{\beta_{n-1} + \beta}\,, \tag{1}$$

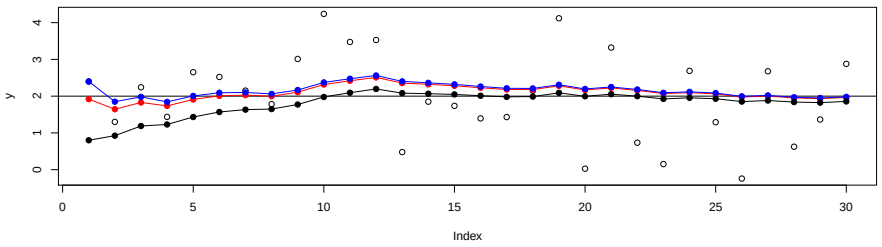


FIGURE 7 – *Espérance conditionnelle pour le modèle d'erreur gaussien. Les cercles vides représentent* $n = 30$ *données* $y_1, \ldots, y_{30}$. *La valeur cachée est* $\theta = 2$ *et la variance du bruit est égale à 1. Les courbes montrent l'apprentissage de la valeur cachée pour différentes valeurs de précision a priori :* $\beta_0 = 0$ *(bleu),* $\beta_0 = .25$ *(rouge),* $\beta_0 = 2$ *(vert). On observe le phénomène d'aplatissement (shrinkage) lorsque la précision* β_0 *grandit.*

où $\beta_{n-1} = \beta_{n-2} + \beta$. Les valeurs obtenues pour des lois a priori différentes les unes des autres sont illustrées dans la figure 7.

Lorsque la loi a priori est non-informative ($\beta_0 = 0$), nous obtenons un résultat très simple

$$\mathrm{E}[\theta|y_1, \ldots, y_n] = m_n = \frac{\sum_{i=1}^{n} y_i}{n} .$$

Cela illumine l'intérêt de considérer la moyenne empirique. Lorsque n tend vers l'infini, nous savons grâce à la loi des grands nombres que la moyenne empirique converge vers la valeur inconnue. Lorsque l'on dispose de n données (n est fini), la moyenne empirique représente la meilleure prédiction que l'on peut faire de la valeur inconnue au sens des moindres carrés.

3.7 Résumé

Résumer les points à retenir et donner quelques exemples illustrant les concepts principaux de la séance.

3.8 Exercices

Pour les exercices suivants, il est nécessaire d'avoir un ordinateur sur lequel on aura installé préalablement le logiciel `R`. Ce logiciel libre est disponible sur le site `http://cran.r-project.org/`.

Exercice 1. Simulation et calcul d'une loi prédictive.

Nous observons la valeur d'une mesure unidimensionnelle continue, y, effectuée avec une erreur. Nous ne connaissons donc pas la véritable valeur de cette mesure, θ, et nous souhaitons l'estimer. Pour cela, nous disposons d'un modèle de l'erreur. En particulier, l'écart-type de l'erreur est connu exactement, et il est égal à σ.

Nous considérons le modèle défini par

$$\theta \sim p(\theta) = \mathrm{N}(0, \sigma_0^2)$$

et

$$y|\theta \sim \mathrm{N}(\theta, \sigma^2)$$

On suppose que σ_0^2 est une valeur finie.

1. Retrouver l'expression de la loi a posteriori, reconnaître cette loi. Ecrire un algorithme de simulation de la loi prédictive du modèle pour la donnée y.
2. On suppose que $\sigma = 1$ et $\sigma_0 = 2$. On observe $y = 3$. Ecrire un algorithme de simulation de la loi prédictive, le programmer en langage `R` et générer 10000 tirages de cette loi.

3. Montrer un histogramme obtenu par la méthode de simulation précédente.

   Calculer la loi prédictive a posteriori théoriquement (facultatif).

Exercice 2. Données poissoniennes.

1. On suppose un modèle d'erreur poissonnien. Dans ce modèle, la variable cachée, θ, est une variable positive. On suppose

$$p(\theta) \propto \frac{1}{\theta}$$

   et

$$p(y|\theta) \propto \frac{\theta^y}{y!} \exp(-\theta).$$

   Calculer l'espérance conditionnelle, $\mathrm{E}[\theta|y]$, lorsque l'on dispose d'une observation,y, supposée non nulle.

2. Calculer l'espérance conditionnelle, $\mathrm{E}[\theta|y_1, \ldots, y_n]$, lorsque l'on dispose de n observations, et on suppose qu'au moins l'une d'entre elles est non-nulle.

3. Calculer la loi prédictive a posteriori lorsque l'on observe $y = y_1 = 1$ $(n = 1)$.

4. Le club d'astro a observé une supernova cette année. Quelle est la probabilité qu'il observe plus de 2 supernovæl'an prochain ?

Exercice 3. Début du TP 2 : Détection d'un changement ponctuel dans une suite binaire.

On observe une suite de longueur n, notée y, composée de signaux binaires correspondant à l'émission d'une source de fréquence inconnue θ_1, susceptible de changer en un instant (point) inconnu de θ_1 à θ_2. Les émissions sont

indépendantes les unes des autres. Par exemple, cette suite peut se présenter de la manière suivante :

$$y = 01100 \ldots 00110 \| 1110011 \ldots 1100111$$

Dans cette représentation, le symbole $\|$ marque le changement ayant lieu au point c, un indice entier compris entre 1 et n. Par convention, $c = 1$ correspond à la situation où il n'y a pas de changement et les n tirages sont de fréquence égale à θ_2. Nous notons

$$\theta = (c, \theta_1, \theta_2).$$

Pour $c = 2, \ldots, n$, nous avons

$$p(y_i|\theta) = \theta_1^{y_i}(1-\theta_1)^{1-y_i} \quad i = 1, \ldots, c-1$$

et

$$p(y_i|\theta) = \theta_2^{y_i}(1-\theta_2)^{1-y_i} \quad i = c, \ldots, n.$$

Nous supposons que

$$p(\theta) = p(c)p(\theta_1)p(\theta_2) = \frac{1}{n}, \quad 0 \leq \theta_1, \theta_2 \leq 1 \; et \; c = 1, \ldots, n.$$

L'objectif de cet exercice est de proposer (et de tester) un algorithme d'apprentissage du point de changement c, permettant le calcul de la loi a posteriori $p(c|y)$.

1. Soit $c = 1$. Donner l'expression de la loi générative $p(y|c = 1, \theta_1, \theta_2)$.
2. Même question pour $c > 1$.
3. On suppose que θ_1 et θ_2 sont connues. Déduire des questions précédentes que

$$\forall c = 2, \ldots, n, \quad \frac{p(c|y)}{p(c=1|y)} = \prod_{j=1}^{c-1} \frac{\theta_1^{y_i}(1-\theta_1)^{1-y_i}}{\theta_2^{y_i}(1-\theta_2)^{1-y_i}}$$

Vérifier que l'on peut calculer le rapport $p(c|y)/p(c=1|y)$ pour tout c en effectuant de l'ordre de $n(n-1)/2$ multiplications.

4. Supposant θ_1 et θ_2 connues, proposer un algorithme permettant de calculer $p(c|y)$ pour tout $c = 1, \ldots, n$ avec une complexité en $O(n)$ (truc : mettre à jour le rapport définit dans la question précédente à l'aide d'une récurrence linéaire).

Exercice 4. Loi jointe du minimum et du maximum de deux variables de loi exponentielle. On considère un couple (y, θ) de densité définie par

$$\forall (y,\theta) \in \mathbb{R}^2\,, \quad p(y,\theta) = \begin{cases} 2\,e^{-\theta-y} & \text{si } 0 < y < \theta\,, \\ 0 & \text{sinon.} \end{cases}$$

1. Rappeler le principe de simulation d'un couple de variables aléatoires de densité $p(y, \theta)$.
2. Calculer la loi marginale, la loi a priori, la loi générative et la loi a posteriori pour le couple (y, θ).
3. Montrer que la loi a posteriori peut se représenter comme la loi de la somme $y + z$, où z est une variable de loi exponentielle de paramètre 1 pour tout $y > 0$.
4. Ecrire un algorithme de simulation du couple (y, θ) faisant explicitement appel à un changement de variables. Prouver la validité de cet algorithme.
5. Soit `unif(0,1)` un générateur aléatoire de loi uniforme. On considère l'algorithme suivant

   `Repeat`

```
u = -log(unif(0,1)) ;
v = -log(unif(0,1)) ;
Jusqu'a (u < v) ;
 y <- u ; theta <- v ;
```

Argumenter (sans calcul) du fait que la loi du couple (y, θ) en sortie de cet algorithme est identique à la loi du couple $(\min(u, v), \max(u, v))$, où u, v sont des variables indépendantes et de loi exponentielle de paramètre 1.

6. En déduire que la loi de y en sortie de l'algorithme précédent est la loi exponentielle de paramètre 2.

7. Soit $y > 0$ et soit $t > y$. En sortie de l'algorithme précédent, justifier que l'on a

$$p(\theta \leq t|y) = \frac{p(y < v \leq t)}{p(v > y)} ,$$

et calculer la densité de la loi conditionnelle $p(\theta|y)$.

8. Démontrer que le couple (y, θ) en sortie de l'algorithme précédent admet $p(y, \theta)$ pour densité jointe.

9. (Facultatif) Soit $y > 0$. En utilisant la question 3, montrer que $\mathrm{E}[\theta|y] = y + 1$. En déduire que $\mathrm{E}[y\theta|\theta] = \theta^2 + \theta$.

10. (Facultatif) Déduire de la question précédente l'espérance $\mathrm{E}[\theta]$ et la covariance $\mathrm{Cov}(y, \theta)$.

4 Séance 4 : Inférence probabiliste pour le modèle gaussien : moyenne et variance.

4.1 Introduction

Dans la séance précédente, nous avons étudié comment il est possible d'apprendre la valeur d'un paramètre caché à partir d'observations séquentielles erronées de cette valeur, mais ayant une erreur de variance connue. Dans cette séance, nous généralisons cette approche. Nous supposons que nous observons simultanément n répliques bruitées de ce paramètre inconnu et discutons le cas où le modèle d'erreur n'est pas connu.

Soit m le paramètre d'intérêt et σ^2 la variance du bruit, que nous supposons tour à tour connue ou inconnue. On dira que σ^2 est un paramètre de nuisance. Nous ne souhaitons pas particulièrement connaître sa valeur, mais nous devons modéliser son incertitude afin d'estimer ou d'apprendre m. Nous supposons que nous observons n réalisations indépendantes de la loi générative $\mathrm{N}(m, \sigma^2)$. Les observations sont notées $y = y_1, \ldots, y_n$.

Nous examinons tour à tour les cas suivants :

1. σ^2 connu, on cherche la loi de $\theta = m$ sachant y,
2. m connu, on cherche la loi de $\theta = \sigma^2$ sachant y,
3. m et σ^2 tout deux inconnus, on cherche la loi de $\theta = (m, \sigma^2)$ sachant y.

4.2 Quelques définitions : Matrice de covariance, loi gaussienne multivariée

Considérons un couple (x_1, x_2) de variables aléatoires réelles. Nous appelons **covariance** du couple (x_1, x_2) la grandeur définie par

$$\mathrm{Cov}(x_1, x_2) = \mathrm{E}[x_1 x_2] - \mathrm{E}[x_1]\mathrm{E}[x_2]\,.$$

La covariance du couple (x_1, x_2) est **nulle** lorsque les variables aléatoires x_1 et x_2 sont **indépendantes**. Nous voyons aussi que

$$\mathrm{Cov}(x_1, x_1) = \mathrm{Var}(x_1).$$

La covariance possède quelques propriétés intéressantes. Tout d'abord, elle est symétrique

$$\mathrm{Cov}(x_1, x_2) = \mathrm{Cov}(x_2, x_1).$$

De plus, la covariance est bi-linéaire

$$\mathrm{Cov}(ax_1 + by_1, cx_2 + dy_2) = ac\mathrm{Cov}(x_1, x_2) + \cdots + bd\mathrm{Cov}(y_1, y_2)\,.$$

De plus, elle vérifie l'inégalité de Cauchy-Schwarz

$$\mathrm{Cov}(x_1, x_2)^2 \leq \mathrm{Var}(x_1)\mathrm{Var}(x_2)\,.$$

Ainsi, on pourra définir le **coefficient de corrélation** entre x_1 et x_2 comme la grandeur comprise entre -1 et $+1$ suivante

$$\rho = \rho(x_1, x_2) = \frac{\mathrm{Cov}(x_1, x_2)}{\sqrt{\mathrm{Var}(x_1)\mathrm{Var}(x_2)}}$$

Matrice de covariance. Nous appelons **matrice de covariance** la matrice C dont le terme général est

$$c_{ij} = \mathrm{Cov}(x_i, x_j)$$

Nous supposons que les valeurs propres de la matrice de covariance sont non-nulles (strictement positives).

Couple gaussien. Nous dirons que le couple (x_1, x_2) est gaussien, de moyenne $m = (m_1, m_2)$ et de matrice de covariance C, si la densité de pro-

babilité de ce couple est donnée par la formule suivante

$$p(x_1, x_2) = \frac{1}{2\pi} \frac{1}{\sqrt{\det(C)}} \exp\left(-\frac{1}{2}(x-m)^T C^{-1}(x-m)\right) ,$$

pour tout $(x_1, x_2) \in \mathbb{R}^2$. Les propriétés des couples et des vecteurs gaussiens seront étudiées en détails plus loin dans le cours de MPA. Pour l'instant, nous admettrons (sans peine) que les lois marginales et conditionnelles d'un couple gaussien sont elles-aussi gaussiennes. Notons encore que la formule ci-dessus peut servir à définir un vecteur gaussien en dimension d, en remplaçant la constante par $1/(2\pi)^{d/2}$.

4.3 Apprentissage probabiliste de la moyenne (m) et de la variance (σ^2)

Nous supposons que nous observons n réalisations indépendantes de la loi générative $\mathrm{N}(m, \sigma^2)$. Ces observations sont notées $y = y_1, \ldots, y_n$. Soit m notre paramètre d'intérêt et σ^2 la variance du bruit dans le modèle d'erreur.

Cas σ^2 connu. Supposons dans un premier temps, que σ^2 est connu. Dans ce cas, nous posons $\theta = m$ et nous nous retrouvons dans une situation familière, déjà rencontrée lors de la dernière séance. Nous supposons que la loi de θ est non-informative ($\beta_0 = 0$)

$$p(\theta) \propto 1 .$$

Les observations $(y_1, \ldots, y_n)$ sont indépendantes, nous avons donc

$$p(y_1, \ldots, y_n|\theta) = \prod_{i=1}^{n} p(y_i|\theta)$$

et nous pouvons écrire l'expression de la loi générative

$$p(y_1, \ldots, y_n|\theta) \propto \prod_{i=1}^{n} \exp\left(-\frac{1}{2\sigma^2}(y_i - \theta)^2\right) .$$

En utilisant la propriété de l'exponentielle, nous voyons que la loi générative est une loi gaussienne multivariée, de matrice de covariance diagonale

$$p(y_1, \ldots, y_n|\theta) \propto \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\theta)^2\right).$$

Pour obtenir la loi a posteriori, considérons la moyenne empirique

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

et développons le carré à l'intérieur de l'exponentielle

$$\sum_{i=1}^{n}(y_i-\theta)^2 = \sum_{i=1}^{n}(y_i-\bar{y})^2 + n(\bar{y}-\theta)^2\,.$$

Le premier terme du second membre est indépendant de θ, nous avons donc

$$p(\theta|y) \propto p(y|\theta)p(\theta) \propto \exp\left(-\frac{n}{2\sigma^2}(\bar{y}-\theta)^2\right).$$

Ainsi, nous voyons que toute l'information utile pour apprendre le paramètre $\theta = m$ est résumée dans la statistique $\bar{y}$

$$p(\theta|y) = p(\theta|\bar{y})\,.$$

En conclusion, la loi a posteriori est la loi normale $\mathrm{N}(\bar{y}, \sigma^2/n)$.

Cas m connu. Nous traitons le cas symétrique au précédent où l'on cherche maintenant à apprendre $\theta = \sigma^2$ sachant la valeur de m. Ce cas n'est pas très réaliste, mais nous verrons que son étude est très utile ensuite.

Rappelons tout d'abord les définitions des lois χ^2_n et inverse χ^2_n à n degrés de liberté. Il s'agit de lois de variables positives. La densité de la loi χ^2_n s'écrit de la manière suivante :

$$p(\theta) \propto \theta^{n/2-1}e^{-\theta/2}\,, \quad \theta > 0\,.$$

La loi Inv-$\chi^2(\nu, s^2)$, est la loi de la variable définie par

$$\theta = \frac{\nu s^2}{Z}\,,$$

où Z suit une loi χ^2_ν, $\nu > 0$. La densité de la loi Inv-$\chi^2(\nu, s^2)$ est proposée en exercice :

$$p(\theta) \propto \theta^{-\nu/2-1} \exp(-\nu s^2/2\theta)\,, \quad \theta > 0\,.$$

Le cas $\nu = 0$ est dégénéré. Il correspond à une loi non-informative

$$p(\log\theta) \propto 1 \quad \text{ou encore} \quad p(\theta) \propto 1/\theta\,.$$

Cela signifie que les valeurs de θ sont uniformément réparties sur une échelle logarithmique.

Afin de traiter le cas de l'apprentissage de $\theta = \sigma^2$ (m connu), nous considérons un modèle dans lequel la loi a priori n'admet pas de densité

$$p(\theta) \propto \frac{1}{\theta}\,.$$

La loi d'échantillonnage est inchangée. On peut l'écrire de la manière suivante

$$p(y_1, \ldots, y_n|\theta) \propto \frac{1}{\theta^{n/2}} \exp\left(-\frac{n}{2\theta} s_n^2\right)$$

où

$$s_n^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - m)^2\,.$$

En multipliant $p(y_1, \ldots, y_n|\theta)$ par l'inverse de θ, nous obtenons la loi a posteriori de la variance $\theta = \sigma^2$. Nous reconnaissons une loi de la famille Inv-χ^2, dont les paramètres sont faciles à identifier

$$\sigma^2|y \sim \text{Inv}{-}\chi^2(n, s_n^2)\,.$$

Apprentissage du couple (m, σ^2). Dans ce paragraphe, ni m ni σ^2 ne sont connus. Cela correspond à la situation la plus réaliste. Nous souhaitons donc apprendre le paramètre composite $\theta = (m, \sigma^2)$. Afin de spécifier un modèle pour le couple (θ, y), nous choisissons une loi a priori non-informative

$$p(m, \sigma^2) \propto \frac{1}{\sigma^2} .$$

Cette formulation implique que la loi a priori est uniforme pour m ($p(m) \propto 1$), et que la loi a priori est uniforme pour $\log(\sigma^2)$. Une loi uniforme sur une échelle logarithmique se justifie parce que la variance est un paramètre d'échelle, et l'ordre de grandeur de la variation est plus importante que la variation elle-même. Notons enfin que supposer une loi uniforme pour σ^2 plutôt que pour son logarithme conduirait à une loi a posteriori non définie.

Les calculs détaillés de la loi a posteriori seront effectués en exercice. Après arrangement des termes présents à l'intérieur de l'exponentielle, nous obtenons

$$p(m, \sigma^2|y) \propto \frac{1}{(\sigma^2)^{n/2+1}} \exp\left(-\frac{(n-1)}{2\sigma^2} s^2_{n-1} - \frac{n}{2\sigma^2}(\bar{y} - m)^2)\right)$$

où l'on a introduit une statistique très courante, correspondant à la **variance empirique** de l'échantillon y

$$s^2_{n-1} = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2 .$$

Contrairement aux cas précédents, nous nous trouvons en présence d'une loi multivariée, dont l'expression est explicite, mais qui ne correspond à **aucune loi connue**. Il est important, à ce niveau, d'insister que cela est la règle générale en apprentissage probabiliste. Les cas où l'on tombe sur des expressions connues restent exceptionnels.

Dans l'esprit général de ce cours, nous aurons recours à la simulation pour étudier la loi $p(m, \sigma^2|y)$. Etant donné l'échantillon y, nous utiliserons l'algorithme suivant :

1. Simuler σ^2 selon la loi $p(\sigma^2|y)$,
2. Sachant σ^2, simuler m selon la loi $p(m|\sigma^2, y)$.

Nous avons déjà résolu l'opération à effectuer lors de la seconde étape de cet algorithme. En effet, nous avons obtenu précédemment que la loi $p(m|\sigma^2, y)$ était une loi normale de moyenne $\bar{y}$ et de variance σ^2/n. En effet cette situation correspond au cas où σ^2 est connu.

Afin de simuler la loi $p(\sigma^2|y)$, nous devons identifier cette loi. Il s'agit de la marginale de la loi a posteriori dont l'expression est donnée plus haut dans ce paragraphe

$$p(\sigma^2|y) = \int p(m, \sigma^2|y) dm \,.$$

Nous devons donc effectuer le calcul d'une intégrale. C'est un exercice que l'on peut résoudre, et nous proposons de le faire en travaux dirigés. Nous trouvons que la loi de σ^2 est une loi Inv-χ^2

$$\sigma^2|y \sim \text{Inv}{-}\chi^2(n-1, s^2_{n-1}) \,.$$

Finalement, la simulation de la loi a posteriori du couple (m, σ^2) se réalise de la manière suivante

1. $\sigma^2|y \sim (n-1)s^2_{n-1}/\chi^2_{n-1}$
2. $m|\sigma^2, y \sim \text{N}(\bar{y}, \sigma^2/n)$

En langage `R`, nous pouvons coder cet algorithme pour effectuer 10000 tirages selon la loi a posteriori de la manière suivante. Dans cette simulation, nous considérons que la variable cachée est égale à 0 et que la variance du modèle

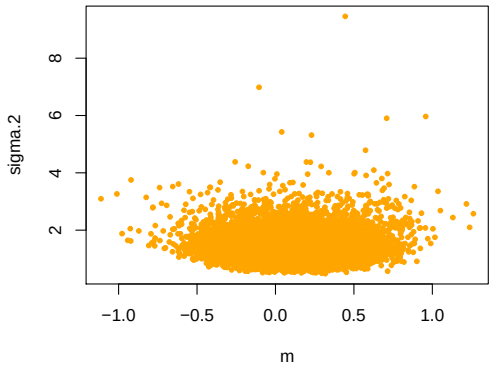


FIGURE 8 – *Echantillonnage de la loi a posteriori* $p(m, \sigma^2|y)$ *pour un échantillon de taille* $n = 20$. *La valeur cachée est égale à 0, et la variance du modèle d'erreur est égale à 1.*

d'erreur est égale à 1. Nous disposons de $n = 20$ répliques indépendantes. Dans ce cas, elles sont tirées au hasard

```
# simulated data

n = 20; y = rnorm(n)

# Posterior distribution sampling

sigma.2 = (n-1)*var(y)/rchisq(10000, n-1)

m = rnorm(10000, mean(y), sd = sqrt(sigma.2/n))
```

La figure 8 décrit un échantillonnage de la loi a posteriori. Toutes les grandeurs d'intérêt peuvent être calculées avec précision à partir de cet échantillon de grande taille. Par exemple, nous estimons que le coefficient de corrélation du couple (m, σ^2) est environ 0.001. Ce résultat confirme la théorie de RA Fisher indiquant l'indépendance de la moyenne et de la variance empirique dans un échantillon gaussien.

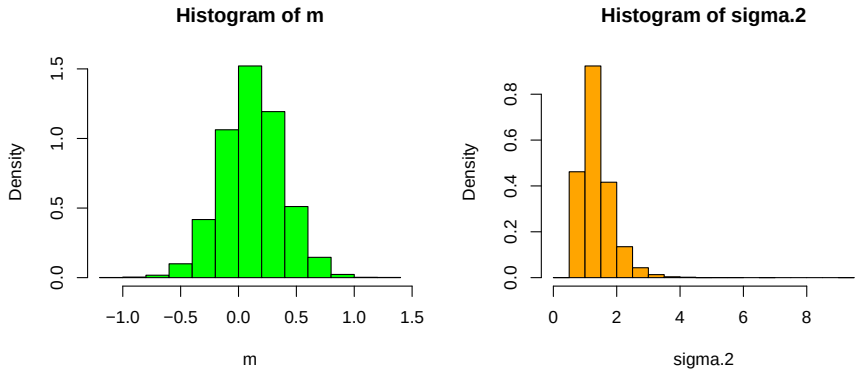


FIGURE 9 – *Histogrammes des lois a posteriori $p(m|y)$ et $p(\sigma^2|y)$ pour un échantillon de taille $n = 20$. La valeur cachée est égale à 0 et la variance du modèle d'erreur est égale à 1.*

Il est possible en particulier de calculer les histogrammes des lois marginales de m et σ^2 sachant y (figure 9). Remarquons que nous avons une expression théorique pour la loi $p(\sigma^2|y)$. Bien que cela ne soit pas nécessaire ici, il est intéressant de noter que nous pouvons aussi déterminer la loi de $p(m|y)$. En effet une série de calcul élémentaire conduit au résultat suivant

$$\sqrt{n}\frac{m-\bar{y}}{s_{n-1}}|y \sim t_{n-1}$$

où t_{n-1} est la loi de Student à $n-1$ degrés de liberté. Cette loi apparait notamment dans la théorie statistique pour tester une valeur moyenne lorsque la variance est inconnue (t-test).

4.4 Pertinence et qualité du modèle d'erreur

Lorsque l'on construit un modèle, nous faisons des hypothèses fortes concernant la nature des données. Par exemple, la loi générative du modèle d'erreur suppose que les données sont gaussiennes. En réalité, nous ne savons rien sur

la nature des données, et bien que la loi gaussienne est très répandue dans la nature, il est nécessaire de critiquer notre modèle pour l'affiner et obtenir de meilleures prédictions.

Considérons un exemple simple où nos observations ne sont pas issues du modèle que nous considérons. Sans le savoir, nous faisons donc une (inévitable) erreur de modélisation. Dans notre exemple, les $n = 20$ données que nous observons sont en fait issues de la loi de Cauchy. La loi de Cauchy à la particularité de n'admettre ni espérance ni écart-type. En langage `R`, nous pouvons créer des observations de la manière suivante

```
y = rcauchy(n)
```

Pour tester notre modèle d'erreur, nous pouvons choisir une statistique de test. Il y a un grand choix, et nous considérons une possibilité parmi tant d'autres. Nous testons notre modèle en utilisant le coefficient de dissymétrie de l'échantillon. Pour nos n observations, ce coefficient est égal à -1.23 (attention, ce résultat n'est pas reproductible !). L'algorithme de Monte Carlo suivant, programmé en langage `R` et disponible parmi les scripts associés au cours, simule la distribution prédictive du coefficient de dissymétrie (skewness) pour nos 20 observations.

```
post.pred = NULL
for (i in 1:1000) {
ind = sample(10000, 1)
post.pred[i] = skewness(rnorm(20, m[ind], sqrt(sigma.2[ind]))) }
hist(post.pred)
skewness(y)
```

Nous observons que la valeur observée du coefficient de dissymétrie (-1.23) se trouve dans les quantiles les plus faibles de la loi prédictive de ce coefficient. Un test rejette donc le modèle d'erreur gaussien pour notre jeu de données. Le rejet du modèle gaussien remet donc en question la validité des estimations faites pour la variable cachée et les calculs d'incertitude liés à cette variable. Nous devons, pour ces données, nous efforcer de rechercher un nouveau modèle d'erreur. Notons que pour exécuter le script précédent, on devra écrire la fonction `skewness` permettant de calculer le coefficient de dissymétrie car celle ci n'est pas disponible dans la bibliothèque de fonctions de base. D'autres choix de statistique sont bien entendu possibles et conseillés !

4.5 Résumé

Résumer les points à retenir et donner quelques exemples illustrant les concepts principaux de la séance.

4.6 Exercices

Pour les exercices suivants, il est nécessaire d'avoir un ordinateur sur lequel on aura installé préalablement le logiciel `R`. Ce logiciel libre est disponible sur le site `http://cran.r-project.org/`.

Exercice 1. Modèle d'erreur gaussien. On considère un paramètre caché, θ de loi $\theta \sim \mathrm{N}(0, \sigma_0^2 = 1/\beta_0)$, et une observation y générée par la loi conditionnelle $y|\theta \sim \mathrm{N}(\theta, \sigma^2 = 1/\beta)$.

1. Montrer que l'on peut représenter le couple (θ, y) de la manière suivante

$$\begin{aligned} \theta &= x_0 \\ y &= x_0 + x_1 \end{aligned}$$

   où x_0 est de loi $\mathrm{N}(0, \sigma_0^2)$, et x_1 est indépendante de x_0 de loi $\mathrm{N}(0, \sigma^2)$.

2. Déduire de la question précédente, l'espérance et la matrice de covariance du couple (θ, y).

Exercice 2. On considère la loi générative suivante

$$p(y_1, \ldots, y_n|\theta) \propto \prod_{i=1}^{n} \exp\left(-\frac{1}{2\sigma^2}(y_i - \theta)^2\right), \quad y_i \in \mathbb{R}$$

où la moyenne, θ, est un paramètre inconnu. On suppose que la variance, σ^2, est connue, et que la loi a priori de θ est dégénérée (non-informative)

$$p(\theta) \propto 1\,.$$

On pose $\bar{y} = \sum_{i=1}^{n} y_i/n$.

1. Montrer que $p(\theta|y) = p(\theta|\bar{y})$ (On peut remarquer que $\bar{y}|\theta \sim \mathrm{N}(\theta, \sigma^2/n)$).

2. Retrouver que la loi a posteriori du paramètre θ est donnée par

$$\theta|y \sim \mathrm{N}(\bar{y}, \sigma^2/n)\,.$$

Exercice 3. On considère le modèle gaussien de loi générative

$$p(y_1, \ldots, y_n|\theta) \propto \frac{1}{\theta^{n/2}} \exp\left(-\frac{n}{2\theta} s_n^2\right)$$

où

$$s_n^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - m)^2\,.$$

Le paramètre θ représente la variance, qui est inconnue. On suppose que la moyenne m est connue et que la loi de θ est dégénérée (uniforme sur une échelle log)

$$p(\theta) \propto \frac{1}{\theta}\,.$$

On pose $\bar{y} = \sum_{i=1}^{n} y_i/n$.

1. On considère une variable aléatoire de loi χ^2_ν, dont la densité est donnée par

$$p(x) \propto x^{\nu/2-1}e^{-x/2}\,, \quad x > 0, \nu > 0\,.$$

   Soit $s^2 > 0$. Montrer que la loi de la variable $\nu s^2/\chi^2_\nu$ admet pour densité

$$p(x) \propto \frac{1}{x^{\nu/2+1}} e^{-\nu s^2/2x}\,, \quad x > 0\,.$$

   Cette loi est notée Inv-$\chi^2(\nu, s^2)$.

2. Retrouver que la loi a posteriori du paramètre $\theta = \sigma^2$ est donnée par

$$\theta|y \sim \mathrm{Inv}{-}\chi^2(n, s_n^2)$$

Exercice 4. Dans cet exercice, m et σ^2 sont des paramètres inconnus et aléatoires. On considère donc $\theta = (m, \sigma^2)$, de loi a priori non-informative

$$p(m, \sigma^2) \propto \frac{1}{\sigma^2}\,.$$

1. Retrouver l'expression de la loi a posteriori

$$p(m, \sigma^2|y) \propto \frac{1}{(\sigma^2)^{n/2+1}} \exp\left(-\frac{1}{2\sigma^2}((n-1)s_{n-1}^2 + n(\bar{y}-m)^2)\right)$$

où s_{n-1}^2 est *l'estimateur sans biais* de la variance :

$$s_{n-1}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2\,.$$

2. Démontrer que la loi marginale a posteriori de σ^2 est

$$\sigma^2|y \sim \text{Inv}\chi^2(n-1, s_{n-1}^2)\,.$$

3. Justifier la validité de l'algorithme de simulation de la loi a posteriori suivant

```
sigma.2 = (n-1)*var(y)/rchisq(10000, n-1)
m = rnorm(10000, mean(y), sd = sqrt(sigma.2/n))
```

4. Programmer cet algorithme en langage `R`, et l'utiliser pour estimer les paramètres de moyenne et variance des longueurs des sépales des iris de Fisher (`data(iris)`). Comparer les résultats obtenus aux résultats théoriques.

5. Que penser de la validité de l'algorithme de simulation de la loi a posteriori répétant les opérations suivantes depuis une condition initiale fixée

```
sigma.2 = sum((y -m)^ 2)/rchisq(1, n)
m = rnorm(1, mean(y), sd = sqrt(sigma.2/n))
```

6. Programmer l'algorithme précédent en langage `R`, et comparer les résultats obtenus aux résultats théoriques pour le même jeu de données que précédemment.

5 Séance 5 : Méthodes de Monte-Carlo par chaînes de Markov

5.1 Introduction

Dans cette séance, nous introduisons une classe de méthodes de Monte-Carlo très utile pour la simulation d'une loi de probabilité dont on ne connaît que le terme général et dont on ne sait pas calculer la constante de normalisation. Pour l'apprentissage probabiliste, cela concerne notamment la loi a posteriori du paramètre θ,

$$p(\theta|y) \propto p(y|\theta)p(\theta)\,,$$

dont la constante de normalisation est donnée par une intégrale qu'il est généralement impossible de calculer exactement

$$p(y) = \int p(y|\theta)p(\theta)d\theta\,.$$

Le principe des méthodes de Monte-Carlo que nous présentons ici repose sur l'utilisation d'un **processus itératif** permettant de générer de manière **approchée** des tirages d'une loi cible quelconque. Nous décrivons dans la section suivante la nature des itérations considérées pour la simulation de la loi cible. En théorie des probabilités, ces processus aléatoires rentrent dans le cadre très étudié des **chaînes de Markov**.

5.2 Notations et définitions

Nous appelons **chaîne de Markov** une suite de variables aléatoires, (θ^t), $t = 0, 1, 2, \ldots$, à valeurs dans un espace d'état E discret ou continu, vérifiant la propriété suivante

$$p(\theta^{t+1}|\theta^t, \cdots, \theta^0) = p(\theta^{t+1}|\theta^t)\,.$$

La chaîne est homogène dans le temps si, de plus,

$$p(\theta^{t+1} = \theta' | \theta^t) = \theta) = p(\theta^1 = \theta' | \theta^0 = \theta)\,.$$

Pour une chaîne de Markov, la loi conditionnelle de la variable θ^{t+1} sachant le passé du processus au temps t ne dépend du passé qu'à travers la dernière observation du processus. Cette propriété est souvent appelée la *propriété de Markov*. La chaîne est homogène si le mécanisme de transition, aussi appelé **noyau de transition**, permettant de générer θ^{t+1} à partir de θ^t ne dépend pas de t.

Dans la suite, nous ne considérons que des chaînes de Markov homogènes dans le temps. Nous notons $k(\theta^0, \theta^1)$ le noyau (*kernel*) de transition

$$k(\theta^0, \theta^1) = p(\theta^1 | \theta^0)\,.$$

Lorsque θ^0 est fixé, un noyau de transition définit donc une loi de probabilité pour la variable θ^1.

Lorsque E est un ensemble d'états fini, nous pouvons numéroter les états. La valeur $k(i, j)$ représente la probabilité que le processus effectue une transition vers l'état j lorsqu'il se trouve dans l'état i. Nous avons donc, pour tout couple (i, j) dans E,

$$0 \leq k(i, j) \leq 1 \quad \text{et} \quad \sum_{j \in E} k(i, j) = 1.$$

La matrice $K = (k(i, j))$ dont le terme général est donné par le noyau de transition s'appelle la **matrice de transition**. Nous étudions les propriétés des matrices de transition et illustrons quelques exemples simples correspondant à d'autres contextes d'applications que l'apprentissage dans une séance ultérieure du cours de MPA.

Lorsque E est un ensemble continu, $k(\theta^0, \theta^1)$ peut définir, pour chaque valeur connue θ^0, une loi de probabilité quelconque sur E . Dans ce cas, un noyau de transition est une fonction positive dont l'intégrale vaut 1

$$k(\theta^0, \theta^1) \geq 0 \quad \text{et} \quad \int_E k(\theta^0, \theta^1) d\theta^1 = 1.$$

La propriété de Markov permet de calculer les lois "fini-dimensionnelles" de la suite (θ^t). En effet, nous avons

$$p(\theta^0, \ldots, \theta^t) = p(\theta^t | \theta^0, \ldots, \theta^{t-1}) p(\theta^0, \ldots, \theta^{t-1}) \,.$$

Par définition, nous avons

$$p(\theta^t | \theta^0, \ldots, \theta^{t-1}) = k(\theta^{t-1}, \theta^t) \,.$$

Par une récurrence élémentaire, nous obtenons que

$$p(\theta^0, \ldots, \theta^t) = p(\theta^0) k(\theta^0, \theta^1) \cdots k(\theta^{t-1}, \theta^t) \,.$$

Par convention, nous notons $\pi(\theta) = p(\theta^0 = \theta)$ la loi initiale de la chaîne de Markov. Le résultat précédent nous indique que le comportement probabiliste d'une chaîne de Markov homogène est entièrement spécifié par la donnée de la loi initiale $\pi(\theta)$ et du noyau de transition $k(\theta^0, \theta^1)$.

Nous disons qu'une loi $\pi(\theta)$ est **stationnaire ou invariante** si

$$p(\theta^1 = \theta) = \pi(\theta) \,.$$

La loi $p(\theta^1)$ peut être calculée comme la loi marginale du couple (θ^0, θ^1)

$$p(\theta^1) = \int p(\theta^0, \theta^1) d\theta^0 = \int \pi(\theta^0) k(\theta^0, \theta^1) d\theta^0 \,.$$

Nous voyons que cette définition implique que la loi initiale $\pi(\theta)$ est stationnaire si elle solution d'une équation de point fixe

$$\pi(\theta^1) = \int \pi(\theta^0) k(\theta^0, \theta^1) d\theta^0 \,.$$

Il existe un cas particulièrement intéressant ou l'on peut vérifier la stationnarité d'une loi sans trop d'effort. On dit qu'une chaîne de Markov est réversible si le sens du temps n'a pas d'influence sur la loi de la chaîne. En d'autres termes, une **chaîne de Markov est réversible** si

$$p(\theta^0 = \theta, \theta^1 = \theta') = p(\theta^0 = \theta', \theta^1 = \theta) \,.$$

Cela se traduit par la condition

$$\pi(\theta^0) k(\theta^0, \theta^1) = \pi(\theta^1) k(\theta^1, \theta^0) \,.$$

Nous pouvons vérifier que la loi initiale d'une chaîne réversible sera invariante pour le noyau de transition. En effet, nous avons

$$\int \pi(\theta^0) k(\theta^0, \theta^1) d\theta^0 = \pi(\theta^1) \int k(\theta^1, \theta^0) d\theta^0 = \pi(\theta^1) \,.$$

Supposons que la chaîne de Markov (θ^t) admette une unique loi stationnaire $\pi^*(\theta)$. Nous admettrons que, sous des conditions faibles, la loi conditionnelle de la variable θ^t sachant la valeur initiale θ^0 converge vers la loi stationnaire

$$p(\theta^t = \theta | \theta^0) \to \pi^*(\theta) \,, \quad t \to \infty.$$

L'idée de la démonstration de ce résultat résulte du phénomène de point fixe déjà évoqué un peu plus haut. En effet, nous avons

$$p(\theta^{t+1=\theta} | \theta^0) = \int p(\theta^t | \theta^0) k(\theta^t, \theta) d\theta^t \,.$$

En passant à la limite dans chacun des termes de l'égalité ci-dessus et par homogénéité des transitions, nous obtenons que

$$\pi^*(\theta) = \int \pi^*(\theta^0) k(\theta^0, \theta) d\theta^0 \,.$$

Nous voyons donc, comme dans une suite récurrente classique, que si la limite existe et est unique alors il s'agit nécessairement d'une solution de point fixe et donc de la loi invariante de la chaîne de Markov.

Exemple. Considérons l'ensemble fini $E = \{(1,1),(1,2),(1,3),(2,2),(2,3),(3,3)\}$ dont les éléments sont rangés dans l'ordre. Soit $\theta^t = (\theta_1^t, \theta_2^t)$, la chaîne de Markov définie de la manière suivante.

La condition initiale est choisie aléatoirement de manière uniforme dans E. Pour tout $t \geq 1$, (θ_1^t, θ_2^t) est mis à jour de la manière suivante

1. Choisir X au hasard uniformément dans $\{1,2\}$.
2. Si $X = 1$, alors θ_1^{t+1} est tiré uniformément dans $\{1, \ldots, \theta_2^t\}$ et $\theta_2^{t+1} = \theta_2^t$.
3. Si $X = 2$, alors $\theta_1^{t+1} = \theta_1^t$ et θ_2^{t+1} est tiré uniformément dans $\{\theta_1^t, \ldots, 3\}$.

La matrice de transition de cette chaîne se calcule de manière explicite et nous obtenons

$$K = (k(i,j)) = \begin{pmatrix} * & 1/6 & 1/6 & 0 & 0 & 0 \\ 1/6 & * & 1/6 & 1/4 & 0 & 0 \\ 1/6 & 1/6 & * & 0 & 1/6 & 1/6 \\ 0 & 1/4 & 0 & * & 1/4 & 0 \\ 0 & 0 & 1/6 & 1/4 & * & 1/6 \\ 0 & 0 & 1/6 & 0 & 1/6 & * \end{pmatrix}.$$

Nous observons que le noyau de transition est symétrique

$$k(i,j) = k(j,i), \quad i,j = 1, \ldots 6$$

Nous voyons donc que la chaîne est réversible, et que la loi uniforme $\pi(i) = 1/6$ est invariante. Nous verrons un peu plus loin que cette chaîne de Markov correspond à un algorithme d'échantillonnage de Gibbs pour la loi uniforme sur E. La méthode employée pourra alors se généraliser pour échantillonner la loi uniforme sur un ensemble quelconque sans pour cela énumérer l'ensemble en question.

5.3 Algorithme de Metropolis

D'après les résultats de la section précédente, une chaîne de Markov, décrite par une loi initiale et un noyau de transition, peut être vue comme **un algorithme itératif permettant d'échantillonner de manière approchée la loi invariante de la chaîne**, supposée unique.

L'algorithme de Metropolis ou *algorithme de Metropolis-Hastings* propose une solution au **problème inverse**. A partir d'une loi cible $\pi(\theta)$ donnée, l'algorithme de Metropolis définit un noyau de transition, $k(\theta^0, \theta^1)$, de sorte que la chaîne associée à ce noyau de transition converge vers $\pi(\theta)$. La propriété remarquable de cet algorithme est de permettre d'**échantillonner la loi** $\pi(\theta)$ **lorsque la constante de normalisation de cette loi est incalculable**.

L'algorithme de Metropolis s'appuie sur un algorithme de rejet appliqué itérativement à la variable θ^t. Pour définir un algorithme de Metropolis, une première étape consiste à choisir un noyau de transition instrumental, $q(\theta^t, \theta^*)$ (proposal kernel). Le noyau instrumental décrit la manière d'explorer l'espace d'états à partir de la valeur courante θ^t. Son rôle/objectif est de proposer des valeurs qui seront évaluées ensuite comme étant acceptables ou non par l'algorithme de rejet.

L'algorithme de Metropolis se définit à partir des étapes suivantes.

1. Initialiser θ^0, $t = 0$
2. A l'itération t, tirer θ^* selon la loi instrumentale $q(\theta^t, \theta^*)$
3. Calculer
$$r = \frac{\pi(\theta^*)q(\theta^*, \theta^t)}{\pi(\theta^t)q(\theta^t, \theta^*)}$$
4. Avec la probabilité $\min(1, r)$, faire $\theta^{t+1} \leftarrow \theta^*$, sinon $\theta^{t+1} \leftarrow \theta^t$.
5. Incrémenter t et retourner en 2.

L'algorithme décrit ci-dessus peut être modélisé par une chaîne de Markov de noyau de transition

$$k(\theta^0, \theta^1) = q(\theta^0, \theta^1) \min\left(1, \frac{\pi(\theta^1)q(\theta^1, \theta^0)}{\pi(\theta^0)q(\theta^0, \theta^1)}\right), \quad \text{pour } \theta^1 \neq \theta^0$$

et

$$k(\theta^0, \theta^0) = 1 - \int_{\theta^1 \neq \theta^0} k(\theta^0, \theta^1) d\theta^1 .$$

L'expression de $k(\theta^0, \theta^1)$ signifie que l'on utilise la loi instrumentale pour proposer une valeur θ^1, puis cette valeur est acceptée avec la probabilité $\min(1, r)$.

On vérifie que la chaîne initialisée par la loi cible $\pi(\theta)$ possède la propriété de réversibilité

$$\pi(\theta^0)k(\theta^0, \theta^1) = \pi(\theta^1)k(\theta^1, \theta^0) .$$

En effet, supposons que $r < 1$, et calculons les termes apparaissant dans chacun des membres de cette équalité. Puisque $r < 1$, nous avons

$$\pi(\theta^0)k(\theta^0, \theta^1) = \pi(\theta^0)q(\theta^0, \theta^1)r,$$

et, cela se simplifie de la manière suivante

$$\pi(\theta^0)k(\theta^0, \theta^1) = \pi(\theta^1)q(\theta^1, \theta^0).$$

Concernant le second membre, le rapport est inversé et nous avons

$$\pi(\theta^1)k(\theta^1,\theta^0) = \pi(\theta^1)q(\theta^1,\theta^0)\min(1,1/r).$$

Cela se simplifie de la manière suivante

$$\pi(\theta^1)k(\theta^1,\theta^0) = \pi(\theta^1)q(\theta^1,\theta^0),$$

et la propriété de réversibilité est ainsi prouvée.

En conséquence, si le noyau instrumental est choisi de manière raisonnable et de sorte à visiter l'ensemble des valeurs à échantillonner, nous avons la garantie que l'algorithme de Metropolis simule des variables aléatoires dont la loi est très proche de la loi cible. En pratique, le choix de la loi instrumentale détermine la qualité de l'approximation faite par l'algorithme de Metropolis. Il sera prudent de ne conserver que les valeurs générées après une phase de *chauffe* de l'algorithme, difficile à calibrer, car dépendante du choix du noyau $q(\theta^0,\theta^1)$. Cette phase, nécessaire pour "atteindre" la loi stationnaire est appelée *burn-in period* en anglais.

5.4 Apprentissage d'une fréquence : modèle bêta-binomial

Dans le contexte de l'apprentissage probabiliste, l'algorithme de Metropolis est très utile pour simuler la loi a posteriori $p(\theta|y)$ et évaluer les grandeurs d'intérêt (histogrammes, quantiles, prédiction) par la méthode de Monte-Carlo.

Afin d'illustrer le comportement de l'algorithme de Metropolis, nous l'appliquons à un cas où la loi a posteriori peut être déterminée explicitement : l'apprentissage d'une fréquence, vu lors de la séance 2. Nous considérons que le paramètre qui nous intéresse est une proportion θ, correspondant à la fréquence d'émission d'une source binaire. Nous supposons que la loi *a priori* est uniforme

sur $(0,1)$, c'est à dire, $p(\theta) = 1$. En supposant que nous observons y signaux égaux à 1 et $n - y$ signaux égaux à 0, la loi générative est la loi binomiale de paramètres n et θ

$$p(y|\theta) = \text{binom}(n, \theta)(y) \propto \theta^y (1-\theta)^{n-y} .$$

La loi a posteriori est une loi de la famille bêta

$$p(\theta|y) = \text{beta}(y+1, n+1-y)(\theta) ,$$

dont la constante de normalisation est connue exactement.

De nombreux choix se présentent pour la loi instrumentale. Nous verrons en travaux dirigés qu'un choix raisonnable consiste à utiliser la loi a priori $q(\theta^0, \theta^1) = p(\theta^1)$. Dans cette section, nous étudions un autre choix, appelé la *marche aléatoire*. L'idée de la marche aléatoire est de proposer de nouvelles valeurs autour de la dernière valeur acceptée par l'algorithme de Metropolis. On entend par là que de telles valeurs auraient plus de chances d'être acceptées que des valeurs proposées au hasard. L'objectif de ce choix est donc d'augmenter le nombre d'acceptations, et donc la vitesse de convergence de l'algorithme de Metropolis.

Pour modéliser une marche aléatoire dans l'intervalle $(0,1)$, nous pouvons utiliser la loi bêta. Par exemple, choisissons $b = 100$ et

$$q(\theta^0, \theta^1) = \text{beta}(b\theta^0/(1-\theta^0), b)(\theta^1)$$

L'espérance (conditionnelle) de θ^1 sachant θ^0 est égale à

$$\text{E}[\theta^1|\theta^0] = \frac{\theta^0/(1-\theta^0)}{\theta^0/(1-\theta^0)+1} = \theta^0$$

La valeur proposée est donc centrée autour de la valeur courante. L'écart-type peut être calculé exactement, nous retiendrons qu'il est faible. Cette loi instrumentale implante donc une recherche locale.

En langage `R`, le rapport r intervenant dans l'algorithme de Metropolis peut se calculer de la manière suivante

```
theta.1<-rbeta(1, b*theta.0/(1-theta.0), b)
ratio1<-(theta.1/theta.0)^  y*((1 - theta.1)/(1 - theta.0))^ (n-y)
ratio2<-dbeta(theta.0,b*theta.1/(1-theta.1),b)/dbeta(theta.1,
b*theta.0/(1-theta.0),b)
ratio <- ratio1*ratio2
```

Nous programmerons l'algorithme de Metropolis en séance de travaux dirigés. Par exemple, pour $n = 20$ et $y = 9$, la loi a posteriori est la loi bêta$(10, 12)$, dont la densité est représentée en rouge dans la figure 10. L'algorithme est initialisé par une valeur proche de 0. Lorsque nous conservons 1000 tirages après une période de chauffe de 100 itérations, nous obtenons une approximation de la loi bêta$(10, 12)$ de très bonne qualité. Si toutefois nous utilisons uniquement les 150 premiers tirages, nous observons que l'algorithme n'a pas convergé, et que l'approximation est de très mauvaise qualité. Cet exemple montre que la durée de la période de chauffe est importante pour la bonne convergence de l'algorithme de Metropolis. La figure montre aussi l'influence du noyau instrumental. Une exploration fondée sur des pas aléatoires aura tendance à générer des plateaux de stagnation et un taux de rejet important. Une recherche fondée sur une marche aléatoire corrige ce défaut, mais la convergence vers le régime stationnaire sera plus lente.

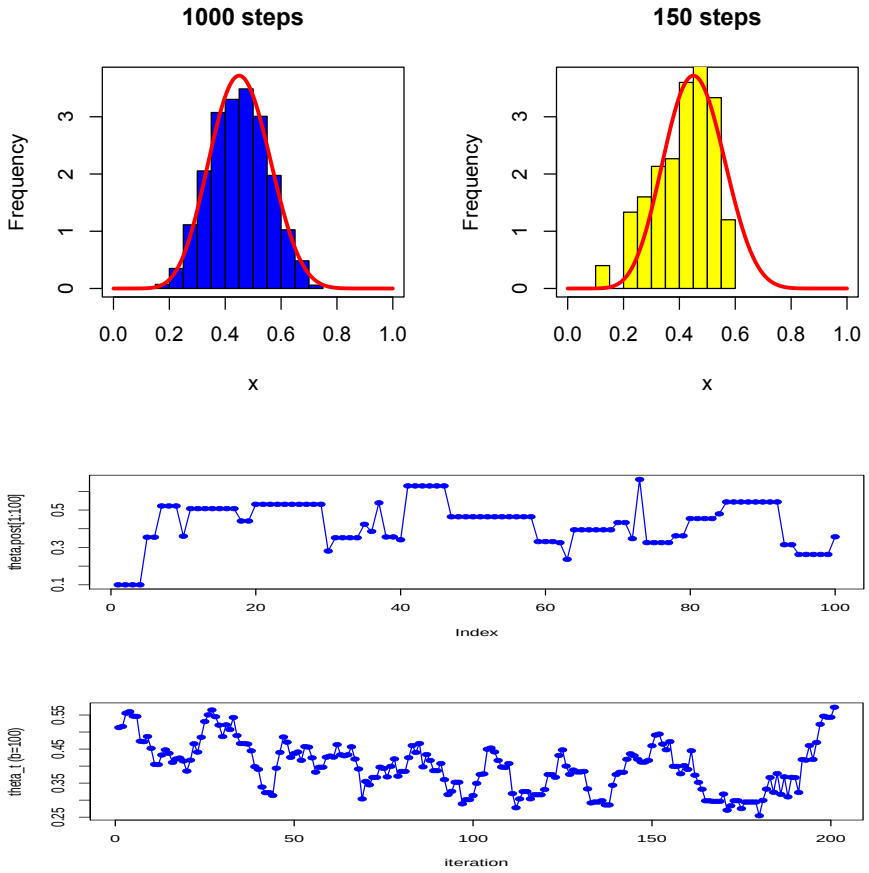


FIGURE 10 – *Histogrammes de la loi a posteriori obtenus par un algorithme de Metropolis pour la fréquence d'émission d'une source binaire* ($n = 20, y = 9$). *L'algorithme est initialisé par une valeur proche de* 0. *En haut à gauche, nous conservons 1000 tirages après élimination des 100 premiers. A droite, nous conservons les 150 premiers tirages. Les graphiques suivants représentent une trajectoire de la chaîne de MH pour un noyau instrumental uniforme (aléatoire), et pour une marche aléatoire de petite variance.*

5.5 Résumé

Résumer les points à retenir et donner quelques exemples illustrant les concepts principaux de la séance.

5.6 Exercices

Pour les exercices suivants, il est nécessaire d'avoir un ordinateur sur lequel on aura installé préalablement le logiciel `R`. Ce logiciel libre est disponible sur le site `http://cran.r-project.org/`.

Exercice 1. Dans cette séance, on considère le modèle de vraisemblance binomiale et de loi a priori choisie dans la famille de lois bêta (modèle bêta-binomial). On note θ le paramètre correspondant à une probabilité inconnue d'un événement donné, et on suppose que l'on observe $y = 9$ réalisations de cet événement lors de la répétition de $n = 20$ épreuves indépendantes. L'objectif de cette séance est de programmer plusieurs algorithmes de Monte Carlo par chaînes de Markov et d'évaluer la convergence de ces algorithmes.

1. On suppose que la loi a priori est uniforme sur $(0, 1)$. Donner un argument de théorie de l'information pour justifier que ce choix exprime le maximum d'incertitude a priori sur θ. Rappeler les expressions mathématiques de la loi générative, de la loi bêta et de la loi a posterior du paramètre θ, aux constantes près.
2. Rappeler l'algorithme d'estimation de Metropolis-Hasting (MH) pour un noyau instrumental $q(\theta, \theta^*)$.
3. On choisit ce noyau de transition égal à la loi a priori (les tirages ne dépendent pas de la valeur en cours θ). Vérifier que le rapport de MH est égal à :
$$r = \left(\frac{\theta^*}{\theta^t}\right)^y \times \left(\frac{1-\theta^*}{1-\theta^t}\right)^{(n-y)}$$
4. Ecrire un script en `R` implantant la simulation de la loi a posteriori

```
theta.1 <- runif(1)
ratio <- (theta.1/theta.0)^ y *((1 - theta.1)/(1 - theta.0))^ (n-y)
if (runif(1) < ratio) theta.0 <- theta.1
```

5. Afficher un histogramme de la loi a posteriori, calculer numériquement un intervalle de crédibilité à 95% pour θ et afficher un histogramme de la loi predictive a posteriori sachant y.

6. Testons maintenant l'algorithme implanté dans le script `R` suivant

```
#Metropolis-Hastings

#kernel = random walk

nit = 1000

theta.0 <- 0.1

theta.post <- theta.0

log.likelihood <- NULL


for (i in 1:nit) {

# proposal

b <- 100

theta.1 <- rbeta(1,b*theta.0/(1-theta.0), b)


# MH ratio

ratio1 <- (theta.1/theta.0)^(y)* ((1 - theta.1)/(1 - theta.0))^(n-y)

ratio2 <-  dbeta(theta.0, b*theta.1/(1-theta.1), b)

/dbeta(theta.1, b*theta.0/(1-theta.0), b)

ratio <- ratio1*ratio2


if (runif(1) < ratio) {theta.0 <- theta.1}

theta.post <- c(theta.post,theta.0)

log.likelihood <- c(log.likelihood, y*log(theta.0) + (n-y)*log(1 - theta.0))

}
```

```
# show results
u <- seq(0, 1, length= 200)
par(mfrow=c(1,3))
plot(u, dbeta(u, 10, 12), type="n", ylab="Frequency",
 main="Posterior distribution (burnin = 200)", xlab = "x")
hist(theta.post[200:1000], col = "blue" , prob = T, add = T)
lines(u, dbeta(u, 10, 12), col = 2, lwd = 3)
plot(u, dbeta(u, 10, 12), type="n", ylab="Frequency",
main="150 sweeps", xlab = "x")
hist(theta.post[1:150], col = "yellow" , prob = T, add = T)
lines(u, dbeta(u, 10, 12), col = 2, lwd = 3)
plot(log.likelihood[1:100], cex = .5, pch = 19, type = "l")
```

7. Quel est le choix de du noyau de transition instrumental $q(\theta^0, \theta^1)$ dans l'algorithme précédent ?

8. Donner l'espérance et la variance de θ^1 sachant la valeur en cours θ^0 (les valeurs pour la loi bêta(a, b) sont $a/(a+b)$ et $ab/(a+b)^2(a+b+1)$).

9. Commenter les résultats affichés par le script `R`. L'algorithme est-il sensible au choix de la valeur initiale (la modifier) ? L'algorithme est-il sensible au choix de la loi instrumentale ? Comment vos observations se traduisent-elles pour le choix de la période de chauffe (*burn-in*) ?

6 Séance 6 : Modèles multi-paramétriques – Echantillonnage de Gibbs

6.1 Introduction

Dans la séance précédente, nous avons introduit un algorithme de Monte Carlo par chaîne de Markov – l'algorithme de Metropolis – permettant d'échantillonner la loi d'une variable θ. L'algorithme de Metropolis est très utile lorsque les constantes de normalisation des lois sont inconnues. Il possède toutefois un inconvénient majeur. Il demande de choisir une loi instrumentale, et si le choix de cette loi n'est pas pertinent, le taux d'acceptation de l'algorithme de rejet peut être faible.

Dans cette section, nous présentons un **nouvel algorithme de simulation par chaîne de Markov**, s'adressant plus particulièrement à la situation d'un paramètre composite ou multi-dimensionnel, $\theta = (\theta_1, \theta_2)$, de loi cible $\pi(\theta_1, \theta_2)$. Cet algorithme est appelé **l'échantillonneur de Gibbs**. Il repose sur le calcul des lois conditionnelles $\pi(\theta_1|\theta_2)$ et $\pi(\theta_2|\theta_1)$, utilisées à tour de rôle comme lois instrumentales. L'échantillonneur de Gibbs se différencie de l'algorithme de Metropolis par le fait qu'il n'effectue aucun rejet.

6.2 Echantillonneur de Gibbs

Nous souhaitons simuler un paramètre composite $\theta = (\theta_1, \theta_2)$ de loi cible $\pi(\theta_1, \theta_2)$. Appliqué à l'apprentissage probabiliste, la loi cible $\pi(\theta)$ que nous considérons est, le plus souvent, la loi a posteriori, $p(\theta|y)$. Rappelons que pour simuler un paramètre composite $\theta = (\theta_1, \theta_2)$, un algorithme élémentaire comporte les deux étapes suivantes

1. Simuler θ_1 selon la loi marginale $\pi(\theta_1)$

2. Etant donné θ_1, simuler θ_2 selon la loi conditionnelle $\pi(\theta_2|\theta_1)$.

Afin d'implanter cet algorithme, il est nécessaire de connaître la loi marginale et de savoir la simuler. Cette loi est définie par

$$\pi(\theta_1) = \int \pi(\theta_1, \theta_2) d\theta_2 \, .$$

Nous voyons qu'intervient le calcul d'une intégrale, dont le calcul peut être impossible à mener à terme.

L'algorithme d'échantillonnage de Gibbs (ou *Gibbs Sampler*), est un algorithme itératif mettant à jour le paramètre $\theta^t = (\theta_1^t, \theta_2^t)$ en répétant le cycle suivant :

GS1. Etant donné θ_2^t, simuler θ_1^{t+1} à partir de la loi conditionnelle $\pi(\theta_1|\theta_2^t)$.

GS2. Etant donné θ_1^{t+1}, simuler θ_2^{t+1} à partir de la loi conditionnelle $\pi(\theta_2|\theta_1^{t+1})$.

Dans cet algorithme, la définition d'un cycle peut être récursive. Si θ_1 est lui-même un paramètre composite, alors nous pouvons mettre à jour ce paramètre en effectuant un sous-cycle du cycle principal. De même, nous pouvons généraliser l'algorithme à des paramètres composites de la forme $\theta = (\theta_1, \ldots, \theta_J)$, en considérant des cycles de longueur $J \geq 2$.

Exemple d'échantillonneur de Gibbs. On souhaite simuler la loi uniforme sur le domaine triangulaire $D = \{0 < \theta_2 < \theta_1 < 1\}$. En utilisant le générateur aléatoire de loi uniforme sur $(0, 1)$, `runif()` et une valeur initiale de θ_1 arbitraire dans $(0, 1)$, nous obtenons (exercice) le cycle de mise à jour suivant :

```
theta.2 = runif() * theta.1

theta.1 = theta.2 + runif()*( 1 - theta.2 )
```

Convergence de l'échantillonneur de Gibbs. L'échantillonneur de Gibbs peut être décrit par une chaîne de Markov dont le noyau de transition est donné par

$$k(\theta^0, \theta^1) = \pi(\theta_2^1|\theta_1^0)\pi(\theta_1^1|\theta_2^1)\,.$$

En effet, le produit des deux lois conditionnelles traduit le fait que nous simulons des variables de manière cyclique à partir de la composante courante du paramètre θ. Notons que seule la valeur initiale de la première composante, θ_1^0, intervient dans la définition du noyau de transition.

L'algorithme se justifie par le fait que la loi cible, $\pi(\theta_1, \theta_2)$, est stationnaire (invariante) pour la chaîne de Markov de noyau de transition $k(\theta^0, \theta^1)$. Cela signifie que la condition suivante est réalisée

$$\pi(\theta_1^1, \theta_2^1) = \int\int \pi(\theta_1^0, \theta_2^0)k(\theta^0, \theta^1)d\theta_1^0 d\theta_2^0\,.$$

Malgré l'aspect laborieux des notations (simplifiées au maximum), cela n'est pas difficile à vérifier. En effet, l'intégrale

$$\int\int \pi(\theta_1^0, \theta_2^0)\pi(\theta_2^1|\theta_1^0)\pi(\theta_1^1|\theta_2^1)d\theta_1^0 d\theta_2^0$$

est égale à

$$\int\int \pi(\theta_1^0, \theta_2^0)\frac{\pi(\theta_1^0, \theta_2^1)}{\pi(\theta_1^0)}\frac{\pi(\theta_1^1, \theta_2^1)}{\pi(\theta_2^1)}d\theta_1^0 d\theta_2^0\,.$$

Dans cette expression, il suffit de décaler les dénominateurs vers la gauche pour obtenir le résultat suivant

$$\left(\int\int \frac{\pi(\theta_1^0, \theta_2^0)}{\pi(\theta_1^0)}\frac{\pi(\theta_1^0, \theta_2^1)}{\pi(\theta_2^1)}d\theta_1^0 d\theta_2^0\right)\,\pi(\theta_1^1, \theta_2^1)\,.$$

Puisque nous avons

$$\int \pi(\theta_2^0|\theta_1^0)\int \pi(\theta_1^0|\theta_2^1)d\theta_1^0 d\theta_2^0 = 1\,.$$

L'équation de stationnarité est prouvée.

Remarquons finalement que, pour justifier complètement la validité de l'algorithme, il est nécessaire que celui-ci converge vers la loi cible. L'invariance de la loi cible ne garantit pas nécessairement la convergence, et il est possible d'exhiber des cas pathologiques où l'échantillonneur de Gibbs ne converge pas. Cela peut arriver lorsque la chaîne ne peut pas explorer l'ensemble de l'espace d'états (on qualifie alors la chaîne de l'adjectif *réductible*). Voir un contre-exemple ci-dessous. Nous retiendrons que ces cas sont rares, et qu'en pratique nous pouvons les détecter en simulant l'algorithme.

Exemple de chaîne non convergente. Considérons le domaine D, inclus dans le carré $[0,1]^2$ et constitué de l'union des domaines D_1 et D_2 définis par $D_1 = (0,1/2)^2$ et $D_2 = (1/2,1)^2$. La loi cible est la loi uniforme sur D. Nous remarquons que les lois conditionnelles sont aussi des lois uniformes (exercice). Si l'échantillonneur de Gibbs est initialisé dans D_1 alors la dynamique de la chaîne est confinée au domaine D_1. Si l'échantillonneur de Gibbs est initialisé dans D_2 alors la dynamique de la chaîne est confinée au domaine D_2. La chaîne est donc non convergente car sa dynamique dépend de la condition initiale choisie.

6.3 Echantillonneur de Gibbs stochastique : Un cas particulier de l'algorithme de Metropolis-Hasting

On montre dans cette section que l'échantillonneur de Gibbs peut être modifié pour ne pas suivre un cycle déterministe de mise à jour des composantes du paramètre θ. En effet, nous considérons une mise à jour aléatoire, où la coordonnée à modifier est choisie de manière uniforme. Cette version de l'algorithme d'échantillonnage de Gibbs est en fait **un cas particulier de l'algorithme de**

Metropolis. Il a le mérite d'être optimal, au sens où le taux d'acceptation est de 100% ($r = 1$).

La mise à jour aléatoire correspond à un noyau de transition instrumental de la forme suivante

$$q((\theta_1, \theta_2), (\theta_1^*, \theta_2)) = \frac{1}{2}\pi(\theta_1^*|\theta_2)$$

et

$$q((\theta_1, \theta_2), (\theta_1, \theta_2^*)) = \frac{1}{2}\pi(\theta_2^*|\theta_1)$$

Dans la première option, on choisit de modifier la coordonnée θ_1 avec la probabilité 1/2 puis on propose θ_1^* en tirant selon la loi conditionnelle $\pi(\theta_1^*|\theta_2)$. Dans la deuxième option, on choisit de modifier la coordonnée θ_2 avec la probabilité 1/2 puis on propose θ_2^* en tirant selon la loi conditionnelle $\pi(\theta_2^*|\theta_1)$. Il n'y a pas d'autre modification autorisée.

Pour ce noyau de transition instrumental, nous pouvons calculer le rapport de Metropolis, r, définit dans la séance précédente. Supposons que le changement se produise sur la première coordonnées (le second cas est symétrique). Dans ce cas, nous avons

$$r = \frac{\pi(\theta_1^*, \theta_2)}{\pi(\theta_1, \theta_2)} \frac{\pi(\theta_1|\theta_2)}{\pi(\theta_1^*|\theta_2)}\,.$$

En développant la loi jointe, nous avons

$$r = \frac{\pi(\theta_1^*|\theta_2)\pi(\theta_2)}{\pi(\theta_1|\theta_2)\pi(\theta_2)} \frac{\pi(\theta_1|\theta_2)}{\pi(\theta_1^*|\theta_2)}\,,$$

et la fraction se simplifie

$$r = 1.$$

L'algorithme de Gibbs peut donc être interprétée comme une version idéale

de l'algorithme de simulation de Metropolis. Dans l'échantillonneur de Gibbs, toutes les transitions proposées seront acceptées.

6.4 Un exemple

À Las Vegas, il y a une proportion inconnue, θ, de tricheurs. Lorsque l'on joue à *pile* ou *face* contre un tricheur, il est impossible de gagner. Dans le cas contraire, on gagne avec la probabilité 1/2. On joue contre n personnes choisies de manière indépendante, et on perd y parties ($0 < y \leq n$). Sans information préalable sur la variable θ, on suppose qu'elle est uniformément répartie entre 0 et 1, $p(\theta) = 1$. Le but de cet exemple est de simuler la loi a posteriori $p(\theta|y)$ en utilisant l'échantillonneur de Gibbs.

L'idée est que le problème est simplifié par la considération d'une variable supplémentaire, z, correspondant au nombre de tricheurs rencontrés lors des n parties jouées. Nous considérons donc le paramètre augmenté (θ, z), et cherchons à déterminer et à simuler la loi jointe $p(\theta, z|y)$. La loi a posteriori se déduit alors par marginalisation. La simulation de la loi jointe permet aussi d'approcher la loi a posteriori de la variable z. Nous déterminerons ainsi la loi du nombre de tricheurs rencontrés sachant que l'on a perdu y parties.

Le modèle devient un modèle hiérarchique, pour lequel

$$p(z|\theta) = \text{bin}(z|n, \theta), \quad z = 0, \ldots, n.$$

et

$$p(y|z) = \text{bin}(y - z|n - z, 1/2), \quad y = z, \ldots, n.$$

où bin dénote la loi binomiale. En effet, si on rencontre z tricheurs, y est supérieur à z. La différence entre ces 2 valeurs représente le nombre de défaites

à la régulière. Cette différence est de loi binomiale de paramètres $n - z$ et $1/2$.

En appliquant la formule de Bayes,

$$p(z|y,\theta) \propto p(y|z,\theta)p(z,\theta) = p(y|z)p(z|\theta)p(\theta).$$

En ne gardant que les termes dépendant de la variable z et utilisant le résultat suivant

$$\binom{n-y}{n-z}\binom{n}{z} = \binom{y}{z}$$

nous obtenons

$$p(z|y,\theta) = c\binom{y}{z}\left(\frac{1-\theta}{2}\right)^{y-z}\theta^z\,.$$

Le calcul de la constante c conduit au résultat suivant

$$c = \sum_{z=0}^{y}\left(\binom{y}{z}\left(\frac{1-\theta}{2}\right)^{y-z}\theta^z\right)^{-1} = \left(\frac{2}{1+\theta}\right)^{y}.$$

Nous ainsi obtenons la loi $p(z|y,\theta)$. Il s'agit de la loi binomiale suivante

$$p(z|y,\theta) = \text{bin}\left(z|y,\frac{2\theta}{(1+\theta)}\right), \quad z = 0,\ldots,y.$$

Par ailleurs, un calcul plus classique nous donne la loi conditionnelle $p(\theta|y,z)$: il s'agit de la loi bêta$(z+1, n-z+1)$. L'algorithme d'échantillonnage de Gibbs repose donc sur les 2 commandes `R` suivantes

```
theta = rbeta(1, z+1, n - z +1)
z = rbinom(1, y, 2*theta/(1+theta))
```

Remarquons que la loi de z sachant y peut aussi être décrite. En effet, nous avons

$$p(z|y) = \int_0^1 p(z,\theta|y)d\theta.$$

or

$$p(z,\theta|y) \propto p(z,y|\theta)p(\theta).$$

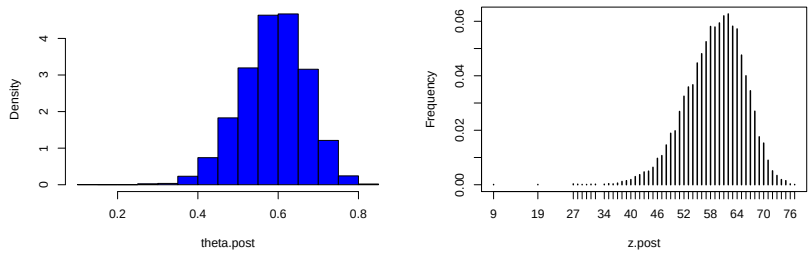
Après intégration et en utilisant les propriétés de la fonction bêta, nous avons

$$p(z|y) \propto \binom{n-z}{y-z} \left(\frac{1}{2}\right)^{y-z}, \quad 0 \leq z \leq y.$$

Cette loi peut être simulée grâce à la commande `sample` de `R`. Nous pouvons donc simuler de manière exacte la loi conditionnelle du couple (θ, z) sachant y. En deux étapes, simule tout d'abord z selon $p(z|y)$, puis on simule θ selon $p(\theta|y, z)$.

Notons enfin que la loi a posteriori du paramètre θ peut être caractérisée de manière explicite. La loi a posteriori de la variable θ correspond à la loi de $2\varphi - 1$ où φ est une variable aléatoire de loi bêta$(y+1, n-y+1)$ conditionnée à être supérieure à $1/2$ (voir Exercice 3, séance 2).

Par exemple, supposons que l'on joue 100 parties de pile ou face à Las Vegas et que l'on en perde 80. Nous pouvons prédire que la proportion de tricheurs est proche de 60%. Cette proportion se trouve dans l'intervalle $(0.42, 0.73)$ avec la probabilité .95, et le nombre de tricheurs rencontrés dans l'ensemble des 100 parties est compris entre 41 et 70 avec la probabilité .95 (l'espérance est égale à 59).



Nombre de tricheurs à Las Vegas. Simulation de l'échantillonneur de Gibbs pour le couple (z, θ) pour $y = 80$ parties perdues sur 100 parties jouées.

6.5 Modèles gaussiens

Reprenons le modèle d'erreur gaussien vu dans la séance précédente. Dans ce modèle, nous cherchons la loi a posteriori du paramètre composite $\theta = (m, \sigma^2)$ étant donné un vecteur de n observations, y. Rappelons qu'un **algorithme de simulation exact** de cette loi produisant 1000 variables est donné par les deux commandes de `R` suivantes :

```
sigma.2 = (n-1)*var(y)/rchisq(10000, n-1)
m = rnorm(10000, mean(y), sd = sqrt(sigma.2/n))
```

Pour ce paramètre bi-dimensionnel, nous avions aussi calculé la loi conditionnelle de la variance σ^2 sachant la moyenne m et y. Nous avions obtenu le résultat suivant

$$\sigma^2 | m, y \sim n s_n^2 / \chi_n^2$$

où s_n^2 est l'estimateur sans biais de la variance lorsque la moyenne est connue. Pour échantillonner la loi a posteriori dans le modèle d'erreur gaussien, il suffit donc d'itérer le cycle suivant

GS1. $\sigma^2 | m, y \sim n s_n^2 / \chi_n^2$

GS2. $m | \sigma^2, y \sim \mathrm{N}(\bar{y}, \sigma^2/n)$

où $\bar{y}$ correspond à la moyenne empirique. En langage `R`, la boucle principale de l'échantillonneur de Gibbs correspond aux commandes suivantes :

```
sigma.2 = sum((y -m)^ 2)/rchisq(1, n)
m = rnorm(1, mean(y), sd = sqrt(sigma.2/n))
```

Après 10100 cycles et après avoir écarté les 100 premières simulations, nous voyons, dans la figure 11 que le résultat est très similaire à celui obtenu par l'algorithme exact. Dans ce cas, l'algorithme de Gibbs possède de très bonnes

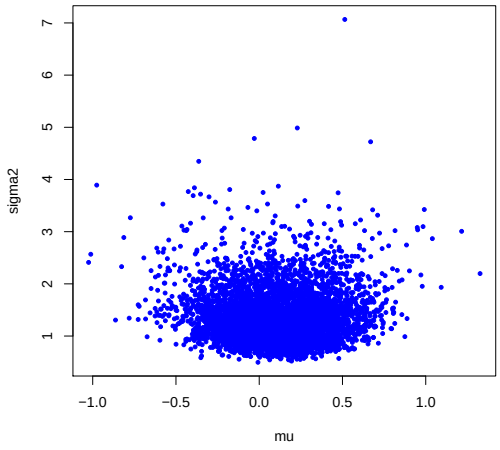


FIGURE 11 – *Simulation de l'échantillonneur de Gibbs pour le couple* (m, σ^2) *pour* $n = 20$ *observations de la loi* N*(0,1).*

propriétés de convergence liées à la très faible dépendance entre m et σ^2.

6.6 Reconstruction d'image

Nous traiterons l'exemple de la reconstruction d'image lors de l'exercice 4 des travaux dirigés. Dans ce cas, le paramètre à apprendre est une image binaire, θ, composée de $K \times L$ pixels à valeurs -1 ou $+1$. La variable observée est à nouveau une image binaire, y, que nous supposons générée à partir d'un modèle d'erreur connu.

L'originalité de la modélisation repose sur une forme de loi a priori très particulière, prenant en compte les dépendances locales qui peuvent exister au sein d'une image binaire. En effet, deux pixels proches de l'image observée ont plus de chance de présenter des niveaux d'intensité identiques que deux pixels éloignés. Cette propriété de mémoire locale est très similaire à la propriété de

Markov vue pour un processus itératif.

La reconstruction d'image consiste à échantillonner θ selon la loi a posteriori, $\mathrm{p}(\theta|y)$, dont les valeurs les plus probables représentent une version débruitée de l'observation y.

6.7 Résumé

Résumer les points à retenir et donner quelques exemples illustrant les concepts principaux de la séance.

6.8 Exercices

Pour les exercices suivants, il est nécessaire d'avoir un ordinateur sur lequel on aura installé préalablement le logiciel `R`. Ce logiciel libre est disponible sur le site `http://cran.r-project.org/`.

Exercice 1. Echantillonnage de Gibbs pour la loi uniforme. Décrire l'algorithme d'échantillonnage de Gibbs pour simuler la loi uniforme sur le disque unité, puis sur un sous-ensemble borné de $\mathbb{R}^2$.

Exercice 2. Iris de Fisher. Pour les longueurs des sépales des iris de Fisher, donner un histogramme des lois a posteriori ainsi que des intervalles de crédibilité à 95% pour les paramètres m et σ^2. Le modèle d'erreur gaussien vous semble-t-il approprié pour l'analyse de ces données ?

Exercice 3. Echantillonnage de Gibbs pour la loi normale. On considère un couple (θ_1, θ_2) de loi gaussienne de moyenne nulle et de matrice de covariance

$$\Lambda = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

On pose $\rho = .99$.

1. Ecrire un algorithme de simulation exact pour le couple (θ_1, θ_2). Le programmer en langage `R`. Que peut-on dire des résultats.
2. Rappeler le principe de l'algorithme d'échantillonnage de Gibbs.
3. Montrer que le script `R` donné ci-dessous implante cet algorithme et commenter sa sortie.

```
rho <- .99
nit <- 2000
theta1 <- -3; theta2 <- 3
theta1.post <- theta1
theta2.post <- theta2

for (i in 1:nit){
theta1 <- rnorm(1, rho*theta2, sd = sqrt(1 - rho^2))
theta2 <- rnorm(1, rho*theta1, sd = sqrt(1 - rho^2))
theta1.post <- c(theta1.post, theta1)
theta2.post <- c(theta2.post, theta2)
}

# show results
qqnorm(theta2.post[100:200])
abline(0,1)
plot(theta1.post[-1], theta2.post[-1])
cor(theta1.post, theta2.post)
hist(theta2.post)
```

4. La vitesse de convergence et les propriétés de mélange de la chaîne de Markov sont-elles dépendantes de ρ ?

Exercice 4 : Reconstruction d'image (d'après Geman et Geman) On considère un signal binaire de longueur n codé sous la forme d'un vecteur (θ_i),

où $\theta_i \in \{-1, +1\}$. Ce signal représente par exemple une image dont les pixels ont deux niveaux de couleur, noir et blanc. Dans ce cas, le codage en vecteur correspond à une matrice binaire ($n = K \times L$).

On observe une réalisation bruitée (y_i) du signal telle que $y_i \in \{-1, +1\}$ pour tout i. On suppose que le modèle probabiliste du bruit est décrit par

$$p(y|\theta) = \prod_{i=1}^{n} p(y_i|\theta_i)$$

où

$$p(y_i = +1|\theta_i = -1) = p(y_i = -1|\theta_i = +1) = (1 + \exp(2\alpha))^{-1}$$

et α est un paramètre réel supposé connu. On suppose que la loi de θ est décrite par un modèle d'Ising

$$p(\theta) \propto \exp(\beta \sum_{i \sim j} \theta_i \theta_j), \quad \beta > 0$$

où la notation $i \sim j$ signifie que i et j sont voisins en un sens à préciser. Pour une image, on peut préalablement définir un voisinage pour chaque pixel. Dans ce cas, cela signifie que l'on somme sur les paires de pixels voisins dans l'image. On suppose aussi que i n'est pas voisin de lui-même.

1. Démontrer la propriété suivante, pour tout $i = 1, \ldots, n$,

$$p(\theta_i|\theta_{-i}) = p(\theta_i|\theta_j,\, j \sim i) \propto \frac{\exp(\beta\theta_i \sum_{j\sim i} \theta_j)}{\exp(-\beta \sum_{j\sim i} \theta_j) + \exp(\beta \sum_{j\sim i} \theta_j)}$$

   Pour quel type d'image le modèle vous paraît-il adapté ?

2. Montrer que la loi a posteriori s'écrit de la manière suivante

$$p(\theta|y) \propto \prod_{i=1}^{n} \exp(\theta_i(\frac{\beta}{2} \sum_{j\sim i} \theta_j + \alpha y_i))$$

3. Ecrire un algorithme d'échantillonnage de Gibbs pour simuler la loi $p(\theta|y)$.

4. Télécharger l'image

   `http://membres-timc.imag.fr/Olivier.Francois/imagetd7.txt`

   et vérifier que l'algorithme suivant, utilisant des conditions de bord 'cycliques' et les 4 pixels les plus proches implante l'algorithme d'échantillonnage de Gibbs (les commandes `moins, plus` définissant les voisinages sont à écrire en langage `R`)

```
image(y<-as.matrix(read.table("imagetd7.txt")))
alpha=log(2);beta=10
theta=y
for (iter in 1:100){
for(i in 1:100)
for(j in 1:100)
{
sij =  theta[moins(i),moins(j)]+ theta[moins(i),plus(j)]+
        theta[plus(i),moins(j)]+ theta[plus(i),plus(j)]
pij = exp(beta*sij+alpha*y[i,j])
p   = exp(beta*sij+alpha*y[i,j])+exp(-beta*sij-alpha*y[i,j])
theta[i,j]=sample(c(-1,+1),1, prob=c(1-pij/p, pij/p)) }
image(theta)}
```

   Vérifier que $\alpha = \log 2$ correspond à environ 20% de pixels bruités. Faire varier la constante β vers des valeurs plus grandes que 10 et plus petites que 1. Qu'observez-vous ?

7 Séance 7 : Modèles hiérarchiques – Modèles de mélange.

7.1 Introduction et Définitions

Dans cette séance, nous rentrons au cœur des techniques d'apprentissage probabiliste en considérant des paramètres composites de grande dimension, dans des modèles structurés hiérarchiquement. Dans ces modèles, les paramètres sont inter-dépendants, et nous modélisons la structure de dépendance des paramètres de sorte à créer des cascades de dépendance. Les hiérarchies considérées sont particulièrement propices à l'utilisation de la formule de Bayes et aux algorithmes d'échantillonnage vus dans les séances précédentes.

En particulier nous abordons un domaine très important de l'apprentissage probabiliste : la **classification non-supervisée**. Ce domaine consiste à grouper des observations en classes séparées, sans avoir préalablement défini les classes et sans disposer d'exemples précis de classement. L'objectif de la classification probabiliste est de quantifier l'incertitude liée à chaque classement établi de cette manière. Pour cela, il est important de pouvoir calculer la probabilité conditionnelle qu'une observation donnée provienne de la classe k connaissant toutes les observations.

Considérons à nouveau en le complexifiant un peu le modèle d'erreur gaussien. Nous supposons désormais que nous observons des données issues en réalité de 2 sources distinctes et non d'une unique source. Par cohérence avec la figure 12, nous appelons les sources source *jaune* et source *bleue*. La valeur cachée émise par la source jaune est égale à $m_1 = -1$ et la valeur cachée émise par la source bleue est égale à $m_2 = 2$. Les écarts types des erreurs produites par

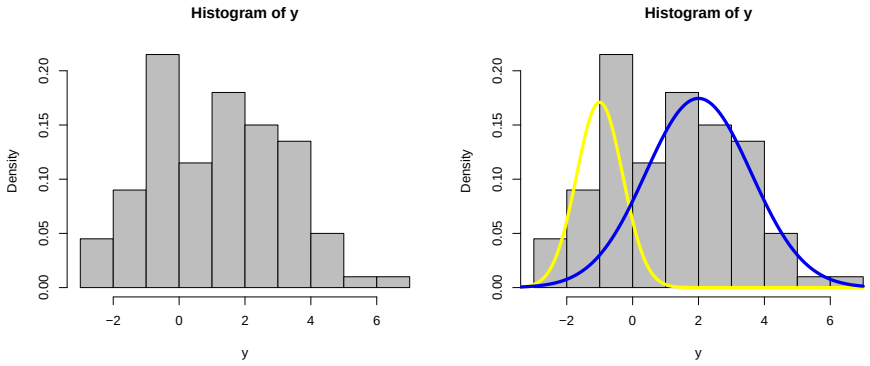


FIGURE 12 – *Observations d'un mélange de deux sources gaussiennes de paramètres respectifs* $m_1 = -1$ *et* $m_2 = 2$. *La partie droite de la figure représente les composantes cachées du mélange.*

les 2 sources sont aussi propres à chacune d'entre elles. De plus les fréquences d'émission des sources ne sont pas égales. La source jaune émet 30% du temps alors que la source bleue émet 70% du temps. Les sources sont aussi appelées *composantes* de mélange.

Bien entendu, les informations précédentes ne sont pas accessibles à un observateur. Ce dernier ne dispose que de données issues de l'observation des deux sources, représentées dans l'histogramme gris de la figure 12 (200 observations). En observant les données, nous ignorons la nature des sources. Le nombre de sources est aussi inconnu. Nous souhaitons donc apprendre les paramètres des sources, tout en nous posant la question suivante. Doit on expliquer les observations à l'aide d'un modèle ayant 1, 2 ou 3 sources ?

En résumé, la classification probabiliste cherche à répondre aux questions suivantes :

– Combien de groupes distincts peut on détecter au sein des observations ?

– Que valent les moyennes intra-groupes et les paramètres de dispersion des groupes détectés ?

– Pour une observation donnée, quelle est la probabilité de provenir de l'un ou l'autre de chaque groupe détecté ?

La classification probabiliste (ou *clustering*) répond à ce cahier des charges en considérant la classe associé à l'observation i, comme une variable non-observée ou latente, notée z_i. Aux paramètres z_i s'ajoutent les paramètres statistiques du modèle, incluant moyennes et variances des classes, résumés dans la variable composite θ. L'apprentissage probabiliste consiste à calculer ou simuler la loi a posteriori

$$p(\theta, z|y) \propto p(y|\theta, z)p(\theta)p(z)$$

et d'en déduire les lois marginales correspondant aux classes latentes $p(z_i|y)$. Notons enfin qu'un modèle à K classes possède $2K + n$ variables cachées, c'est à dire plus que le nombre des observations.

7.2 Exemple de modèle hiérarchique

La construction de modèle hiérarchique est particulièrement adaptée au développement d'algorithmes d'apprentissage. Un modèle hiérarchique décrit une loi de probabilité pour les données y et un paramètre composite que nous notons (θ, ψ) afin de mettre en évidence la structure hiérarchique. Le modèle est structuré de la manière suivante. Sans changement, les données y sont échantillonnées selon la loi générative

$$y|\theta, \psi \sim p(y|\theta)$$

Notons l'absence de dépendance de ψ dans cette définition. La loi de θ est définie conditionnellement à ψ

$$\theta|\psi \sim p(\theta|\psi)\,.$$

Le paramètre ψ est appelé *hyper-prior*. Il est échantillonné selon la loi

$$\psi \sim p(\psi)\,.$$

Ainsi, la loi a priori du paramètre composite (θ, ψ) est définie par $p(\theta, \psi) = p(\theta|\psi)p(\psi)$, et la loi a posteriori est donnée par la formule de chainage suivante

$$p(\theta, \psi|y) \propto p(y|\theta)p(\theta|\psi)p(\psi).$$

Les modèles hiérarchiques se prêtent bien à l'implantation de l'algorithme d'échantillonnage de Gibbs. En effet, nous avons les simplifications suivantes

$$p(\theta|\psi, y) \propto p(y|\theta)p(\theta|\psi)\,,$$

et

$$p(\psi|\theta, y) \propto p(y|\theta)p(\theta|\psi)p(\psi)\,.$$

Notons que les deux expressions font intervenir les mêmes grandeurs. Toutefois, les lois peuvent être très différentes losque l'on regarde la fonction de θ ou de ψ. Partant d'une condition initiale (θ^0, ψ^0), l'algorithme de Gibbs peut donc s'appuyer sur les cycles suivants :

Pour $t = 1, \ldots, T$

GS1 : Simuler θ^t selon $p(\theta|\psi^{t-1}, y)$,

GS2 : Simuler ψ^t selon $p(\psi|\theta^t, y)$

FinPour

7.3 Modèles de mélanges

Dans cette section, nous définissons les modèles de mélanges pour la classification non-supervisée, puis nous décrivons l'algorithme pour l'apprentissage des variables cachées de ces modèles.

Mélange de lois univariées. Considérons une unique observation, $y \in \mathbb{R}$. Une loi de mélange est définie comme la combinaison convexe de lois élémentaires, appelées **composantes du mélange**. Les **proportions** ou **coefficients** de mélange, $(p_k)_{k=1,\ldots,K}$, sont des nombres strictement positifs tels que

$$\sum_{k=1}^{K} p_k = 1 \,.$$

Les composantes du mélange peuvent être vues comme des lois génératives de paramètres θ_k, $k = 1, \ldots, K$. Nous définissons une loi de mélange de la manière suivante

$$p(y|\theta) = \sum_{k=1}^{K} p_k p(y|\theta_k)$$

Les paramètres des mélanges de gaussiennes correspondent à la moyenne et la variance des composantes, $\theta_k = (m_k, \sigma_k^2)$, et nous avons dans ce cas

$$p(y|\theta_k) = \mathrm{N}(y|m_k, \sigma_k^2) \,.$$

Les mélanges de lois gaussiennes sont facile à simuler. L'exemple représenté dans la figure 12 correspond à un mélange de lois gaussiennes, dans lequel nous avons $K = 2$ et les proportions de mélange sont égales à .3 et .7. La simulation de la figure 12 a été effectuée en échantillonnant exactement 30% des données depuis la source jaune. Pour un échantillon de taille totale $n = 200$, nous avons utilisé les commandes du langage `R` suivantes

```
z.truth = c(rep(1,60), rep(2, 140))

m.truth = c(-1, 2)

s.truth = c(.7, 1.6)

for (i in 1:200){

y[i]=rnorm(1,m.truth[z.truth[i]],sd=s.truth[z.truth[i]]) }
```

Pour un échantillon de taille n, $y = (y_1, \ldots, y_n)$, une approche naïve de l'apprentissage des paramètres consiste à écrire directement la loi générative du modèle

$$p(y|\theta) = \prod_{i=1}^{n} \left(\sum_{k=1}^{K} p_k p(y_i|\theta_k) \right)$$

Cette approche n'est pas inexacte, mais elle comporte de nombreux inconvénients algorithmiques. Pour l'apprentissage probabiliste dans les modèles de mélange, nous utilisons une technique d'**augmentation** des données, qui consiste à introduire une variable de classe cachée, $z_i \in \{1, \ldots, K\}$, associée à chacune des observations y_i. Cela nous amène à considérer, pour tout $i = 1, \ldots, n$, la loi générative suivante

$$p(y_i|\theta, z_i) = p(y_i|\theta_{z_i}).$$

Nous voyons que cette représentation nous conduit bien à un modèle de mélange :

$$p(y_i|\theta) = \sum_{k=1}^{K} p(y_i|\theta_k) p(z_i = k)\,.$$

Pour construire un modèle de mélange de lois gaussiennes, nous définissons donc un vecteur de variables cachées $z = (z_1, \ldots, z_n)$. Notre modèle (y, z, θ) s'écrit de la manière suivante

$$y_i|\theta, z_i = k \sim \mathrm{N}(y_i|m_k, \sigma_k^2)$$

$$p(z) \propto 1$$

$$p(\theta) \propto \tfrac{1}{\sigma_1^2} \tfrac{1}{\sigma_2^2} \cdots \tfrac{1}{\sigma_K^2}\,.$$

Dans cette expression, la loi uniforme $p(z)$ signifie simplement que $p(z_i = k) = 1/K$ pour i et pour tout $k \in \{1, \ldots, K\}$. Une alternative à la construction du modèle précédent est de considérer un modèle hiérarchique plus complexe, où les fréquences des sources sont aussi des paramètres

$$p(z_i = k) = p_k \, .$$

Pour la loi jointe des variables p_k, on choisit souvent la loi de Dirichlet, généralisant la loi bêta à plusieurs dimensions. Nous traiterons le modèle plus simple décrit en premier lieu.

7.4 Echantillonnage de Gibbs pour le modèle de mélange de lois gaussiennes

Afin d'écrire l'algorithme de Gibbs pour simuler la loi a posteriori du modèle de mélange de lois gaussiennes, nous devons considérer le paramètre composite suivant

$$(z, \theta) = (z_1, \ldots, z_n, m_1, \ldots, m_K, \sigma_1^2, \ldots, \sigma_K^2)$$

Ce paramètre comporte exactement $n + 2K$ dimensions. La loi a posteriori est proportionnelle à

$$p(z, \theta | y) \propto p(y|\theta, z) p(z) p(\theta) \, .$$

Les calculs peuvent se détailler de la manière suivante

$$\begin{aligned} p(z, \theta|y) &\propto \left(\prod_{i=1}^{n} p(y_i|\theta, z_i) p(z_i) \right) p(\theta) \\ &\propto \prod_{i=1}^{n} p(y_i|\theta, z_i) \prod_{k=1}^{K} \frac{1}{\sigma_k^2} \\ &\propto \prod_{i=1}^{n} \frac{1}{\sqrt{\sigma_{z_i}^2}} \exp(-\frac{1}{2\sigma_{z_i}^2}(y_i - m_{z_i})^2) \prod_{k=1}^{K} \frac{1}{\sigma_k^2} \end{aligned}$$

L'échantillonnage de Gibbs s'appuie sur les itérations d'un cycle que nous décomposons de la manière suivante

GS1. Pour $i = 1, \ldots, n$, mettre à jour $z_i \sim p(z_i|y_i, \theta)$ (mise à jour simultanée)

GS2. Pour $k = 1, \ldots, K$, mettre à jour $m_k \sim p(m_k|y, z, m_{-k}, \sigma^2)$ où le vecteur m_{-k} est privé de la composante m_k

$$m_{-k} = (\ldots, m_{k-1}, m_{k+1}, \ldots)$$

GS3. Pour $k = 1, \ldots, K$, mettre à jour $\sigma_k^2 \sim p(\sigma_k^2|y, z, m, \sigma_{-k}^2)$ où le vecteur σ_{-k}^2 est privé de la composante σ_k^2

$$\sigma_{-k}^2 = (\ldots, \sigma_{k-1}^2, \sigma_{k+1}^2, \ldots)\,.$$

Dans la suite de cette section, nous détaillons chacune des 3 étapes du cycle de l'échantillonnage de Gibbs.

Gibbs sampler GS1. Soit $z_i \in \{1, \ldots, K\}$. D'après la formule de Bayes, nous avons

$$p(z_i|y, \theta) = \frac{p(y|z_i, \theta)p(z_i|\theta)}{p(y|\theta)}$$

Or, nous avons

$$p(y|z_i, \theta) = p(y_i|z_i, \theta) \prod_{j \neq i} p(y_j|\theta)$$

et

$$p(y|\theta) = p(y_i|\theta) \prod_{j \neq i} p(y_j|\theta)\,.$$

Par indépendance, nous avons

$$p(z_i|\theta) = p(z_i) = \frac{1}{K}\,.$$

En reportant ces équations dans la formule de Bayes, et en utilisant la formule des probabilités totales, nous obtenons

$$p(z_i|y,\theta) = p(z_i|y_i,\theta) = \frac{p(y_i|z_i,\theta)}{p(y_i|\theta)} = \frac{p(y_i|z_i,\theta)}{\sum_k p(y_i|z_i=k,\theta)}\,.$$

En utilisant l'expression des densités gaussiennes, nous obtenons finalement

$$p(z_i|y_i,\theta) = \frac{\frac{1}{\sqrt{\sigma^2_{z_i}}}\exp(-\frac{1}{2\sigma^2_{z_i}}(y_i - m_{z_i})^2)}{\sum_{k=1}^K \frac{1}{\sqrt{\sigma^2_k}}\exp(-\frac{1}{2\sigma^2_k}(y_i - m_k)^2)}\,.$$

Le code `R` implantant cette phase du cycle utilise la commande `sample`. Cette commande prend en entrée un vecteur des probabilités exprimées à la constante de normalisation près

```
for (i in 1:n) {
p = exp(-(y[i] - m)^ 2/2/sigma2 )/sqrt(sigma2)
z[i] = sample(1:K, 1, prob=p) }
```

Gibbs sampler GS2. Les étapes GS2 et GS3 découlent de considérations vues dans les séances précédentes. Nous détaillerons les résultats en travaux dirigés. La phase GS2 concerne la simulation de la loi conditionnelle des moyennes intra-groupes sachant les variances (et les données). Soit n_k l'effectif courant de la classe k,

$$n_k = \#\{i : z_i = k\}\,,$$

supposé non-nul. Nous avons alors

$$p(m_k|y,z,m_{-k},\sigma^2) = \mathrm{N}(m_k|\bar{y}_k, \frac{\sigma^2_k}{n_k})$$

où $\bar{y}_k$ est la moyenne empirique calculée pour la classe k :

$$\bar{y}_k = \frac{1}{n_k}\sum_{i:z_i=k} y_i.$$

Ce résultat est identique à celui d'une séance précédente pour un échantillon gaussien de taille n_k et de variance connue. Nous le trouvons de la manière suivante :

$$p(m_k|y, z, m_{-k}, \sigma^2) \propto p(y|z, \theta)p(m_k|z, m_{-k}, \sigma^2).$$

Or m_k est indépendant de m_{-k} et de σ^2. Nous avons donc

$$p(m_k|z, m_{-k}, \sigma^2) = p(m_k|z) \propto p(z|m_k)p(m_k)\,.$$

EN remplaçant dans la formule précédente, nous obtenons

$$p(m_k|y, z, m_{-k}, \sigma^2) \propto \prod_{i:z_i=k} p(y_i|m_k, \sigma_k^2)\,.$$

Cela établi l'équivalence avec le résultat connu pour un échantillon de taille n_k.

Le code R implantant cette phase du cycle utilise la commande `rnorm`

```
nk[k]=sum(z==k)
m[k]=rnorm(1,mean(y[z==k]),sd=sqrt(sigma2[k]/nk[k]))
```

Gibbs sampler GS3. La phase GS3 concerne la simulation de la loi conditionnelle des variances intra-groupes sachant les moyennes (et les données). Nous avons

$$p(\sigma_k^2|y, z, m, \sigma_{-k}^2) = \text{Inv}{-}\chi^2(\sigma_k^2|n_k, s_k^2)$$

où s_k^2 représente la variance empirique au sein de la classe k lorsque la moyenne est connue

$$s_k^2 = \frac{1}{n_k} \sum_{i:z_i=k} (y_i - m_k)^2.$$

Le code R implantant cette phase du cycle utilise la commande `rchisq`.

```
sigma2[k] = sum((y[z==k]-m[k])^ 2)/rchisq(1, nk[k])
```

Finalement nous pouvons écrire une fonction en langage R permettant de simuler l'échantillonneur de Gibbs pour les mélanges de lois gaussiennes.

```
mcmc.mix=function(y,niter=1000,m.o= c(0,1),sigma2.o=c(1,1)) {
n=length(y);K=length(m.o);m=m.o;s2=sigma2.o
z=NULL;p=NULL;nk = NULL
p.mcmc=NULL;m.mcmc=NULL;sigma2.mcmc=NULL;z.mcmc=NULL
for(nit in 1:niter) {
for(i in 1:n) {
p=exp(-(y[i]-m)^ 2/2/s2)/sqrt(s2)
z[i]=sample(1:K,1,prob = p) }
z.mcmc=rbind(z.mcmc,z)
for(k in 1:K) {
nk[k]=sum(z==k)
m[k]=rnorm(1,mean(y[z==k]),sd=sqrt(s2[k]/nk[k]))
s2[k]=sum((y[z==k]-m[k])^ 2)/rchisq(1,nk[k]) }
m.mcmc=rbind(m.mcmc,m)
sigma2.mcmc=rbind(sigma2.mcmc,s2) }
return(list(z=z.mcmc,m=m.mcmc,sigma2=sigma2.mcmc)) }
```

7.5 Quelques résultats

Nous pouvons utiliser l'algorithme de Gibbs pour analyser les données présentées dans la figure 12. Nous effectuons 200 cycles de l'algorithme, précédés de 100 cycles de chauffe (burn-in). Il est pratique de représenter les lois a posteriori des classes individuelles ($p(z_i|y)$) en utilisant un diagramme en barres colorées. Dans ce diagramme, chaque segment vertical correspond à l'une des classes cachées, et

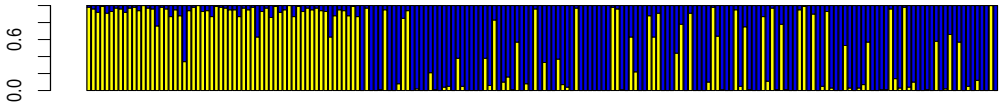


FIGURE 13 – *Diagramme en barres pour les lois a posteriori $p(z_i|y)$ correspondant à l'histogramme simulé au début de la section. Résultats pour $K = 2$ classes.*

la longueur du segment jaune ou bleu correspond à la probabilité a posteriori que le classement soit jaune ou bleu. En langage `R`, nous pouvons obtenir les probabilités a posteriori de chaque classe en comptant l'occurrence cette classe le long de la chaîne de Markov associée à l'échantillonnage de Gibbs. La représentation graphique est obtenue grâce à la commande `barplot`. Les résultats sont reportés dans les figures 13 et 14.

Pour $K = 2$, concernant la classe jaune, le taux de faux positifs est de l'ordre de 20% et le taux de fausse découverte est de l'ordre de 40% (résultats obtenus en considérant la valeur la plus probable a posteriori). Pour $K = 3$, nous observons que l'algorithme conserve la classe jaune. Il partitionne la seconde classe en attribuant des probabilités sensiblement égales aux classes bleue et rouge. Le résultat semble indiquer que $K = 2$ est un meilleur choix que $K = 3$.

Nous utilisons ensuite la loi prédictive de deux statistiques pour tester le modèle à trois classes : les coefficients d'aplatissement (kurtosis) et d'asymétrie (skewness). Dans le script suivant, l'objet `obj` est obtenu après application de la fonction `mcmc.mix` aux données de mélange.

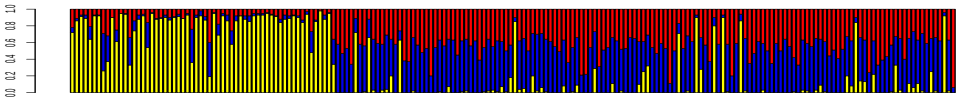


FIGURE 14 – *Diagramme en barres pour les lois a posteriori $p(z_i|y)$ correspondant à l'histogramme simulé au début de la section. Résultats pour $K = 3$ classes.*

```
post.stat = NULL

for (i in 1:100){

zz = obj$z[i+100,];mm = obj$m[i+100,];ss2 = obj$sigma2[i+100,]

yy = c(rnorm(sum(zz==1),mm[1],sqrt(ss2[1])),

rnorm(sum(zz==2),mm[2],sqrt(ss2[2])),

rnorm(sum(zz==3),mm[3],sqrt(ss2[3])))

post.stat=c(post.stat,kurtosis(yy))}
```

La figure 15 décrit un histogramme de la loi prédictive pour la statistique d'aplatissement (kurtosis) dans le modèle avec $K = 3$ classes. Nous voyons que cette statistique ne permet pas de rejeter fermement un modèle à trois classes pour nos données simulées. Cela montre que le choix de modèle peut être très difficile à réaliser. Dans ce cas, nous savons qu'il y a deux sources, mais un modèles à trois sources peut donner des résultats que l'on ne peut pas critiquer.

Enfin, notons que l'algorithme n'empêche pas le phénomène de permutation des étiquettes de classes (*label switching*). En effet, lorsqu'il y a plusieurs classes, la numérotation des classes est arbitraire, et on peut convenir que la classe bleue est la classe 1 ou 2 indifféremment. Par exemple, la figure 16 montre une exécution de l'algorithme sur les données de Fisher (longueur des sépales des

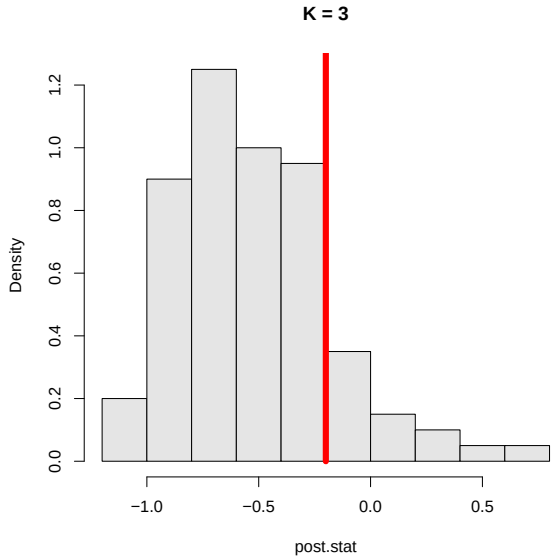


FIGURE 15 – *Loi prédictive pour la statistique d'aplatissement (kurtosis, $K = 3$).*

iris) pour $K = 2$. On voit clairement, que l'algorithme permute les étiquettes de classes en passant la bissectrice. Ce phénomène est clairement diagnostiqué en examinant la trajectoire de la chaîne de Markov, et les estimations correctes sont obtenues en début de run. Pour éviter que l'algorithme permute les classes, on peut simplement contraindre les moyennes $m_1 < m_2$ ou effectuer des runs courts.

7.6 Résumé

Résumer les points à retenir et donner quelques exemples illustrant les concepts principaux de la séance.

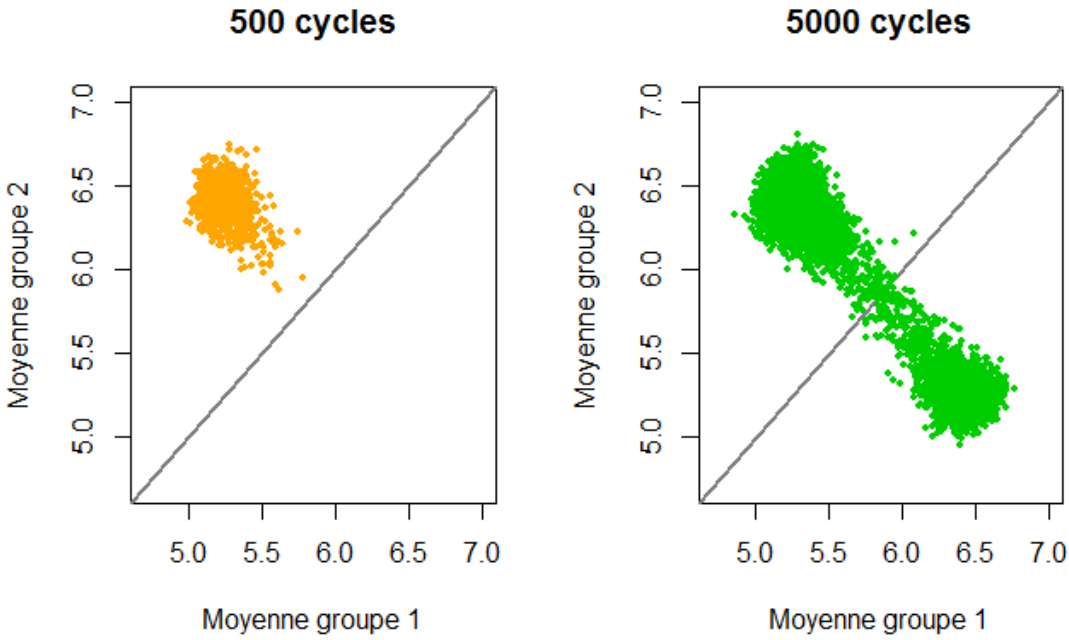


FIGURE 16 – *Run pour $K = 2$. Le phenomène de label switching peut etre analysé en examinant le run, et évité en répétant des runs courts. A gauche (run court), l'ordre des deux classes est conservé au cours du run. A droite (run l'ordre), l'ordre des deux classes n'est pas conservé au cours du run et il y a permutation des classes.*

7.7 Exercices

Pour les exercices suivants, il est nécessaire d'avoir un ordinateur sur lequel on aura installé préalablement le logiciel `R`. Ce logiciel libre est disponible sur le site `http://cran.r-project.org/`.

Exercice 1 : Mélange de 2 gaussiennes. On considère un échantillon $y = y_1, \ldots, y_n$ un échantillon constitué de n données réelles. On suppose les données indépendantes et provenant de deux sources gaussiennes de moyennes et de variances inconnues $\theta_1 = (m_1, \sigma_1^2)$ et $\theta_2 = (m_2, \sigma_2^2)$. On définit un vecteur $z = (z_1, \ldots, z_n)$, $z_i \in \{1, 2\}$, (correspondant aux classes non-observées) et un modèle de la manière suivante

$$y_i | z_i = k, \theta \sim \mathrm{N}(y_i | \theta_k), \quad k = 1, 2$$

$$p(z) \propto 1$$

$$p(\theta) \propto 1/\sigma_1^2 \sigma_2^2$$

1. Montrer que la loi a posteriori vérifie

$$p(z, \theta | y) \propto \prod_{k=1,2} \frac{1}{(\sigma_k^2)^{1+n_k/2}} \prod_{i \in I_k} \exp\left(-\frac{1}{2\sigma_k^2}(y_i - m_k)^2\right)$$

   où I_k est l'ensemble des indices tels que $z_i = k$ et n_k est le nombre d'éléments dans I_k (supposé non nul).

2. Pour tout $k = 1, 2$, montrer que la loi conditionnelle de z_i sachant (y, θ) vérifie

$$p(z_i = k | y, \theta) = \frac{\exp\left(-\frac{1}{2\sigma_k^2}(y_i - m_k)^2\right) / \sqrt{\sigma_k^2}}{\sum_{\ell=1}^{2} \exp(-\frac{1}{2\sigma_\ell^2}(y_i - m_\ell)^2) / \sqrt{\sigma_\ell^2}}$$

3. En vous appuyant sur les résultats d'une séance de travaux dirigés précédente,

justifier que, pour tout $k = 1, 2$,

$$m_k|y, z, \theta_{-m_k} \sim \mathrm{N}(m_k|\bar{y}_k, \sigma_k^2/n_k)$$

où $\bar{y}_k = \sum_{i \in I_k} y_i/n_k$.

4. De même justifier que, pour tout $k = 1, 2$,

$$\sigma_k^2|y, z, \theta_{-\sigma_k^2} \sim \mathrm{Inv}{-}\chi^2(\sigma_k^2|n_k, s_k^2)$$

où $s_k^2 = \sum_{i \in I_k} (y_i - m_k)^2/n_k$ est l'estimateur de la variance dans la classe k.

5. Ecrire un algorithme d'échantillonnage de Gibbs pour la loi a posteriori de ce modèle en langage `R`. On pourra utiliser la fonction `mcmc.mix` du script téléchargeable à l'adresse suivante

   `http://membres-timc.imag.fr/Olivier.Francois/script_R.R`

6. À quoi correspondent les lois marginales a posteriori, $p(z_i|y)$, $i = 1, \ldots, n$, pour l'analyse des donnés y ? Comment peut-on les visualiser sous `R` ? (aide : `barplot`)

7. Charger le jeu de données des iris de Fisher et extraire les longueurs des sépales. Simuler la loi a posteriori du modèle précédent et comparer les classifications obtenues pour chaque plante à l'appartenance de chaque plante à l'espèce *I. setosa*. On pourra utiliser la commande `barplot` pour visualiser les lois marginales a posteriori, $p(z_i|y)$, pour tout $i = 1, \ldots, n$.

Exercice 2. Modèle hiérarchique, effet aléatoire et *shrinkage*. Dans un modèle hiérarchique, les observations sont modélisées conditionnellement à

certains paramètres, et ces paramètres sont eux-mêmes décrits par des lois de probabilités dépendant d'autres paramètres, appelés *hyper-paramètres*.

L'objectif de cet exercice est d'estimer le risque pour un groupe de rats de décéder d'une tumeur cancéreuse à l'issue d'un traitement thérapeutique spécifique (les rats ont été sélectionnés pour développer la tumeur). Nous disposons pour cela de 71 groupes de rats, dont le nombre d'individus varie entre 14 et 52. Pour tenir compte de différences dans les échantillons considérés, nous souhaitons utiliser un modèle hiérarchique pour le risque. On note θ_i , i variant de 1 à 71, la probabilité qu'un rat du groupe i développe une tumeur, et on considère le vecteur $\theta = (\theta_i)$. Dans un groupe de rats de taille n_i, on modélise le nombre y_i de rats avec une tumeur par une loi binomiale de paramètres n_i et θ_i.

A priori, on modélise θ_i par une loi Beta de paramètres α et β. Les réels positifs α et β sont les hyperparamètres du modèle. La figure ci dessous représente la structure du modèle hiérarchique que l'on souhaite ajuster aux données.

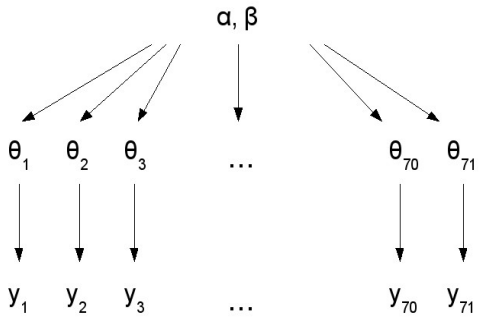


Figure 1: *Structure du modèle hiérarchique*

Il reste maintenant à spécifier les paramètres α et β. Afin d'effectuer un traitement bayésien du modèle, c'est-à-dire apprendre la loi a posteriori jointe de tous les paramètres du modèle, il est important de spécifier une loi a priori pour α et β. Comme nous n'avons pas d'information sur l'ensemble des paramètres, il est naturel de chercher à décrire une loi a priori non informative. Une loi uniforme ne convient pas, car elle conduit à une loi a posteriori non intégrable. Pour des raisons que l'on admettra ici, il est raisonnable de choisir une loi a priori qu'on voudra uniforme pour le couple

$$(\mu, \nu) = \left(\frac{\alpha}{\alpha + \beta}, \frac{1}{\sqrt{\alpha + \beta}} \right).$$

Afin d'apprendre la loi a posteriori de tous les paramètres du modèle, nous souhaitons programmer une méthode de Monte-Carlo par chaîne de Markov hybride. Il s'agit de mettre à jour les paramètres du modèle de manière séquentielle, pour un nombre fixé de pas. Voici une description synthétique de la méthode de

MCMC à programmer :

Etape 1. Partir de valeurs arbitraires α_0 , β_0 et θ_0 et fixer un nombre de pas pour l'algorithme, N. Pour $t = 1, \ldots, N$, répéter les opérations suivantes :

Etape 2. Mettre à jour les paramètres α_t , β_t à partir de leur valeur au pas $t-1$ à l'aide de l'algorithme de Métropolis-Hasting.

Etape 3. Mettre à jour le vecteur de paramètres θ_t à partir de la valeur précédente θ_{t-1} par échantillonage de Gibbs.

Etape 4. Après N balayages de l'ensemble des paramètres, choisir un nombre $b < N$ et conserver les valeurs de θ_t pour $t > b$ (b pour la période de *burn-in*).

1. Télécharger le fichier de données `rats.asc` à l'adresse suivante

   `http://www.stat.columbia.edu/~gelman/book/data/`

2. Calculer analytiquement le terme général de la densité $p(\alpha, \beta)$ par un changement de variables. Nous pouvons remarquer que cette densité est impropre (la densité n'est pas intégrable).

3. Donner l'expression de la vraisemblance $p(y|\theta)$ et de la loi conditionnelle $p(\theta|\alpha, \beta, y)$.

4. Calculer le rapport de Metropolis permettant la mise à jour du couple d'hyper-paramètres (α, β) (étape 2). Pour cette question, on pourra choisir librement la loi de transition instrumentale servant à proposer de nouvelles valeurs des paramètres.

5. Rappeler brièvement le principe de l'échantillonnage de Gibbs. Décrire

l'algorithme de mise à jour du paramètre θ (étape 3).

6. Programmer l'algorithme MCMC dans le langage `R`. Donner une estimation ponctuelle (moyenne conditionnelle, médiane et mode a posteriori) pour chacun des paramètres θ_i. Calculer les intervalles de crédibilité à 95% de ces paramètres .

7. Comparer ces estimations avec les estimations de maximum de vraisemblance $\hat{\theta}_i = y_i/n_i$. On pourra trier les valeurs y_i/n_i par ordre croissant, et afficher les moyennes a posteriori des lois marginales (les θ_i) en utilisant le même ordre. Commenter le phénomène de régularisation des risques observé dans cette courbe.

8 Séance 8 : Lois gaussiennes

8.1 Introduction

Dans cette séance, nous définissons les lois gaussiennes multivariées à partir des *composantes principales*. Nous établissons la forme générale de la densité de cette famille de lois, puis, dans le cas bi-varié, nous calculons l'espérance et la variance conditionnelle d'une coordonnée sachant l'autre coordonnée. Nous montrons que le calcul de l'espérance conditionnelle est similaire au calcul de la *régression linéaire* et justifions que le carré du coefficient de corrélation – ou coefficient de détermination – représente exactement le pourcentage de variance expliquée par la variable observée. Nous discutons ensuite la simulation de variables gaussiennes multivariées ainsi que l'échantillonnage de Gibbs dans ce contexte.

8.2 Définitions

On dit qu'un vecteur aléatoire $y \in \mathbb{R}^d$ d'espérance nulle est **un vecteur gaussien** s'il existe une matrice de rotation U, vérifiant $UU^T = \mathrm{Id}$, telle que les coordonnées du vecteur $z = Uy$ sont indépendantes et de loi normale : $z_j \sim \mathrm{N}(0, \lambda_j)$ pour tout $j = 1, \ldots, d$.

La matrice de covariance d'un vecteur aléatoire, y, est la matrice C dont le terme général c_{ij} correspond à la covariance du couple (y_i, y_j). Elle peut s'écrire sous la forme suivante

$$C = \mathrm{E}[(y - \mathrm{E}[y])(y - \mathrm{E}[y])^T] \,.$$

Pour une transformation linéaire d'un vecteur aléatoire y de matrice de cova-

riance C, notée $x = Ay$, la matrice de covariance de x, notée $\tilde{C}$, est égale à

$$\tilde{C} = ACA^T \, .$$

En appliquant la formule ci-dessus à $y = U^T z$, nous obtenons en toute généralité que la matrice covariance de y est égale à

$$C = U^T \Lambda U \, ,$$

où Λ est une matrice diagonale dont le j^{e} terme diagonal est égal à λ_j. Nous voyons ainsi que le déterminant de C, égal à celui de Λ, est strictement positif. Les variances λ_j correspondent aux valeurs propres de C.

Un vecteur aléatoire gaussien $y \in \mathbb{R}^d$ d'espérance m est défini comme la somme d'un vecteur gaussien d'espérance nulle et du vecteur m. Dans ce cas, cela revient à prendre $z = U(y - m)$ dans la définition précédente.

Nous pouvons remarquer que toute combinaison linéaire des coordonnées d'un vecteur gaussien suit une loi normale. Plus précisément, soit z_a une combinaison linéaire des coordonnées y_j, telle que a représente les coefficients de cette combinaison

$$z_a = a^T y = \sum_{j=1}^{d} a_j y_j \, .$$

Alors, z_a suit la loi normale de moyenne $m_a = a^T m$ et de variance $\sigma_a^2 = a^T C a$. En effet, nous avons

$$z_a = (Ua)^T z = b^T z = \sum_{j=1}^{d} b_j z_j \, .$$

où $b = Ua$ Le résultat découle alors du fait que les variables $b_j z_j$ sont des variables de loi normale indépendantes entre elles. Nous savons, en effet, que la somme de variables aléatoires indépendantes de loi normale est encore une

variable aléatoire de loi normale (le vérifier à l'aide de l'expression des densités). Cela implique qu'une transformation matricielle linéaire, $x = Ay$, du vecteur gaussien y est un vecteur gaussien. Lorsque le déterminant de la matrice de covariance est non-nul, la dimension de x est égale à d. Dans le cas contraire elle est égale au rang de A.

Une combinaison particulière des coordonnées d'un vecteur gaussien consiste à considérer la première variable, y_1, extraite du vecteur y. Dans ce cas, le vecteur a est le vecteur unitaire correspondant à la première coordonnée, $a = (1, 0, \ldots, 0)$. Le résultat précédent nous indique que y_1 est de loi normale. Généralement, **les lois marginales d'un vecteur gaussien sont toujours gaussiennes**.

8.3 Densité de probabilité d'un vecteur gaussien.

Dans ce paragraphe, nous décrivons la forme générale de la loi de probabilité d'un vecteur gaussien de moyenne m et de matrice de covariance C, en supposant que le déterminant de C est strictement positif.

Expression de la densité d'un vecteur gaussien. Soit $y \in \mathbb{R}^d$ un vecteur aléatoire ayant d coordonnées. On dit que la loi de y est une loi gaussienne de moyenne m et de matrice de covariance C si

$$p(y|m, C) = \frac{1}{(2\pi)^{d/2}} \frac{1}{(\det C)^{1/2}} \exp(-\frac{1}{2}(y-m)^T C^{-1}(y-m)).$$

On note $\Delta^2 = (y-m)^T C^{-1}(y-m)$ la distance de Mahanalobis entre m et y. La densité de la loi normale de moyenne m et de matrice de covariance C est notée $\mathrm{N}(y|m, C)$.

Il est important de bien comprendre la forme de cette loi de probabilité.

Pour cela, notons que C est une matrice symétrique dont les valeurs propres sont positives. D'après le théorème spectral et la définition donnée au début de cette séance, nous pouvons considérer les vecteurs propres orthonormés, (u_j), associés aux valeurs propres, (λ_j), de C

$$Cu_j = \lambda_j u_j \, .$$

Ces vecteurs propres s'appellent les **composantes principales** de la matrice de covariance. Nous pouvons écrire les représentations spectrales associées à la matrice de covariance

$$C = \sum_{j=1}^{d} \lambda_j u_j u_j^T \, ,$$

ainsi qu'à son inverse

$$C^{-1} = \sum_{j=1}^{d} \frac{1}{\lambda_j} u_j u_j^T \, .$$

Nous voyons que la distance de Mahanalobis entre m et y s'écrit de la manière suivante

$$\Delta^2 = \sum_{j=1}^{d} \frac{1}{\lambda_j} z_j^T z_j = \sum_{j=1}^{d} \frac{1}{\lambda_j} z_j^2$$

où $z_j = u_j^T (y - m)$. En écriture matricielle, nous venons d'effectuer le changement de variable linéaire suivant

$$z = U(y - m) = \varphi^{-1}(y) \, .$$

La valeur absolue du jacobien de ce changement de variable linéaire est égale au déterminant de U, c'est à dire, égale à 1. En rappelant que le déterminant d'une matrice carrée correspond au produit des valeurs propres de cette matrice, nous obtenons par la formule de changement de variable l'expression de la densité de

la variable z

$$p(z) = \frac{1}{(2\pi)^{d/2}} \frac{1}{(\lambda_1 \dots \lambda_d)^{1/2}} \exp(-\sum_{j=1}^{d} \frac{1}{\lambda_j} z_j^2).$$

Cette densité peut se factoriser de la manière suivante

$$p(z) = \prod_{j=1}^{d} \mathrm{N}(z_j|0, \lambda_j)\,.$$

Les variables z_j sont donc indépendantes et de loi normale $\mathrm{N}(0, \lambda_j)$. Nous vérifions ainsi la définition d'un vecteur gaussien, et au passage, que la matrice covariance de ce vecteur est bien égale à C.

Une conséquence importante de la définition d'un vecteur gaussien est que pour un tel vecteur, les **coordonnées, $y_1, \dots, y_d$, sont indépendantes si et seulement si la matrice covariance du vecteur y est diagonale**.

Exemple. Considérons le vecteur $y = (y_1, y_2)$ de moyenne nulle et de matrice de covariance C égale à

$$C = \frac{1}{2} \begin{pmatrix} 3 & -1 \\ -1 & 3 \end{pmatrix}$$

Le déterminant de C est égal à 2 et nous avons

$$4\Delta^2 = 3y_1^2 + 2y_1y_2 + 3y_2^2\,.$$

Nous pouvons réécrire cette expression sous la forme d'une somme de carrés

$$\Delta^2 = z_1^2 + \frac{z_2^2}{2}$$

où

$$\begin{cases} z_1 &= (y_1 + y_2)/\sqrt{2} \\ z_2 &= (y_2 - y_1)/\sqrt{2} \end{cases}$$

Ainsi, nous venons de mettre en évidence la rotation permettant le changement de base $z = Uy$ (angle $\pi/4$). D'après l'expression de Δ^2, nous pouvons vérifier que z_1 suit la loi N(0,1) et que z_2 suit la loi N(0,2). De plus les deux variables sont indépendantes. Nous pouvons de plus vérifier que $C = U^T \Lambda U$ où Λ est la matrice diagonale dont les valeurs propres sont égales à $(1, 2)$.

Bien que cette méthode n'est pas la plus aisée, il est enfin possible grâce à la décomposition spectrale que nous venons d'effectuer, d'écrire un algorithme de simulation du couple (y_1, y_2). En langage `R`, nous pouvons effectuer 1000 tirages de cette loi bi-variée de la manière suivante

```
z1 = rnorm(1000) ; z2 = rnorm(1000, sd = sqrt(2))
y1 = (z1 - z2)/sqrt(2) ; y2 = (z1 + z2)/sqrt(2)
```

8.4 Lois conditionnelles d'un vecteur gaussien – Régression linéaire

Dans cette section, nous calculons l'espérance et la variance conditionnelle au sein d'un couple de variables gaussiennes, et nous établissons une connexion avec la régression linéaire. Le terme **régression** est un synonyme du terme **espérance conditionnelle**. Cette grandeur représente la meilleure prédiction d'une variable au sens des moindres carrés sachant une autre variable. Dans cette section, nous établissons deux résultats importants : **dans le cas gaussien, la régression est linéaire, et le coefficient de corrélation représente la fraction de la variance d'une variable expliquée par l'autre variable**. Comme il est simple de le vérifier grâce à l'expression des densités, nous admettons que les **lois conditionnelles d'un vecteur gaussien sont elles mêmes gaussiennes**.

Pour bien comprendre le conditionnement gaussien, considérons un couple (y, θ) où θ est la variable cachée que nous cherchons à prédire et y est une observation bruitée résultant d'un plan d'expérience gaussien. Nous supposons que le couple (y, θ) est gaussien, d'espérance m et de matrice de covariance C. Afin de calculer l'espérance conditionnelle $\mathrm{E}[\theta|y]$, nous cherchons une combinaison linéaire de θ et y indépendante de θ. Pour cela, considérons les variables suivantes

$$\left\{ \begin{array}{ll} z_1 & = y \\ z_2 & = y + a\theta \end{array} \right.$$

où a est un scalaire que choisissons de sorte que

$$\mathrm{Cov}(y, y + a\theta) = 0\,.$$

Nous obtenons

$$a = -\left[\frac{\mathrm{Cov}(y,\theta)}{\mathrm{Var}(y)}\right]^{-1}\,.$$

Pour cette valeur de a, z_1 et z_2 sont indépendantes. Nous pouvons donc calculer facilement $\mathrm{E}[z_2|y]$. Par indépendance, cette espérance ne dépend pas de y. Elle est égale à

$$\mathrm{E}[z_2|y] = \mathrm{E}[z_2] = \mathrm{E}[y] + a\mathrm{E}[\theta].$$

D'autre part, nous avons

$$\mathrm{E}[z_2|y] = \mathrm{E}[y + a\theta|y] = y + a\mathrm{E}[\theta|y]\,.$$

En comparant les deux expressions de $\mathrm{E}[z_2|y]$, nous obtenons une formule générale donnant $\mathrm{E}[\theta|y]$

$$\mathrm{E}[\theta|y] = \mathrm{E}[\theta] + \frac{\mathrm{Cov}(\theta, y)}{\mathrm{Var}(y)}(y - \mathrm{E}[y])\,.$$

Cette formule correspond à la formule classique permettant de calculer la **régression linéaire**. On pourrait retrouver cette formule directement par la résolution d'un

problème de moindres carrés. Nous nous passerons de ce calcul, car nous avons justifié le résultat lors d'une séance précédente.

Nous pouvons maintenant calculer la variance conditionnelle de θ sachant y. En utilisant les résultats précédents, nous avons

$$\mathrm{Var}(z_2|y) = \mathrm{Var}(y + a\theta)\,.$$

En développant cette expression nous obtenons

$$\mathrm{Var}(z_2|y) = \left(\frac{\mathrm{Var}^2(y)\mathrm{Var}(\theta)}{\mathrm{Cov}(y,\theta)^2}\right) - \mathrm{Var}(y) = a^2\mathrm{Var}(\theta) - \mathrm{Var}(y)\,,$$

alors qu'en effectuant le calcul direct, nous avons

$$\mathrm{Var}(z_2|y) = \mathrm{Var}(a\theta|y) = a^2\mathrm{Var}(\theta|y)\,.$$

Nous pouvons comparer les deux expressions de $\mathrm{Var}(z_2|y)$, et cela conduit au résultat suivant

$$\mathrm{Var}(\theta|y) = \mathrm{Var}(\theta) - \frac{\mathrm{Cov}(y,\theta)^2}{\mathrm{Var}(y)} = \mathrm{Var}(\theta)(1-\rho^2)\,,$$

où

$$\rho = \frac{\mathrm{Cov}(y,\theta)}{\sigma(y)\sigma(\theta)}.$$

En d'autres termes, **la part de variance de θ expliquée par l'observation y est égale au carré du coefficient de corrélation**

$$\frac{\mathrm{Var}(\theta) - \mathrm{Var}(\theta|y)}{\mathrm{Var}(\theta)} = \rho^2\,.$$

A nouveau, il s'agit d'un résultat classique pour la régression linéaire, justifiant l'utilisation du carré du coefficient de corrélation dans ce contexte.

8.5 Simulation d'un vecteur gaussien

Nous souhaitons simuler un couple gaussien de moyenne nulle et de matrice de covariance C. Les techniques évoquées ici se généralisent pour des vecteurs de dimension d. Pour simplifier, nous nous limitons à décrire le cas $d = 2$.

D'après la définition d'un vecteur gaussien, nous avons la possibilité d'effectuer la simulation de ce vecteur en utilisant les composantes principales de la matrice de covariance. Une méthode élémentaire pour simuler un vecteur gaussien consiste à calculer les vecteurs propres orthonormés de la matrice de covariance C (matrice U), de simuler des variables z_j de loi normale $\mathrm{N}(0, \lambda_j)$ où λ_j est la j^{e} valeur propre de C, puis enfin de transformer z en $y = U^T z$. Dans cette section, nous présentons deux méthodes différentes. La première méthode, la plus utile, s'appuie sur la **décomposition de Cholesky de la matrice de covariance**. La seconde méthode s'appuie sur l'échantillonnage de Gibbs, donné ici comme illustration pédagogique du cours de MPA.

Décomposition de Cholesky. La matrice C est une matrice symétrique semi-definie positive. D'après Cholesky, il existe une matrice L, triangulaire inférieure telle que

$$C = LL^T .$$

La matrice L peut être obtenue grâce à l'algorithme de **factorisation de Cholesky**. L'algorithme de simulation consiste alors à tirer les variables z_j selon la loi $\mathrm{N}(0, 1)$ et de manière indépendante, puis d'effectuer la transformation

$$y = Lz .$$

Nous voyons que cette méthode est très similaire à l'algorithme classique de

simulation d'un couple de variable aléatoire pour les lois gaussiennes.

En dimension $d = 2$, la loi de la variable y_1 est une loi normale

$$y_1 \sim \mathrm{N}(0, \sigma_1^2)\,,$$

où σ_1^2 est la variance de y_1. En effet, dans un vecteur gaussien, les lois marginales sont normales. D'après la section précédente, la loi conditionnelle de y_2 sachant y_1 est aussi normale. En posant $c = \mathrm{Cov}(y_1, y_2)$, nous avons

$$y_2|y_1 \sim \mathrm{N}(cy_1/\sigma_1^2, \sigma_2^2(1-\rho^2)).$$

Pour simuler le couple (y_1, y_2), nous définissons z_1, z_2, deux variables aléatoires indépendantes de loi $\mathrm{N}(0,1)$, par exemple à l'aide du générateur aléatoire `rnorm()` de `R`. Nous pouvons écrire

$$\left\{ \begin{array}{ll} y_1 & = \sigma_1 z_1 \\ y_2 & = cz_1/\sigma_1 + \sigma_2\sqrt{1-\rho^2}z_2 \end{array} \right.$$

Cela revient à considérer la transformation linéaire inverse triangulaire inférieure suivante

$$L = \left(\begin{array}{cc} \sigma_1 & 0 \\ c/\sigma_1 & \sigma_2\sqrt{1-\rho^2} \end{array} \right).$$

Nous pouvons facilement vérifier que $C = LL^T$. L'implantation de l'algorithme classique et exact de la simulation d'un couple de variable aléatoire est réalisable par changement de variable linéaire. Il est similaire à l'algorithme de factorisation de Cholesky. Un algorithme de simulation exact de cette loi gaussienne permettant 1000 tirages simultanés peut s'écrire en langage `R` de la manière suivante

```
theta1 = rnorm(1000, sd = sigma1)
```

```
theta2 = sigma2 * rnorm(1000, rho*theta1, sd = sqrt(1 - rho^ 2))
```

Echantillonnage de Gibbs. L'algorithme d'échantillonnage de Gibbs pour un couple de loi gaussienne consiste à alterner la simulation des lois conditionnelles,

$$y_2|y_1 \sim \mathrm{N}(\frac{\mathrm{Cov}(y_1,y_2)}{\sigma_1^2}y_1, \sigma_2^2(1-\rho^2))$$

et

$$y_1|y_2 \sim \mathrm{N}(\frac{\mathrm{Cov}(y_1,y_2)}{\sigma_2^2}y_2, \sigma_1^2(1-\rho^2)).$$

Afin de comparer l'échantillonneur de Gibbs à l'algorithme exact décrit dans le paragraphe précédent, nous souhaitons simuler une loi gaussienne de moyenne $(0,0)$ et de matrice de covariance

$$C = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

Dans cette matrice de covariance, le coefficient ρ est le coefficient de corrélation des deux variables simulées, θ_1 et θ_2. L'algorithme de simulation exact de cette loi gaussienne peut s'écrire en langage `R` de la manière suivante

```
theta1 = rnorm(1000)
theta2 = rnorm(1000, rho*theta1, sd = sqrt(1 - rho^ 2))
```

Dans ce cas l'algorithme d'échantillonnage de Gibbs consiste à itérer le cycle suivant

```
theta1 = rnorm(1, rho*theta2, sd = sqrt(1 - rho^ 2))
theta2 = rnorm(1, rho*theta1, sd = sqrt(1 - rho^ 2))
```

Pour la valeur $\rho = 70\%$, les résultats des deux algorithmes sont quasiment identiques (figure 16). Nous verrons, en expérimentant cet algorithme lors

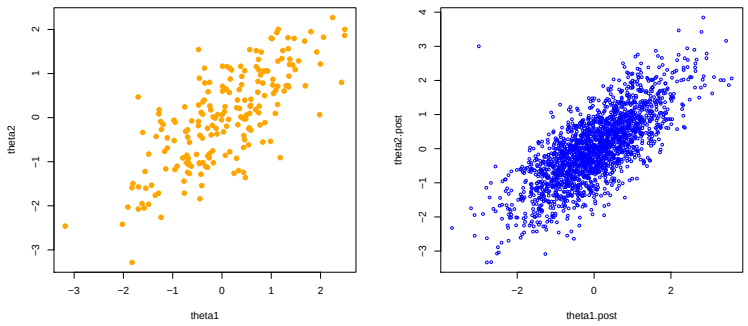


FIGURE 17 – *Simulation d'une loi gaussienne par l'algorithme exact et par échantillonnage de Gibbs.*

de la séance de travaux dirigés que la vitesse de convergence de l'algorithme d'échantillonnage de Gibbs dépend de la corrélation entre les composantes du paramètre à simuler. Plus la valeur de ρ est grande, plus il devient difficile d'obtenir un échantillon de la loi cible tels que les paramètres simulés peuvent être considérés indépendants les uns des autres.

8.6 Résumé

Résumer les points à retenir et donner quelques exemples illustrant les concepts principaux de la séance.

8.7 Exercices

Pour les exercices suivants, il est nécessaire d'avoir un ordinateur sur lequel on aura installé préalablement le logiciel `R`. Ce logiciel libre est disponible sur le site `http://cran.r-project.org/`.

Exercice 1. Résumé On considère un couple gaussien (y_1, y_2) de moyenne $m = (1, 0)$ et de matrice de covariance

$$C = \begin{pmatrix} 3 & -1 \\ -1 & 2 \end{pmatrix}.$$

1. Décrire la densité du couple (y_1, y_2) et ses lois marginales.
2. Soit c une constante non nulle. Décrire la loi du vecteur (z_1, z_2) ci-dessous

$$\begin{cases} z_1 = y_1 \\ z_2 = y_1 + cy_2 \end{cases}$$

3. Démontrer que les variables z_1 et z_2 sont indépendantes si et seulement si $c = 3$.
4. Soit $y_1 \in \mathbb{R}$ et $c = 3$. En utilisant l'indépendance de z_2 et y_1 (ou z_1), montrer que

$$\mathrm{E}[z_2] = y_1 + 3\,\mathrm{E}[y_2|y_1]\,.$$

   En déduire l'espérance conditionnelle de y_2 sachant y_1.

5. Par un argument similaire à la question précédente, montrer que

$$\mathrm{Var}(z_2) = c^2 \mathrm{var}(y_2|y_1)\,.$$

6. On note

$$R^2 = \frac{\mathrm{Var}(y_2) - \mathrm{Var}(y_2|y_1)}{\mathrm{Var}(y_2)}$$

la part de variance de y_2 expliquée par y_1. Vérifier que R^2 est égal au carré du coefficient de corrélation (aussi applelé le *coefficient de détermination*).

7. Décrire la loi conditionnelle de y_2 sachant y_1.
8. Proposer un algorithme de simulation du couple (y_1, y_2) à partir d'un simulateur `rnorm` de la loi $\mathrm{N}(0,1)$, et effectuer des simulations pour vérifier les résultats précédents à l'aide du logiciel `R`, et en utilisant la commande `lm`.

Exercice 2. Soit (x, y) un couple de variables aléatoires de densité

$$p(x,y) \propto \exp(-\frac{4x^2 - 2xy + y^2}{6}).$$

1. Montrer que le couple (x, y) est gaussien, de moyenne $m = (0, 0)$ et de matrice de covariance
$$C = \begin{pmatrix} 1 & 1 \\ 1 & 4 \end{pmatrix}.$$
En déduire la valeur de la constante de proportionnalité de la loi $p(x, y)$.
2. Déterminer la loi de x, de y. Quelle est la valeur de $cov(x, y)$? Les variables x et y sont-elles indépendantes ?
3. On pose
$$\begin{cases} z_1 &= x - y \\ z_2 &= x \end{cases}$$
Montrer que le couple (z_1, z_2) est gaussien et caractériser sa loi. Les variables z_1 et z_2 sont-elles indépendantes ?
4. Ecrire un algorithme de simulation pour le couple (x, y). (On prendra soin de démontrer que le couple en sortie de cet algorithme est bien de densité $p(x, y)$.)

5. Déterminer la loi conditionnelle de y sachant x .
6. À la suite de la question précédente, écrire un nouvel algorithme de simulation du couple (x, y).
7. Ecrire un algorithme d'échantillonnage de Gibbs pour simuler la loi de ce vecteur.
8. programmer les algorithmes précédents en langage `R` et comparer les résultats des simulations.

Exercice 3. On considère la matrice symétrique suivante

$$\Lambda = \begin{pmatrix} 3 & c \\ c & 4 \end{pmatrix},$$

où c est un scalaire quelconque.

1. Pour quelles valeurs de la constante c, la matrice Λ est-elle une matrice de covariance ? *Lorsque la condition précédente est satisfaite, on définit un couple gaussien* (θ_1, θ_2) *de moyenne nulle et de matrice de covariance* Λ. *On pose* $\rho = c/2\sqrt{3}$.
2. Écrire la densité du couple (θ_1, θ_2). Caractériser les lois marginales et les lois conditionnelles en fonction de ρ à partir des formules du cours.
3. Proposer un algorithme de simulation exacte du couple (θ_1, θ_2) à partir de la commande `rnorm` du langage `R`.
4. Rappeler le principe de l'algorithme d'échantillonnage de Gibbs pour le couple (θ_1, θ_2) et écrire cet algorithme en langage `R`.
5. On initialise l'algorithme d'échantillonnage de Gibbs avec la valeur $\theta_2^0 = 0$

et le tirage de θ_1^0. Montrer qu'au cycle t de l'algorithme, on peut écrire

$$\begin{aligned} \theta_1^{t-1} &= \frac{\sqrt{3}}{2}\rho\,\theta_2^{t-1} + \epsilon_2^{t-1} \\ \theta_2^{t} &= \frac{2}{\sqrt{3}}\rho\,\theta_1^{t-1} + \epsilon_1^{t} \end{aligned}$$

où les variables ϵ_i^t sont des variables gaussiennes indépendantes de moyenne et de variance à préciser.

6. Montrer, à l'aide de la question précédente, que l'on peut écrire pour $t \geq 1$

$$\theta_2^t = \epsilon^t + \rho^2\epsilon^{t-1} + \rho^4\epsilon^{t-2} + \cdots + (\rho^2)^{t-1}\epsilon^1$$

où les variables ϵ^t sont indépendantes de loi normale de moyenne 0 et de variance $4(1-\rho^4)$.

7. En déduire que θ_2^t est une variable aléatoire gaussienne de moyenne nulle et de variance égale à

$$\mathrm{Var}(\theta_2^t) = 4(1-\rho^{4t})$$

Que peut-on dire du comportement de l'algorithme lorsque ρ est proche de la valeur 1 ?

Exercice 4. On considère un vecteur gaussien x en dimension 3 de moyenne nulle et de matrice de covariance

$$C = \begin{pmatrix} \sigma_1^2 & 0 & c_{13} \\ 0 & \sigma_2^2 & c_{23} \\ c_{13} & c_{23} & \sigma_3^2 \end{pmatrix}$$

où $\det C > 0$. Les coordonnées de x sont notées x_1, x_2 et x_3. On définit le vecteur

y de la manière suivante

$$\begin{cases} y_1 = x_1 \\ y_2 = x_2 \\ y_3 = x_3 - c_{13}x_1/\sigma_1^2 - c_{23}x_2/\sigma_2^2 \end{cases}$$

1. Montrer que le vecteur y est gaussien et de moyenne nulle.
2. Calculer $\mathrm{Cov}(y_1, y_3)$ et $\mathrm{Cov}(y_2, y_3)$. Montrer que les coordonnées y_1, y_2 et y_3 sont indépendantes.
3. Montrer que

   $$\mathrm{Var}(x_3) = \mathrm{Var}(y_3) + c_{13}^2/\sigma_1^2 + c_{23}^2/\sigma_2^2$$

   En déduire la loi de y.
4. On dispose d'un générateur aléatoire `rnorm()` retournant des variables indépendantes de loi $\mathrm{N}(0,1)$. Écrire un algorithme de simulation exact d'un vecteur gaussien de moyenne $m = (0,1,0)$ et de matrice de covariance

   $$K = \begin{pmatrix} 1 & -1 & 0 \\ -1 & 9 & 2 \\ 0 & 2 & 1 \end{pmatrix}$$

5. Ecrire un algorithme d'échantillonnage de Gibbs pour simuler la loi de ce vecteur.
6. Programmer les algorithmes précédents en langage `R` et comparer les résultats des simulations.

Énoncés des Travaux Pratiques (TP) de MPA

Les TPs de MPA peuvent être effectués en binôme d'un même groupe, et donnent lieu à un compte-rendu noté à déposer sur l'application TEIDE. Toute journée de retard dans la réception du compte-rendu sur l'application entraîne un décompte de 4 points sur la note du compte-rendu. Le compte-rendu ne dépassera pas 6 pages incluant les formules, figures, tableaux et les principales commandes du langage `R` utilisées. Les formats doc ou PDF sont acceptés. Le barème de notation prend en compte l'exactitude, la qualité de la présentation, mais aussi les commentaires et la discussion des résultats. Les légendes des figures et leurs axes devront être explicites et lisibles sans référence systématique au texte. La description des algorithmes devra garantir que l'on puisse les programmer en `R` sans ambiguité. La qualité de la rédaction, en français ou en anglais, sera aussi notée.

TP 1 : Simulation et programmation en `R` – A remettre le lundi de la semaine 2.

On considère la loi de probabilité définie sur $\mathbb{R}^2$ par

$$p(y,\theta) = \frac{1}{16\pi}\exp\left(-\frac{1}{32}(8y^2 - 4y\theta + \theta^2)\right) .$$

On rappelle que la loi normale $\mathrm{N}(m,\sigma^2)$, définie sur $\mathbb{R}$, admet pour densité la fonction suivante

$$p(y) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2\sigma^2}(y-m)^2\right) .$$

1. Sans effectuer de calcul intégral, identifier la loi générative $p(y|\theta)$, puis la loi a priori $\mathrm{p}(\theta)$. Indication : utiliser uniquement des relations de pro-

portionnalité, puis déduire les constantes de normalisation de la formule générale de la loi normale.

2. Identifier la loi a posteriori $p(\theta|y)$, puis la loi marginale $\mathrm{p}(y)$.
3. Ecrire un programme en langage `R` permettant de simuler des tirages selon la loi de densité $p(y, \theta)$.
4. Calculer l'espérance de la loi conditionnelle $p(\theta|y)$.
5. Effectuer 1000 tirages de loi $p(y, \theta)$. Représenter graphiquement les tirages en couleur grise, et superposer en couleur bleue la courbe qui, à y, associe l'espérance $\mathrm{E}[\theta|y]$.
6. Superposer en couleur orange la droite de régression linéaire de la variable θ par la variable y (commande `lm`).
7. Effectuer 100000 tirages de la loi $p(y, \theta)$. Programmer en langage `R` un algorithme rejetant les valeurs de la variable θ telles que la variable y se trouve en dehors de l'intervalle $1.99, 2.01$.
8. Tracer l'histogramme des valeurs retenues par l'algorithme. Interpréter l'histogramme ainsi obtenu.
9. Justifier le fait que la loi de densité $\mathrm{p}(y, \theta)$ est caractérisée par la donnée des lois conditionnelles $p(y|\theta)$ et $p(\theta|y)$.
10. On initialise la variable θ à une valeur θ^0 arbitraire (par exemple, $\theta^0 = 0$).
    Pour tout $t \geq 1$, on considère l'itération du cycle suivant
    - Simuler y^t selon la loi $p(y|\theta^{t-1})$
    - Simuler θ^t selon la loi $p(\theta|y^t)$.

    Cet algorithme est appelé algorithme d'*échantillonnage de Gibbs*. Ecrire un programme en langage `R` permettant de simuler l'algorithme d'*échantillonnage*

de Gibbs. Effectuer 1000 tirages du couple (y, θ) en prenant 1000 valeurs successives après avoir éliminé les premières valeurs afin de perdre la condition initiale. Commenter le résultat. Est-il sensible au nombre de valeurs éliminées ?

TP 2 : Détection d'un changement ponctuel dans une suite binaire – A remettre avant le lundi de la semaine 6. On observe une suite de longueur n, notée y, composée de signaux binaires correspondant à l'émission d'une source de fréquence inconnue θ_1, susceptible de changer en un instant (point) inconnu de θ_1 à θ_2. Les émissions sont indépendantes les unes des autres. Par exemple, cette suite peut se présenter de la manière suivante :

$$y = 01100 \ldots 00110 \| 1110011 \ldots 1100111$$

Dans cette représentation, le symbole $\|$ marque le changement ayant lieu au point c, un indice entier compris entre 1 et n. Par convention, $c = 1$ correspond à la situation où il n'y a pas de changement et les n tirages sont de fréquence égale à θ_2. Nous notons

$$\theta = (c, \theta_1, \theta_2).$$

Pour $c = 2, \ldots, n$, nous avons

$$p(y_i|\theta) = \theta_1^{y_i}(1 - \theta_1)^{1-y_i} \quad i = 1, \ldots, c-1$$

et

$$p(y_i|\theta) = \theta_2^{y_i}(1 - \theta_2)^{1-y_i} \quad i = c, \ldots, n.$$

Nous supposons que

$$p(\theta) = p(c)p(\theta_1)p(\theta_2) = \frac{1}{n}, \quad 0 \leq \theta_1, \theta_2 \leq 1 \; et \; c = 1, \ldots, n.$$

L'objectif de cet exercice est de proposer (et de tester) un algorithme d'apprentissage du point de changement c, permettant le calcul de la loi a posteriori $p(c|y)$.

1. Soit $c = 1$. Donner l'expression de la loi générative $p(y|c = 1, \theta_1, \theta_2)$.
2. Même question pour $c > 1$.
3. On suppose que θ_1 et θ_2 sont connues. Déduire des questions précédentes que
$$\forall c = 2, \ldots, n, \quad \frac{p(c|y)}{p(c = 1|y)} = \prod_{j=1}^{c-1} \frac{\theta_1^{y_i}(1 - \theta_1)^{1-y_i}}{\theta_2^{y_i}(1 - \theta_2)^{1-y_i}}.$$
Vérifier que l'on peut calculer le rapport $p(c|y)/p(c = 1|y)$ pour tout c en effectuant de l'ordre de $n(n - 1)/2$ multiplications.
4. Supposant θ_1 et θ_2 connues, proposer un algorithme permettant de calculer $p(c|y)$ pour tout $c = 1, \ldots, n$ avec une complexité en $O(n)$. Indication : mettre à jour le rapport définit dans la question précédente à l'aide d'une récurrence linéaire.
5. On suppose désormais que θ_1 et θ_2 sont inconnues. A partir de valeurs initiales arbitraires, proposer un algorithme itératif, de type échantillonnage de Gibbs, permettant de calculer $p(c|y)$ pour tout $c = 1, \ldots, n$ en combinant la simulation de la loi $p(c|y, \theta_1, \theta_2)$ et la simulation de valeurs des paramètres θ_1 et θ_2.
6. Tester la convergence de votre algorithme en examinant la sensibilité aux conditions initiales choisies arbitrairement.

7. Analyser les jeux de données envoyés en pièce jointe par l'enseignant. Décrire l'incertitude sur le(s) point(s) de changement pour ces jeux de données (localisation des points des changements et intervalles contenant chacun des points avec une probabilité supérieure à 50%).

TP 3 : Modèle de Poisson et mélanges – A remettre avant le lundi de la semaine 9. On souhaite estimer le nombre moyen d'occurrences d'un phénomène donné, correspondant par exemple au nombre de clics journaliers sur un type de produit spécifique dans un site de vente en ligne. Pour cela, on dispose de n observations entières positives ou nulles, notées $y_1, \ldots, y_n$. Au moins l'une de ces observations est non nulle.

Première partie.

1. On suppose les observations indépendantes et de loi de Poisson de paramètre $\theta > 0$. Déterminer la loi générative des données, y, dans ce modèle.
2. On suppose que la loi a priori est non informative

$$p(\theta) \propto \frac{1}{\theta}, \quad \theta > 0\,.$$

   Déterminer la loi a posteriori du paramètre θ. Quelle est l'espérance de la loi a posteriori ?
3. Rappeler le principe de l'algorithme de Metropolis-Hasting. Ecrire dans un programme en `R` une version de cet algorithme pour simuler la loi a posteriori du paramètre θ. On pourra, par exemple, choisir une loi instrumentale exponentielle.

4. Fournir une estimation ponctuelle du paramètre θ (moyenne et médiane a posteriori). Donner un intervalle de crédibilité à 95% pour ce paramètre.

5. Représenter graphiquement l'histogramme de la loi a posteriori du paramètre θ obtenu suivant l'algorithme de Metropolis-Hasting. Superposer la loi obtenue à la question 2.

6. Déterminer la loi prédictive *a posteriori* d'une nouvelle donnée $\tilde{y}$. Ecrire un algorithme de simulation de cette loi et comparer l'histogramme des résultats simulés aux données. Quelles critiques pouvez-vous faire du modèle proposé pour les données et le paramètre θ.

Seconde partie. Soit $y = y_1, \ldots, y_n$ un échantillon constitué de n données de comptage, entiers positifs ou nuls. On suppose les données indépendantes et provenant de K sources poissonniennes de moyennes inconnues $\theta = (\theta_1, \ldots, \theta_K)$. On définit un vecteur de variables non-observées $z = (z_1, \ldots, z_n)$, $z_i \in \{1, \ldots, K\}$, et un modèle pour (y, z, θ) de la manière suivante

$$p(y_i|z_i, \theta) = (\theta_{z_i})^{y_i} e^{-\theta_{z_i}} / y_i!$$

$$p(z) \propto 1$$

$$p(\theta_k) \propto 1/\theta_k \, .$$

On suppose que les $K+n$ variables $(\theta_k, z_i)_{k,i}$ sont mutuellement indépendantes.

1. Décrire la loi $p(y|z, \theta)$ en séparant les produits faisant intervenir les ensembles d'indices $I_k = \{i : z_i = k\}$, $k = 1, \ldots, K$.

2. Décrire la loi a posteriori du vecteur de variables (z, θ).

3. Soit $i \in \{1, \ldots, n\}$, calculer la probabilité conditionnelle $p(z_i = k|y, \theta)$.

4. Soit n_k le nombre d'éléments dans I_k et $\bar{y}_k$ la moyenne empirique des

données y_i pour $i \in I_k$. On note $\theta_{-k} = (\ldots, \theta_{k-1}, \theta_{k+1}, \ldots)$ le vecteur θ privé de sa coordonnées k. Montrer que

$$p(\theta_k|\theta_{-k}, y, z) \propto \theta_k^{n_k \bar{y}_k - 1} \exp(-n_k \theta_k).$$

5. Décrire l'implantation d'un cycle de l'algorithme d'échantillonnage de Gibbs pour la loi a posteriori en langage `R`. On pourra utiliser la commande `rgamma` pour simuler des réalisations de la loi gamma.

6. À l'issue d'une exécution de l'algorithme précédent, comment peut-on estimer l'espérance de θ_k sachant y ? Comment peut-on estimer les proportions de chacune des composantes du mélange ?

7. On considère à nouveau l'échantillon y donné au début de l'énoncé. Pour cet échantillon, représenter un histogramme et superposer la loi de mélange obtenue par l'algorithme d'échantillonnage de Gibbs pour diverses valeurs de K. Quelle valeur de K vous semble-t-elle la plus appropriée pour modéliser l'échantillon de données ? Proposer une ou plusieurs statistiques pour vérifier le modèle et en calculer les lois prédictives (code `R` à donner). Critiquer la modélisation.

Ensimag 2A **2014-15**

Modèles probabilistes pour l'apprentissage

Décembre 2014

Documents manuscrits et polycopiés autorisés. Durée 3h.

Les zéros jouent toujours 2 fois.

Une source émet n signaux binaires indépendants, $z_i \in \{0, 1\}$, de fréquence inconnue, θ (i de 1 à n). Pour chaque émission telle que $z_i = 0$, la source émet un second signal binaire de fréquence identique et indépendant du premier signal. Ce nouveau signal remplaçe alors le premier signal. On note z la somme des valeurs obtenues lors de la première émission

$$z = \sum_{i=1}^{n} z_i \,.$$

Nous supposons que

$$p(z_i = 1|\theta) = \theta \quad \text{et} \quad p(z_i = 0|\theta) = 1 - \theta.$$

et que

$$p(\theta) = 2(1 - \theta), \quad 0 \leq \theta \leq 1.$$

Soit y la somme des n valeurs résultant de l'émission répétée. Pour tout i, nous avons

$$p(y_i = 1|z_i = 1, \theta) = 1, \qquad p(y_i = 1|z_i = 0, \theta) = \theta,$$

et

$$y = \sum_{i=1}^{n} y_i.$$

Dans ce problème, les variables z et θ sont **inconnues**, seule la variable y est observée. On cherche à simuler, puis à calculer la loi a posteriori des variables θ et z sachant l'observation y. La première partie propose un algorithme de simulation par chaîne de Markov. La seconde partie propose un algorithme de simulation exacte. La troisième partie propose une simplification et une représentation exacte de la loi a posteriori.

Note importante : *Il est fortement conseillé de commencer le problème en traitant les questions préliminaires et la partie 1. On choisira alors entre la partie 2 ou la partie 3 en indiquant clairement le choix effectué sur la copie. La partie restante pourra être traitée de manière facultative. Elle ne sera corrigée que lorsque l'on aura répondu à toutes les questions de la partie choisie.*

Questions préliminaires. Construction du modèle probabiliste.

1. Montrer que la loi conditionnelle $p(z|\theta)$ est égale à

$$p(z|\theta) = \binom{n}{z}\theta^z(1-\theta)^{n-z}, \quad z = 0, \ldots, n.$$

2. Montrer que la loi générative $p(y|z,\theta)$ est égale à

$$p(y|z,\theta) = \binom{n-z}{y-z}\theta^{y-z}(1-\theta)^{n-y}, \quad y = z, \ldots, n.$$

Première partie. Méthode de simulation par chaîne de Markov.

1. Montrer que la loi conditionnelle $p(z|\theta,y)$ est égale à

$$p(z|\theta,y) = \binom{y}{z}\frac{(1-\theta)^{y-z}}{(2-\theta)^y}, \quad z = 0, \ldots, y.$$

   Note. On pourra avantageusement remarquer qu'avec les valeurs du problème, nous avons

$$\binom{n-z}{y-z}\binom{n}{z} = \binom{n}{y}\binom{y}{z}$$

   et

$$\sum_{z=0}^{y}\binom{y}{z}(1-\theta)^{y-z} = (2-\theta)^y.$$

2. Montrer que la loi conditionnelle $p(\theta|z,y)$ est la loi bêta$(y+1, 2(n+1)-y-z)$.

3. Ecrire en langage `R` un algorithme de simulation de la loi a posteriori de la variable θ à l'aide d'une chaîne de Markov (utiliser la fonction `choose(n,k)` pour calculer les coefficients binomiaux).

4. Décrire une méthode de Monte-Carlo permettant d'évaluer la loi a posteriori, $p(z|y)$, de la variable cachée z, pour tout $z = 0, \ldots, y$.

5. Donner sous la forme d'une intégrale – que l'on ne cherchera pas à calculer – la probabilité a posteriori prédictive de l'événement "Obtenir un nouveau 1'' sachant y. En quel sens cette prédiction est-elle optimale ?

Deuxième partie. Loi de z et méthode de simulation exacte.

1. Montrer que la loi jointe $p(\theta, z|y)$ vérifie

$$p(\theta, z|y) \propto 2\binom{n}{y}\binom{y}{z}\theta^y(1-\theta)^{2n+1-y-z}, \quad 0 < \theta < 1, \quad z = 0, \ldots, y.$$

2. En déduire que la loi $p(z|y)$ satisfait à

$$p(z|y) \propto \frac{1}{(2n-z+2)}\binom{y}{z}/\binom{2n-z+1}{y}, \quad z = 0, \ldots, y.$$

Note. On se rappelera que la constante de normalisation de la loi bêta(α, β) est égale à

$$\int_0^1 \theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}, \quad \alpha, \beta > 1,$$

et que $\Gamma(n) = (n-1)!$ pour tout $n \geq 1$.

3. Ecrire un algorithme de simulation de la loi $p(z|y)$ en langage `R` utilisant les fonctions `choose(n,k)` et `sample`.

4. En déduire un algorithme de simulation exact de la loi a posteriori $p(\theta|y)$ en 2 étapes. Ecrire cet algorithme en langage `R`.

Troisième partie. Calcul exact de la loi a posteriori.

1. Soit $\alpha, \beta > 1$. Soit φ une variable aléatoire de loi bêta(α, β) et

$$\theta = 1 - \sqrt{1-\varphi}.$$

En introduisant la fonction de répartition de la variable θ, $p(\theta \leq t)$, puis en dérivant par rapport à t, montrer que

$$p(\theta) = 2\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}((2-\theta)\theta)^{\alpha-1}(1-\theta)^{2\beta-1}, \quad 0 < \theta < 1.$$

2. On considère 2 échantillons indépendants de longueur n, émis simultanément par la source de fréquence θ. On note ces deux échantillons $(z_1^1, \ldots, z_n^1)$ et $(z_1^2, \ldots, z_n^2)$. Montrer que la loi de l'échantillon $(y_1, \ldots, y_n)$ est identique à la loi de l'échantillon constitués des maxima pris terme à terme :

$$p(y_i = 1) = p(\max(z_i^1, z_i^2) = 1) = (2-\theta)\theta, \quad i = 1, \ldots, n.$$

3. Déduire la question précédente que la loi de y est donnée par

$$p(y|\theta) = \binom{n}{y}(2-\theta)^y\theta^y(1-\theta)^{2(n-y)}, \quad y = 0, \ldots, n.$$

4. Donner l'expression analytique de la densité de la loi a posteriori $p(\theta|y)$ à une constante près. Puis donner l'expression analytique exacte de la densité de la loi a posteriori $p(\theta|y)$ (la constante de proportionalité pourra être explicitée par identification à une loi connue).

5. Montrer que la loi a posteriori est la loi d'une variable θ obtenue par un changement de variable à partir d'une variable φ de loi bêta$(y+1, n-y+1)$. En déduire un algorithme de simulation exact de la loi a posteriori en langage `R`.

.

Ensimag 2A **2015-16**

Modèles probabilistes pour l'apprentissage

Décembre 2015

Documents manuscrits et polycopiés autorisés. Durée 3h.

L'exercice et le problème sont indépendants. Les points noirs indiquent l'importance accordée aux questions par les correcteurs.

Exercice – Corrélation entre estimation et prédiction. Un expert déclare qu'une variable inconnue, θ, suit la loi normale de moyenne m_0 et de variance σ_0^2 (loi a priori). Cette variable inconnue est observée avec une erreur aléatoire, ϵ, que l'on suppose indépendante de θ et de loi normale de moyenne nulle et de variance connue, σ^2. On note y la valeur observée :

$$y = \theta + \epsilon\,.$$

Pour simplifier la lisibilité des calculs, on pose $\beta = 1/\sigma^2$ et $\beta_0 = 1/\sigma_0^2$.

1. • • ○ Montrer que le couple (θ, y) est gaussien. Déterminer son espérance et sa matrice de covariance. En déduire que la loi a posteriori, $p(\theta|y)$, est identique à la loi de la variable aléatoire suivante

   $$\theta_1 = \frac{\beta_0 m_0 + \beta y}{\beta_0 + \beta} + \epsilon'$$

   où ϵ' est une variable aléatoire de normale de moyenne nulle et de variance

égale à $(\beta_0 + \beta)^{-1}$.

2. ●●○ Sans chercher à l'identifier, montrer que la loi prédictive a posteriori, $p(y_1|y)$, est identique à la loi de la variable aléatoire suivante

$$y_1 = \frac{\beta_0 m_0 + \beta y}{\beta_0 + \beta} + \epsilon' + \epsilon$$

où ϵ est une variable de loi normale indépendante de ϵ'.

3. ●○○ Identifier la loi prédictive, $p(y_1|y)$.

4. ●●○ Montrer que le couple (θ_1, y_1) est un couple gaussien. Identifier sa moyenne et sa matrice de covariance.

5. ●●● Quelle est la part de variance de la valeur prédite, y_1, expliquée par la valeur estimée, θ_1 ?

Problème de Monty Hall. Dans un célèbre jeu télévisé états-unien (*Let's Make a Deal*), un candidat se trouve face à trois portes fermées. Derrière l'une des trois portes (la bonne porte) se trouve un cadeau. Le candidat désigne l'une des trois portes. Le présentateur ouvre alors devant le candidat l'une des 2 portes non-désignées qu'il sait être vide, et lui pose la question suivante : *souhaitez vous changer de porte ?*

Questions probabilistes préliminaires faciles. On note B l'événement "le candidat choisit initialement la bonne porte", C l'événement "le candidat change de porte". G l'événement "le candidat ouvre finalement la bonne porte et gagne le jeu".

1. ∘∘∘ Donner sans calcul les probabilités conditionnelles suivantes : $p(G|B \cap C)$ et $p(G|\bar{B} \cap C)$, puis $p(G|B \cap \bar{C})$ et $p(G|\bar{B} \cap \bar{C})$.
2. ∘∘∘ On suppose B et C indépendants. Montrer que la probabilité conditionnelle de l'événement G sachant que le candidat change de porte est égale à $2/3$. Montrer que la probabilité conditionnelle de G sachant que le candidat conserve son choix initial est égale à $1/3$.

Analyse bayésienne. Après n répétitions du jeu, on sait que k candidats ont gagné le jeu et c'est la seule information dont on dispose. On souhaite alors estimer la probabilité, θ, qu'un candidat pris au hasard dans la population (états-unienne) change de porte. Cette probabilité est inconnue, et supposée être une variable aléatoire de loi uniforme sur $(0, 1)$ (loi a priori). On suppose que les candidats sont indépendants les uns des autres.

1. •∘∘ Montrer que la probabilité qu'un candidat pris au hasard dans la

population gagne le jeu est égale à $(1+\theta)/3$.

2. ●●○ Montrer que

$$p(\theta|k) \propto (1+\theta)^k (2-\theta)^{n-k} \quad 0 \leq k \leq n.$$

3. ○○○ Montrer que la densité de la loi a posteriori admet un maximum au point

$$\hat{\theta} = 3k/n - 1.$$

4. ●○○ Soit φ une variable aléatoire de loi bêta$(k+1, n-k+1)$. Calculer le terme général de la densité de la loi de la variable aléatoire $3\varphi - 1$.

5. ●●○ Montrer que la loi a posteriori de la variable θ correspond à la loi de la variable $3\varphi - 1$, où φ est une variable aléatoire de loi bêta$(k+1, n-k+1)$ conditionnée à se trouver dans l'intervalle $(1/3, 2/3)$.

6. ●●● Ecrire un algorithme de simulation de la loi a posteriori de la variable θ.

7. ●●○ Proposer une ligne de commande `R` permettant de déterminer un intervalle I tel que $p(\theta \in I|k) \approx 95\%$. Expliquer la nature de l'approximation effectuée.

Echantillonnage de Gibbs. Pour tout $i = 1, \ldots, n$, on sait désormais si le candidat i a gagné ($y_i = 1$) ou non ($y_i = 0$). A nouveau, c'est la seule information dont on dispose. On note $y = (y_1, \ldots, y_n)$.

Par ailleurs, on définit la variable indicatrice $z_i = 1$ si le candidat i change de porte et $z_i = 0$ sinon. On note $z = (z_1, \ldots, z_n)$ et

$$\bar{z} = \sum_{i=1}^{n} z_i \, .$$

La variable $\bar{z}$ représente le nombre de candidats ayant changé de porte à l'issue des n répétitions du jeu.

1. $\bullet\bullet\circ$ Montrer que la loi conditionnelle $p(\theta|z, y)$ est la loi bêta($\bar{z}+1, n-\bar{z}+1$).
2. $\bullet\bullet\circ$ Pour tout $i = 1, \ldots, n$, montrer que

$$p(z_i = 1|\theta, y_i = 0) = \frac{\theta}{2 - \theta} \, .$$

3. $\bullet\bullet\circ$ Pour tout $i = 1, \ldots, n$, montrer que

$$p(z_i = 1|\theta, y_i = 1) = \frac{2\theta}{1 + \theta} \, .$$

4. $\bullet\bullet\bullet$ Ecrire un algorithme d'échantillonnage de Gibbs permettant de simuler la loi $p(\theta, z|y)$.

Bouquet final pour les meilleurs. $\bullet\circ\circ$ Généraliser les solutions proposées dans les parties précédentes au problème à m portes, $m \geq 3$, ne comportant qu'un seul cadeau, où le présentateur propose un seul changement.

.