Une introduction aux statistiques inférentielles

Christophe Lalanne

Sommaire

1 Quelques rappels utiles de probabilités

Quelques-uns des concepts fondamentaux en statistique théorique nécessitent pour le lecteur de s'être bien approprié certains éléments de la Théorie des probabilités. Dans cette perspective, on se contentera de rappeler les axiomes du calcul des probabilités, les notions d'indépendance et de probabilités conditionnelles, celles-ci nous amenant directement à exposer le principe de Bayes. Enfin, nous définirons les variables aléatoires, à valeurs dans $\mathbb{N}^m$ ou $\mathbb{R}^m$, avec les lois de probabilité usuelles et les règles de manipulation qui leur sont associées. Les distributions de probabilité les plus utiles en biologie sont présentées dans des ouvrages généraux (Billingsley, 1995, Feller, 1968, Fisz, 1963, Johnson et al., 1994 and Kendall et al., 2004).

1.1 Les axiomes fondamentaux

Une probabilité est une fonction qui associe un nombre appartenant à l'intervalle $[0, 1]$ à un ensemble A. On dit que $\Pr(A)$ est la probabilité d'un ensemble, ou d'un événement, A. Généralement, on considère que les événements font partie d'une famille $\mathcal{A}$ de sous-ensembles d'un espace probabilisé, dénoté Ω. Si cette famille est close par rapport à la complémentation ($A \in \mathcal{A} \Rightarrow A^c \in \mathcal{A}$) ainsi qu'à la sommation dénombrable ($A_i \in \mathcal{A}, i = 1, 2, \ldots \Rightarrow (\bigcup_{i=1}^{\infty} A_i) \in \mathcal{A}$), et si elle contient l'ensemble vide $\emptyset$ (et donc $\Omega = \emptyset^c$), on parle d'une σ-algèbre de sous-ensembles de Ω.

L'axiomatique de Kolmogorov conduit à définir les propriétés suivantes :

i. $P(A) \in [0, 1],\ A \in \mathcal{A}$
ii. $P(\Omega) = 1$ (Ax. de normalisation)
iii. $P\left(\sum_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i),\ A_i \in \mathcal{A},\ A_i \cap A_j = \emptyset,\ i, j = 1, 2, \ldots$ (Ax. d'additivité)

L'additivité finie est une conséquence de (iii), et on retrouve le résultat bien connu :

$$P(A \cup B) = P(A) + P(B),\ A, B \in \mathcal{A},\ A \cap B = \emptyset.$$

La probabilité associée à la réunion de deux ensembles disjoints est la somme des probabilités associées à chacun de ces ensembles. On a également

$$P(\Omega) = P(A) + P(A^c) = 1,$$

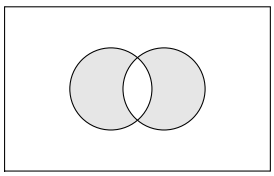
d'où également

$$P(A^c) = 1 - P(A).$$

Enfin, dans le cas plus général, où les événements A et B ne sont pas nécessairement mutuellement exclusifs, on a

$$P(A \cup B) = P(A) + P(B) - P(A \cap B),$$

où $A \cap B$ dénote l'intersection des ensembles A et B.



1.2 Indépendance, probabilités conditionnelles

À présent, les bases essentielles qui permettent de rendre une probabilité intuitivement consistante ont été posées et l'on peut définir la notion d'indépendance et de probabilité conditionnelle. Mathématiquement, deux événements A et B sont dits indépendants si et seulement si

$$P(A \cap B) = P(A)P(B).$$

La probabilité de A sachant B (i.e. conditionnellement à l'observation de l'événement B) est définie comme

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}.$$

Si les événements $B_1, B_2, \ldots$ sont mutuellement exclusifs ($B_i \cap B_j = \emptyset$) et sont collectivement exhaustifs ($\bigcup_{k=1}^{\infty} B_k = \Omega$), alors on peut décomposer $P(A) = P(A \cap \Omega)$ comme suit :

$$P(A) = \sum_{k=1}^{\infty} P(A \cap B_k) = \sum_{k=1}^{\infty} P(A \mid B_k)P(B_k). \tag{1}$$

L'expression ci-dessus est appelée *loi des probabilités totales*. La propriété d'exhaustivité de $B_1, B_2, \ldots$ n'est toutefois pas indispensable et l'on peut se contenter du fait que les ensembles sont tous disjoints, avec $A \subset \bigcup_{k=1}^{\infty} B_k$.

La probabilité conditionnelle $P(B_k \mid A)$ se calcule aisément, comme :

$$P(B_k \mid A) = \frac{P(A \cap B_k)}{P(A)} = \frac{P(A \mid B_k)P(B_k)}{\sum_{i=1}^{\infty} P(A \mid B_k)P(B_k)}$$

Dans ce cadre, $P(B_k \mid A)$ est appelée probabilité a posteriori de B_k, et l'expression ci-dessus est connue comme étant la *seconde formule de Bayes*.

1.3 Variables aléatoires

Contrairement à un cadre purement déterministe dans lequel une variable se voit attribuer une valeur unique, l'univers probabiliste repose sur des variables aléatoires (v.a.) qui peuvent prendre différentes valeurs aléatoires. Plus formellement, on définira une variable aléatoire comme une application de l'espace Ω vers l'ensemble $\mathbb{R}$ des réels.

Considérons dans un premier temps une v.a. discrète, X, à valeurs dans un ensemble fini (ou infini) dénombrable de $\mathbb{R}$. Une telle variable prend des valeurs $x_0, x_1, \ldots, x_k, \ldots$ avec probabilité $p_0, p_1, \ldots, p_k, \ldots$, sous la condition (de normalisation)

$$\sum_{k=0}^{\infty} p_k = 1.$$

La série (finie ou infinie) $\{p_0, p_1, \ldots\}$ est appelée la distribution de X.

Une v.a. continue, au contraire, prend ses valeurs dans un sous-intervalle de $\mathbb{R}$, et c'est sa fonction de répartition, notée $F_X(x)$, qui joue le rôle de distribution de probabilité. Celle-ci se définit comme

$$F_X(x) = P(X \leq x),$$

et exprime, pour un x donné, la probabilité de l'événement $X \leq x$. Les propriétés de $F_X(\cdot)$ sont : (i) $F_X(\cdot)$ est croissante, (ii) $F_X(-\infty) = 0$ et (iii) $F_X(+\infty) = 1$. Les intervalles sur lesquels $F_X(x)$ est constante correspondent aux intervalles pour lesquels la probabilité de X n'est pas définie, tandis que les sauts de $F_X(x)$ coïncident avec les masses discrètes de la distribution de probabilité de X.

Si $F_X(x)$ est différentiable, sa dérivée est appelée la fonction de densité de probabilité et on la note $f_X(x)$, avec

$$f_X(x) = \lim_{\Delta x \to 0} \frac{F(x < X \leq x + \Delta x)}{\Delta x} = \frac{dF_X(x)}{dx}.$$

On notera que l'inégalité est stricte à gauche. On a également

$$\int_{-\infty}^{x} f_X(\xi) d\xi = F_X(x),$$

et comme $F_X(+\infty) = 1$, on obtient facilement la condition de normalisation pour la distribution d'une v.a. continue X :

$$\int_{-\infty}^{+\infty} f_X(x) dx = \lim_{x \to +\infty} F_X(x) = 1.$$

Vecteurs aléatoires

Lorsque l'on est placé face à plusieurs distributions de v.a. et que l'on souhaite les analyser conjointement, on est amené à travailler avec des vecteurs aléatoires. Dans le cas discret, où X et Y prennent les valeurs $x_0, x_1, \ldots, x_k, \ldots$ et $y_0, y_1, \ldots, y_k, \ldots$, respectivement, la loi de probabilité conjointe est donnée par

$$p_{ij} = P(X = x_i, Y = y_j),$$

sous la condition

$$\sum_{i=0}^{\infty}\sum_{j=0}^{\infty} p_{ij} = 1.$$

Dans le cas continu, la fonction de répartition conjointe de (X, Y) est

$$F_{X,Y}(x, y) = P(X \leq x, Y \leq y),$$

et la densité de probabilité conjointe correspondante ($F_{X,Y}(x, y)$ est supposée uniformément continue) est donnée par

$$\begin{aligned} f_{X,Y}(x, y) &= \lim_{\Delta x \to 0, \Delta y \to 0} \frac{P(x < X \leq x + \Delta x, y < Y \leq y + \Delta y)}{\Delta x \Delta y} \\ &= \frac{\partial^2 F_{X,Y}(x, y)}{\partial x \partial y} \end{aligned}$$

sous la condition

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f_{X,Y}(x, y)dxdy = \lim_{x \to +\infty, y \to +\infty} F_{X,Y}(x, y) = 1$$

Distributions marginales

Les distributions bi-dimensionnelles, et plus généralement multi-dimensionnelles, peuvent être réduites à des distributions uni-dimensionnelles en calculant leurs distributions marginales. Pour une v.a. discrète X, distribuée de manière conjointe à Y, on a

$$p_i = P(X = x_i) = \sum_{j=0}^{\infty} p_i j,$$

tandis que pour une v.a. continue, la distribution marginale s'exprime sous la forme

$$F_X(x) = F_{X,Y}(x, \infty),$$

avec pour fonction de densité

$$f_X(x) = \int_{-\infty}^{+\infty} f_{X,Y}(x, y)dx.$$

On généralisera aisément les formules ci-dessus aux dimensions supérieures.

Opérations sur les variables aléatoires

Pour des v.a. X et Y indépendantes, leur loi de probabilité jointe satisfait

$$p_{ij} = p_i p_j$$

dans le cas discret, et

$$F_{X,Y}(x,y) = F_X(x)F_Y(y) \quad \text{ou} \quad f_{X,Y}(x,y) = f_X(x)f_Y(y)$$

dans le cas continu.

La distribution conditionnelle de X sachant $Y = y$ est donnée par une formule identique à celle exposée précédemment,

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

Lorsque l'on travaille avec des distributions conditionnelles, la formule suivante (« règle de la chaîne ») se révèle très utile :

$$f_{X,Y|Z}(x,y \mid z) = f_{X|Y,Z}(x \mid y,z)f_{Y|Z}(y \mid z).$$

Les opérations algébriques les plus fréquemment rencontrées se résument souvent au calcul de la distribution d'une v.a. définie par une relation sur d'autres v.a., indépendantes ou non, ou à laquelle on applique une transformation.

Soient X et Y deux v.a. dont la distribution conjointe est donnée par $f_{X,Y}(x,y)$. On peut définir une nouvelle v.a., Z, telle que

$$Z = X + Y,$$

et la distribution de Z peut être obtenue en intégrant sur la densité $f_{X,Y}(x,y)$, soit

$$f_Z(z) = \int\int_{x+y=z} f_{X,Y}(x,y) = \int_{-\infty}^{+\infty} f_{X,Y}(x, z-x)dx. \tag{2}$$

Lorsque X et Y sont indépendants, l'**intégrale 2** devient une simple intégrale de convolution :

$$f_Z(z) = \int_{-\infty}^{+\infty} f_X(x)f_Y(z-x)dx.$$

On peut également s'intéresser à la transformation d'une v.a. par une fonction g. Soit X une v.a. dont la fonction de densité est donnée par $f_X(x)$, et $Y = g(X)$. On se demande quelle est la loi de probabilité de Y ? Si l'on suppose que $g(\cdot)$ est strictement monotone, alors $g(\cdot)$ est inversible :

$$y = g(x) \Rightarrow x = g^{-1}(y).$$

À l'aide de cette fonction inverse, on peut représenter la fonction de répartition de Y, $F_Y(y)$, en fonction des réalisations de X, et par conséquent également en termes de la fonction de répartition de X, $F_X(x)$. On a alors

$$\begin{aligned} F_Y(y) &= P(Y \leq y) = P[g(X) \leq y] \\ &= \begin{cases} P[X \leq g^{-1}(y)] = F_X[g^{-1}(y)] & \text{pour } g(x) \text{ croissante} \\ P[X \geq g^{-1}(y)] = 1 - F_X[g^{-1}(y)] & \text{pour } g(x) \text{ décroissante.} \end{cases} \end{aligned}$$

En termes de densités, si elles existent, on a le résultat suivant :

$$f_Y(y) = \left| \frac{d}{dy} g^{-1}(y) \right| f_X\left[g^{-1}(y)\right].$$

1.4 Espérance mathématique et moments

À partir de maintenant, on se permettra « d'alléger » la notation, en considérant que si X désigne une v.a., ses réalisations possibles seront dénotées x, et sa loi de probabilité indexée par x, $f_X(x)$, sera notée simplement $f(x)$ lorsque cela ne prête à aucune confusion.

L'espérance d'une fonction $g(x)$ par rapport à la distribution d'une v.a. X discrète, telle que définie à la **page 3**, est donnée par

$$\mathbb{E}\left[g(X)\right] = \sum_{k=1}^{\infty} p_k g(x_k).$$

Dans le cas d'une v.a. x continue (i.e. par rapport à sa distribution $f_X(x)$), on a

$$\mathbb{E}\left[g(X)\right] = \int_{-\infty}^{+\infty} g(x) f_X(x) dx. \tag{3}$$

Lorsque $g(x) = x$, l'**expression 3** devient l'espérance de X, encore appelée moment d'ordre 1 de la v.a. X. On a alors

$$\mathbb{E}(X) = \sum_{k=0}^{\infty} p_k x_k \quad \text{(cas discret)}$$

et

$$\mathbb{E}(X) = \int_{-\infty}^{+\infty} x f_X(x) dx \quad \text{(cas continu)}.$$

Les moments d'ordre supérieur de X se définissent de manière analogue, en prenant $g(X) = X^n$ pour le moment d'ordre n et $g(X) = [X - \mathbb{E}(X)]^n$ pour le *moment centré* d'ordre n de la v.a. X. Le second moment centré figure parmi les plus intéressants puisqu'il correspond à ce que l'on nomme la *variance* d'une v.a. :

$$\mathbb{V}(X) = \sum_{k=0}^{\infty} p_k \left[x_k - \mathbb{E}(X)\right]^2, \text{ dans le cas discret,}$$

$$\mathbb{V}(X) = \int_{-\infty}^{+\infty} \left[X - \mathbb{E}(X)\right]^2 f_X(x) dx, \text{ dans le cas continu.}$$

La variance permet de mesurer la dispersion de la v.a. autour de son espérance mathématique. La racine carrée de la variance s'appelle l'écart-type et on le note

$$\sigma(X) = \sqrt{\mathbb{V}(X)}.$$

Il correspond au facteur d'échelle de la distribution de $X - \mathbb{E}(X)$.

L'espérance ou les moments d'une fonction d'une v.a. n'existent que si la série ou l'intégrale associée est convergente. Par exemple, dans l'**intégrale 3**, si la fonction $g(x)$ croît trop rapidement par rapport à x, celle-ci ne sera pas finie. De même, si la distribution d'une v.a. possède des queues de distribution trop épaisses, certains moments ne peuvent être définis, comme c'est le cas avec les distributions de Cauchy ou du t de Student.

Enfin, mentionnons les deux propriétés les plus importantes de ces opérateurs. L'espérance de la somme de deux v.a. est la somme de leurs espérances,

$$\mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y)$$

(quelles que soient les lois de X et Y !), et la variance de la somme de deux v.a. *indépendantes* est la somme de leur variance,

$$\mathbb{V}(X+Y) = \mathbb{V}(X) + \mathbb{V}(Y).$$

Dans le cas où X et Y ne sont pas indépendantes, il faudra associer le terme (signé) de covariance à la somme précédente.

1.5 Fonctions génératrices et fonctions caractéristiques

Les transformations vues aux paragraphes précédents se révèlent souvent suffisantes pour la plupart des situations que l'on rencontre dans le domaine des sciences expérimentales (Ditkin and Prudnikov, 1965 and Wilf, 1990). Elles servent à calculer les lois de probabilités, les moments et les fonctions de répartition d'une vaste gamme de v.a.. Elles sont également utilisées pour démontrer des propriétés de convergence en loi. Toutefois, on peut adopter une approche différente pour retrouver la distribution d'une v.a.

À une v.a. discrète X prenant des valeurs (x_i) avec probabilité p_i, on associe une fonction $P_X(z)$ d'un argument complexe z telle que

$$P_X(z) = \sum_{k=0}^{\infty} z^k p_k. \tag{4}$$

La fonction $P(z)$ ci-dessus est appelée fonction génératrice de X. En utilisant la propriété de normalisation des distributions de probabilité discrètes, on vérifie que $P(z)$ est bien définie pour tout z dans le disque unité. De **4**, on déduit $P(1) = 1$ et

$$\frac{d}{dz} P_X(z)\bigg|_{z=1} = \sum_{k=0}^{\infty} k p_k = \mathbb{E}(X),$$

de sorte que la différentielle de $P_X(z)$ (évaluée au point $z = 1$) nous donne l'espérance de X. De même, les dérivées successives permettent de calculer les moments d'ordre supérieur. Si l'on se donne deux v.a. indépendantes, X et Y, alors la fonction génératrice de leur somme est le produit de leurs fonctions génératrices :

$$P_{X+Y}(z) = P_X(z) P_Y(z). \tag{5}$$

Pour une v.a. continue X, de fonction de densité $f(x)$, on définit sa fonction caractéristique associée $F(j\omega)$ par

$$F_X(\omega) = \int_{-\infty}^{+\infty} f(x) \exp(-j\omega x) dx,$$

où j est le nombre imaginaire $\sqrt{-1}$ et ω un réel. La fonction caractéristique de X n'est autre que la transformée de Fourier de sa densité de probabilité et possède des propriétés similaires à celles démontrées plus haut, dans le cas des fonctions génératrices. Spécifiquement, on a $F_X(j0) = 1$ et

$$\frac{d}{d\omega} F_X(\omega)\bigg|_{\omega=0} = \int_{-\infty}^{+\infty} jx f(x) = j\mathbb{E}(X),$$

ainsi que

$$F_{X+Y}(\omega) = F_X(\omega) F_Y(\omega),$$

pour X et Y indépendantes.

En guise d'illustration, considérons deux v.a. discrètes indépendantes, X et Y suivant toutes les deux une loi géométrique de paramètre $p = 0.5$ et $p = 0.2$, respectivement. On cherche la loi suivie par la v.a. $X + Y$. On a

$$P_X(z) = \frac{0.5}{1 - 0.5z}, \quad P_Y(z) = \frac{0.2}{1 - 0.8z}$$

(cf. **section 1.6**), et

$$P_{X+Y}(z) = \frac{0.1}{(1 - 0.5z)(1 - 0.8z)}$$

d'après la **propriété 5**. Si l'on développe l'expression ci-dessus sous forme fractionnelle, on a

$$\frac{0.1}{(1 - 0.5z)(1 - 0.8z)} = \frac{A}{1 - 0.5z} + \frac{B}{1 - 0.8z},$$

d'où l'on déduit que $A = -1/6$ et $B = 4/15$. Ceci amène à conclure que

$$P\left[(X = Y) = k\right] = \frac{4}{15} 0.8^k - \frac{1}{6} 0.2^k, \quad k = 0, 1, 2, \ldots.$$

□

1.6 Lois de probabilités usuelles

Schéma de Bernoulli et loi binomiale

Le schéma de Bernoulli est sans doute l'un des schémas d'échantillonnage les plus courants en statistique. Un essai de Bernoulli est une expérience aléatoire dans laquelle deux

issues sont possibles, et on les dénomme souvent succès/échec. La distribution binomiale (**Figure 1**, a) décrit les probabilités p_k d'obtenir k succès sur un ensemble de K essais indépendants, sans considération de l'ordre des tirages,

$$p_k = \binom{K}{k} p^k (1-p)^{K-k}, \tag{6}$$

où p est la probabilité de succès d'un essai. Dans l'expression ci-dessus, $\binom{K}{k}$ désigne le nombre de combinaisons que l'on peut former avec k éléments pris parmi K ; il s'agit du nombre binomial défini comme

$$\binom{K}{k} = \frac{K!}{k!(K-k)!}.$$

La v.a. X peut être représentée comme une somme d'événements élémentaires, tous indépendants :

$$X = \sum_{k=1}^{K} X_k, \tag{7}$$

où X_k sont des v.a. de Bernoulli, avec $P(X_k = 1) = p$ et $P(X_k = 0) = q = 1-p$. Le nombre de succès dans une série d'expériences répétées (e.g. lancers d'une pièce ou d'un dé) est généralement modélisé par une distribution binomiale. Qui plus est, la loi binomiale sert de brique de base à la construction d'autres lois de probabilité discrète et possède des propriétés asymptotiques qui la relie aux distributions pour variables continues.

Les moments associés à la distribution binomiale se définissent comme suit :

$$\mathbb{E}(X) = Kp, \qquad \mathbb{V}(X) = Kp(1-p),$$

et sa fonction génératrice est

$$P(z) = (q + pz)^K \quad (q = 1-p).$$

Loi géométrique

Une v.a. discrète suit une loi géométrique lorsqu'elle prend les valeurs $0, 1, \ldots, k, \ldots$ avec probabilités

$$p_k = (1-p)^k p. \tag{8}$$

La distribution géométrique correspond typiquement à une situation où l'on répète une expérience de Bernoulli jusqu'à observer le premier succès. L'événement $X = k$, dont la probabilité est donnée en 8, peut être assimilé à la série de k échecs suivi d'un succès.

Les moments d'une v.a. X distribuée géométriquement sont

$$\mathbb{E}(X) = \frac{1-p}{p}, \qquad \mathbb{V}(X) = \frac{1-p}{p^2},$$

et la fonction génératrice correspondante est

$$P(z) = \frac{p}{1-(1-p)z}.$$

Loi binomiale négative

Comme pour la loi géométrique, on peut relier directement la loi binomiale négative (**Figure 1**, c) à la loi binomiale exposée plus haut. Ici, X vaut le nombre d'essais nécessaires pour observer r succès. La probabilité de l'événement $X = k$ (le $r$ième succès au $k$ième essai) est alors

$$p_k = \binom{k-1}{r-1} p^r (1-p)^{k-r},$$

qui découle du fait que l'événement « le $r$ième succès au $k$ième essai » est l'intersection (ou le produit) de deux événements indépendants : (a) $r-1$ succès en $k-1$ essais, soit $\binom{k-1}{r-1} p^{r-1}(1-p)^{k-r}$, et (b) un succès au $k$ième essai.

Les moments d'une v.a. X distribuée selon cette loi sont

$$\mathbb{E}(X) = r\frac{1-p}{p}, \qquad \mathbb{V}(X) = r\frac{1-p}{p^2},$$

et la fonction génératrice correspondante est

$$P(z) = \left(\frac{pz}{1-(1-p)z}\right)^r.$$

Loi de Poisson

Les v.a. de Poisson sont souvent utilisées pour modéliser des expériences impliquant l'observation du nombre d'occurences d'un événement survenant aléatoirement sur une période donnée de temps. Par exemple, il peut s'agir du nombre de clics dans un compteur Geiger, du nombre d'appels aux services des urgences ou du nombre d'accidents de voitures. Il est possible de dériver des v.a. de Poisson à partir d'un mécanisme stochastique pur que l'on appelle le *processus de Poisson* (Kingman, 1993). Dans ce schéma, des événements discrets interviennent sur un intervalle fini I tels que

1. les nombres d'événements survenant dans deux intervalles disjoints sont indépendants,
2. la probabilité qu'un événement survienne dans l'intervalle $(t, t+\Delta t)$ vaut $\lambda\Delta t + o(\Delta t)$, où λ est un paramètre d'intensité et $\lim_{\Delta t \to 0} o(\Delta t)/\Delta t = 0$.

Une v.a. de Poisson prend les valeurs $0, 1, \ldots, k, \ldots$ avec probabilités

$$p_k = P(X = k) = \exp(-\lambda)\frac{\lambda^k}{k!}, \tag{9}$$

où λ est le paramètre caractéristique de la loi, encore appelé cadence. Il existe une relation duale entre la loi de Poisson et la loi binomiale. Si (1) on a une séquence infinie de v.a. binomiales

$$X_1, X_2, \ldots, X_n, \ldots \tag{10}$$

avec les paramètres p_n et K_n qui décrivent les probabilités de succès et le nombre d'essais pour la distribution des X_n, et si (2) la séquence des paramètres obéit aux propriétés suivantes : $\lim_{n\to\infty} p_n = 0$, $\lim_{n\to\infty} K_n = \infty$ et $\lim_{n\to\infty} p_n K_n = \lambda$, alors la séquence **10** se comporte asymptotiquement comme une v.a. de Poisson. La distribution de Poisson est illustrée dans la **Figure 1** (b).

Les moments d'une v.a. X distribuée selon cette loi sont

$$\mathbb{E}(X) = \lambda, \qquad \mathbb{V}(X) = \lambda,$$

et la fonction génératrice correspondante est

$$P(z) = \exp\big(\lambda(z-1)\big).$$

Loi multinomiale

Il s'agit d'une généralisation de la loi binomiale dans laquelle chaque expérience aléatoire (indépendante) possède M issues possibles, avec les probabilités

$$p_m = P(\text{résultat de l'expérience} = m), \quad m = 1, 2, \ldots, M,$$

où

$$\sum_{m=1}^{M} p_m = 1. \tag{11}$$

Un vecteur aléatoire $X = [X_1, \ldots, X_M]$ suit une loi multinomiale avec les paramètres $p_1, \ldots, p_M$ et K répétitions si les v.a. X_m décrivent le nombre d'événements m en K essais. La distribution multinomiale est de la forme

$$P(k_1, k_2, \ldots, k_M) = \frac{K!}{k_1! k_2! \ldots k_M!} p_1^{k_1} p_2^{k_2} \ldots p_M^{k_M},$$

où $k_1, k_2, \ldots, k_M$ sont les nombres d'occurences et $\sum_{m=1}^{M} k_m = K$.

Loi hypergéométrique

La distribution hypergéométrique décrit le nombre de succès dans un schéma d'échantillonnage aléatoire sans remise, à partir d'une population finie avec deux types d'individus, notés 1 et 0. Pour une v.a. X suivant la loi hypergéométrique de paramètres N, M, n, l'événement $X = k$ est interprété comme k caractères de type 1 dans un échantillon de taille n, tirés aléatoirement dans une population de N individus, parmi lesquels M sont

de type 1 et $N-M$ sont de type 0. La distribution hypergéométrique (**Figure 1**, d) a la forme

$$p_k = P(X=k) = \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}. \tag{12}$$

L'**équation 12** découle du fait que parmi l'ensemble des résultats observables (au total, il y en a $\binom{N}{n}$) ceux avec k succès sont obtenus en combinant k individus de type 1 tirés d'un ensemble de M individus avec $n-k$ individus de type 0 tirés parmi les $N-M$ individus restants. La condition de normalisation pour la distribution hypergéométrique devient ainsi

$$\sum_{k=0}^{\min(n,M)} \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}} = 1.$$

Les moments d'une v.a. X décrit par une loi hypergéométrique sont

$$\mathbb{E}(X) = n\frac{M}{N}, \qquad \mathbb{V}(X) = n\frac{M(N-M)(N-n)}{N^2(N-1)},$$

La fonction génératrice peut être obtenue à partir d'une série hypergéométrique.

Loi normale (Laplace-Gauss)

La loi « normale » est sans doute la loi continue la plus importante. Son rôle essentiel se retrouve dans le théorème central limite (TCL) qui permet d'affirmer que la somme de plusieurs composantes aléatoires indépendantes de variances finies se distribue approximativement selon une loi gaussienne. En conséquence, les variables décrivant les erreurs de mesure, de même que certains paramètres décrivant des individus d'une population, comme les tailles, les poids ou les surfaces, sont modélisés à l'aide de ce type de loi.

De **7**, on peut voir que la distribution binomiale, lorsque K est grand, converge vers la loi normale. Les sommes de v.a. normales indépendantes sont également gaussiennes.

La distribution gaussienne (**Figure 2**, a) prend pour support la droite réelle, $\mathbb{R}$, et la fonction de densité d'une v.a. X gaussienne est

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right), \tag{13}$$

où μ et σ sont les paramètres pour l'espérance et l'écart-type, respectivement.

Les moments d'une v.a. X gaussienne sont

$$\mathbb{E}(X) = \mu, \qquad \mathbb{V}(X) = \sigma^2,$$

et la fonction caractéristique correspondante est

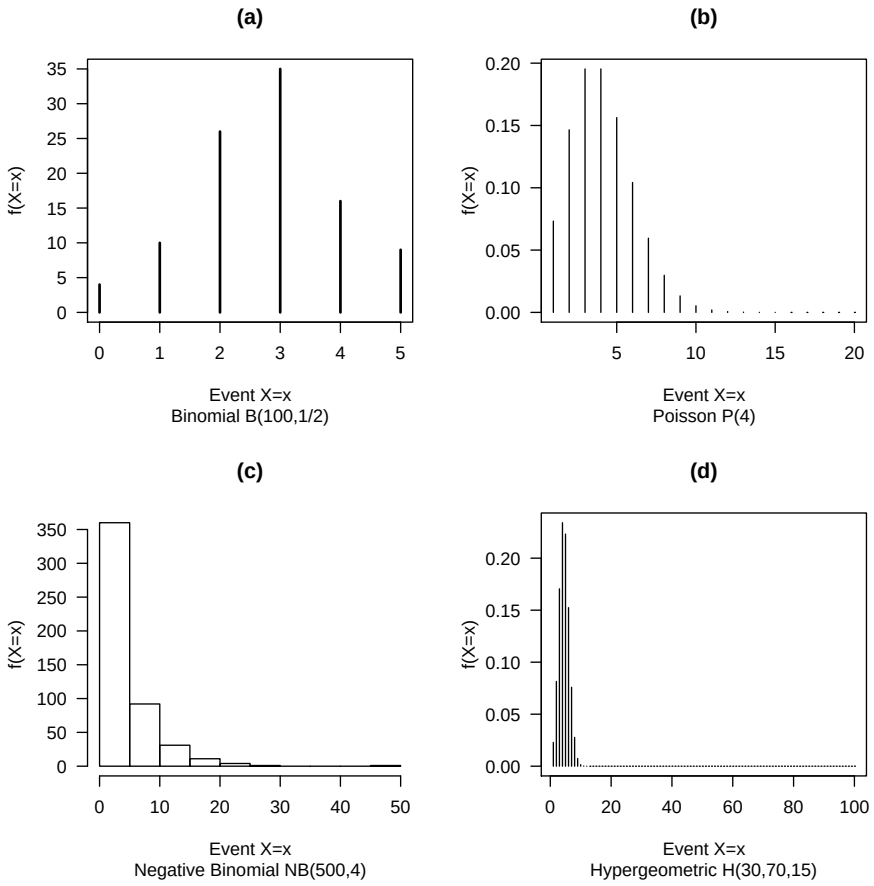


Figure 1 Quelques distributions de probabilités pour des v.a. discrètes

$$F(\omega) = \exp(j\mu\omega - \frac{\omega^2\sigma^2}{2}).$$

Loi exponentielle

La distribution exponentielle est la loi duale de la distribution géométrique décrite précédemment. On l'utilise généralement pour modéliser des intervalles de temps aléatoires, par exemple des temps d'attente, le temps entre deux échecs ou des durées de survie. Le temps entre l'occurence de deux événements successifs dans un processus poissonien se distribue également selon une loi exponentielle. La distribution exonentielle est définie sur l'intervalle $[0, \infty[$ et la fonction de densité d'une v.a. exponentielle prend la forme :

$$f(t) = a \exp(-at). \tag{14}$$

Le paramètre $a > 0$ est appelé la paramètre de cadence.

Les moments d'une v.a. T exponentielle sont

$$\mathbb{E}(T) = \frac{1}{a}, \qquad \mathbb{V}(T) = \frac{1}{a^2},$$

et la fonction caractéristique correspondante est

$$F(\omega) = \frac{a}{a - j\omega}.$$

Loi Gamma

La distribution Gamma (**Figure 2**, b) est le pendant de la loi binomiale négative, dans le cas continu. Elle est définie sur l'intervalle $[0, \infty[$ et peut être interprétée comme un temps aléatoire avec une structure composite ; par exemple, la somme de K v.a. exponentielle indépendantes et identiquement distribuées (i.i.d.) est une v.a. dont la loi est une loi Gamma. La densité de probabilité associée à une variable X qui suit une loi Gamma est données par :

$$f(x) = x^{k-1} \frac{\exp(-x/\theta)}{\theta^k \Gamma(k)}. \tag{15}$$

Dans l'expression ci-dessus, $\Gamma(k)$ est la fonction gamma eulérienne

$$\Gamma(z) = \int_0^\infty t^{z-1} \exp(-t) dt, \tag{16}$$

et $k > 0$, $\theta > 0$ sont les paramètres de la distribution Gamma, appelés respectivement paramètres de forme et d'échelle. Lorsque $k = 1$, **15** représente une densité de probabilité exponentielle. Si au contraire, $k = n/2$ et $\theta = 2$, on obtient la fonction de desnité d'une distribution du χ^2 à n degrés de liberté.

Les moments d'une v.a. X distribuée selon une loi Gamma sont

$$\mathbb{E}(X) = k\theta, \qquad \mathbb{V}(X) = k\theta^2,$$

et la fonction caractéristique correspondante est

$$F(\omega) = \frac{1}{(1 - j\theta\omega)^k}.$$

Loi Beta

La loi Beta (**Figure 2**, c) est définie sur l'intervalle $[0, 1]$. La densité de probabilité correspondante est donnée par

$$f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1} \tag{17}$$

où $x \in]0, 1[$ et $a > 0$, $b > 0$ sont des paramètres, tandis que Γ est toujours la fonction eulérienne introduite dans le cas de la loi Gamma (**Eq. 16**). En modifiant a et b, on fait varier la forme du graphe de la fonction de densité ci-dessus. Lorsque $a > 1$, $b > 1$, la

densité de probabilité a une forme parabolique, alors que lorsque $a < 1$, $b < 1$, la fonction de densité est en forme de U. Lorsque $a = 1$, $b = 1$, la densité de probabilité 17 décrit une distribution uniforme sur l'intervalle $[0, 1]$.

Les moments d'une v.a. X distribuée selon une loi Beta sont

$$\mathbb{E}(X) = \frac{a}{a+b}, \qquad \mathbb{V}(X) = \frac{ab}{(a+b)^2(a+b+1)}.$$

La fonction caractéristique associée à une loi Beta est donnée par une somme de séries hypergéométriques.

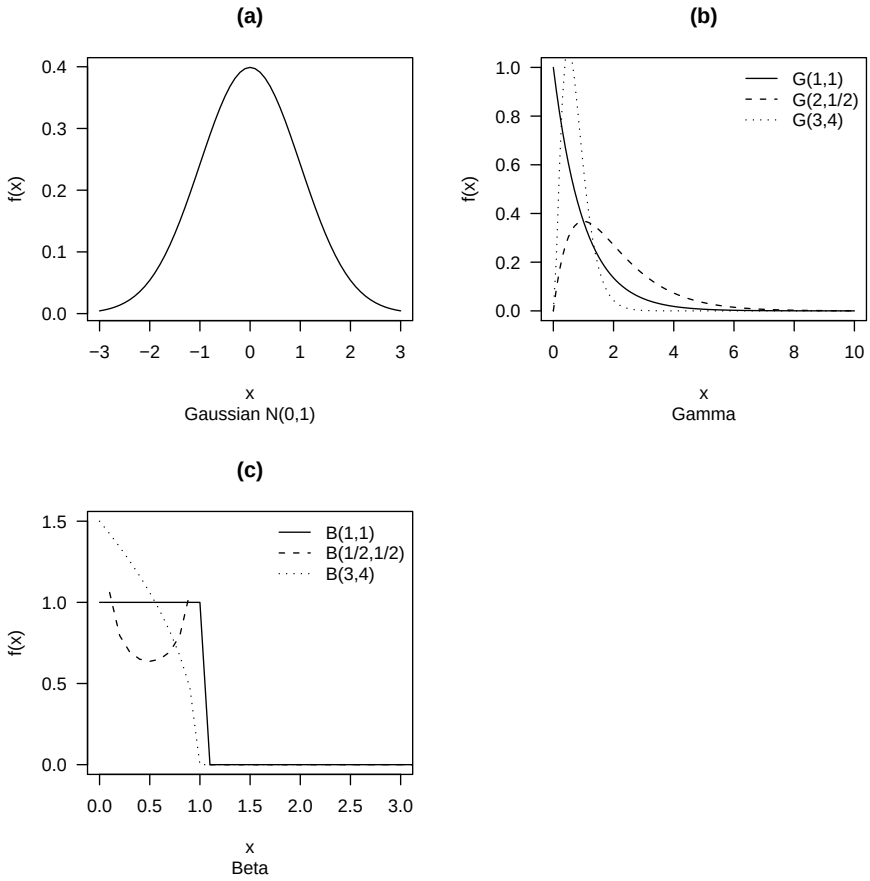


Figure 2 Quelques fonctions de densité de probabilités pour des v.a. continues

2 Méthode d'estimation de paramètres

2.1 Maximisation de la vraisemblance

Il est assez fréquent de chercher à déterminer de quelle distribution les observations dont nous disposons ont été échantillonnées. Il s'agit d'un problème d'*estimation*. La théorie

de l'estimation constitue une partie importante des statistiques, et plusieurs méthodes permettent d'estimer les paramètres d'une loi. En pratique, la méthode du maximum de vraisemblance (MV) est celle qui est le plus souvent utilisée. Nous en présentons ici le principe général. Dans sa forme paramétrique, on suppose que les observations ont été tirées d'une distribution appartenant à une famille de lois paramétriques. En d'autres termes, les observations $x_1, x_2, \ldots, x_N$ sont des réalisations i.i.d. d'une variable aléatoire X de distribution $f(x, p)$, où $f(\cdot, \cdot)$ peut être une loi discrète, continue ou une fonction de répartition. Lorsque l'on traite la fonction $f(x_n, p)$ comme une fonction du paramètre p pour un x_n fixé, elle est appelée la vraisemblance de l'observation x_n. La fonctionnelle $f(\cdot, \cdot)$ est connue mais pas la valeur du ou des paramètres p.

Pour estimer les paramètres p d'une distribution de probabilité à partir des réalisations observées d'une v.a. ou d'un vecteur aléatoire X, on utilisera le principe du MV qui repose sur l'idée que puisque les événements avec une probabilité élevée surviennent plus fréquemment que ceux pour qui la probabilité de survenue est faible, alors il est naturel de considérer que *ce qui est arrivé était le plus probable*. Par conséquent, le meilleur estimateur de p est la valeur $\hat{p}$ qui maximise la vraisemblance de l'échantillon, i.e.

$$L(p, x) = L(p) = f(x_1, x_2, \ldots, x_N, p) = \prod_{n=1}^{N} f(x_n, p),$$

où $x = x_1, x_2, \ldots, x_N$ et le produit des fonctions de densité individuelles découle de l'indépendance des observations. Mathématiquement,

$$\hat{p} = \arg\max \prod_{n=1}^{N} f(x_n, p).$$

Il est plus facile de travailler avec le log de la fonction de vraisemblance, et on parle alors de la log-vraisemblance $\ell(x_1, x_2, \ldots, x_N, p)$,

$$\ell(x_1, x_2, \ldots, x_N, p) = \ln\big(L(x_1, x_2, \ldots, x_N, p)\big) = \sum_{n=1}^{N} \ln\big(f(x_n, p)\big),$$

qui transforme les produits en sommes, et grâce à la monotonicité de la fonction logarithme donne le même $\hat{p}$ que lorsqu'on travaille directement avec $L(x_1, x_2, \ldots, x_N, p)$. Ce principe d'applique dans le cas discret comme dans le cas continu.

Les exemples suivants permettront sans doute de mieux comprendre le principe général de l'estimation par MV.

Distribution binomiale

Pour une v.a. X distribuée selon une loi binomiale comme en 6, en supposant que la réalisation observée inclut k succès parmi K essais, on peut maximiser la vraisemblance par rapport à p, on obtient l'EMV

$$\hat{p} = \frac{k}{K}.$$

De manière plus générale, on peut considérer une expérience avec K expériences de Bernoulli, répétées N fois, sachant qu'on enregistre les nombres de succès $k_1, k_2, \ldots, k_N$. Cela amène à la log-vraisemblance suivante :

$$\ell(k_1, k_2, \ldots, k_N, p) = \sum_{n=1}^{N} \left[k_n \ln p + (K - k_n) \ln(1-p) + \ln \binom{K}{k_n} \right]. \tag{18}$$

En maximisant l'**expression 18** par rapport à p, on obtient l'estimateur

$$\hat{p} = \frac{\sum_{n=1}^{N} k_n}{NK}.$$

Distribution multinomiale

On montre aisément que la log-vraisemblance correspondant à la distribution multinomiale prend la forme :

$$\ell(k_1, k_2, \ldots, k_M, p_1, p_2, \ldots, p_M) = \ln \frac{K!}{k_1! k_2! \ldots k_M!} + \sum_{m=1}^{M} k_m \ln p_m.$$

La maximisation de cette expression par rapport aux paramètres ne peut se faire qu'en prenant en considération la contrainte **11**, ce qui amène à construire la fonction de Lagrange suivante :

$$\begin{aligned} L &= (k_1, k_2, \ldots, k_M, p_1, p_2, \ldots, p_M, \lambda) \\ &= \ln \frac{K!}{k_1!, k_2!, \ldots, k_M!} + \sum_{m=1}^{M} k_m \ln p_m - \lambda \left(\sum_{m=1}^{M} p_m - 1 \right) \end{aligned}$$

où λ désigne le coefficient de Lagrange, et l'EMV est alors :

$$\hat{p}_m = \frac{k_m}{K}.$$

Distribution de Poisson

Soit X une v.a. de Poisson dont la distribution est donnée en **9**. Pour N réalisations indépendantes $k_1, k_2, \ldots, k_N$ de X, on a la fonction de log-vraisemblance suivante :

$$\ell(k_1, k_2, \ldots, k_N, \lambda) = \sum_{i=1}^{N} \big(-\lambda + k_i \ln(\lambda) - \ln(k_i!) \big),$$

qui prend son maximum en

$$\hat{\lambda} = \frac{\sum_{i=1}^{N} k_i}{N}.$$

Distribution géométrique

Si l'on considère une v.a. X suivant une loi géométrique, le paramètre à estimer est $p \in [0,1]$. Soient N réalisations indépendantes, $k_1, k_2, \ldots, k_N$, de X. Alors la log-vraisemblance prend la forme :

$$\ell(k_1, k_2, \ldots, k_N, p) = \sum_{n=1}^{N} \big((k_n - 1)\ln(1-p) + \ln p\big),$$

et l'EMV, $\hat{p}$, est

$$\hat{p} = \begin{cases} \frac{N}{\sum_{n=1}^{N} k_n} & \text{si } \sum_{n=1}^{N} k_n \geq 1 \\ 1 & \text{si } \sum_{n=1}^{N} k_n = 0. \end{cases}$$

Distribution gaussienne

La fonction de densité de probabilité d'une loi normale est fournie en **13**. La fonction de log-vraisemblance résultant de l'observation de N réalisations indépendantes $x_1, x_2, \ldots, x_N$ de X est alors

$$\ell(x_1, x_2, \ldots, x_N, \mu, \sigma) = \sum_{n=1}^{N} \left[-\frac{1}{2}\ln(2\pi) - \ln\theta - \frac{(x_n - \mu)^2}{2\sigma^2} \right].$$

Le maximum de $\ell(x_1, x_2, \ldots, x_N, \mu, \sigma)$ est atteint pour les valeurs $\hat{\mu}$ et $\hat{\sigma}$ données par la moyenne et la variance empiriques, soient :

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad \text{et} \qquad \hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{\mu})^2. \tag{19}$$

Distribution exponentielle

Soient N réalisations $t_1, t_2, \ldots, t_N$ d'une v.a. exponentielle T. La log-vraismeblance correspondant à l'échantillon observé est

$$\ell(t_1, t_2, \ldots, t_N, a) = \sum_{n=1}^{N}(-at_n + \ln a),$$

qui donne, après maximisation, l'EMV suivant :

$$\hat{a} = \frac{\sum_{n=1}^{N} t_n}{N}.$$

2.2 Autres méthodes d'estimation

Si l'on considère souvent la méthode MV comme une méthode pratique et efficace, il existe d'autres méthodes d'estimation qui peuvent être employées, justement la méthode par MV atteint ses limites (e.g. problèmes de grande complexité, coût en temps de calcul élevé, existence de maxima locaux multiples). D'autre part, les estimateurs obtenus par la méthode MV, i.e. les EMV, ne sont qu'asymptotiquement sans biais.

L'une des méthodes alternatives à l'EMV est la *méthode des moments*. Celle-ci repose sur la loi des grands nombres. Soit une v.a. X, dont la densité de probabilité est donnée par $f_X(x,p)$ et dépend d'un paramètre p. L'espérance de X, $\mathbb{E}(X,p) = \int x f_X(x,p)dx$, dépend de la valeur de p et peut être estimée par la moyenne empirique. La loi des grands nombres nous assure, sous certaines conditions de régularité, que pour de grands échantillons la moyenne empirique sera proche de l'espérance de X. Par conséquent, le moment de l'estimateur $\hat{p}$ peut être obtenu en résolvant l'équation suivante par rapport à p :

$$\frac{1}{N}\sum_{n=1}^{N} x_n = \int_{-\infty}^{+\infty} x f_X(x,p)dx.$$

Dans tous les eemples exposés précédement, l'estimateur obtenu par la méthode des moments coïncide avec l'EMV. Toutefois, ce n'est pas toujours le cas. Les exemples suivants devraient permettre de mettre en évidence les différences entre ces deux approches.

Distribution uniforme

On considère une distribution uniforme, représentée dans la Figure ??. L'intervalle sur lequel est définie la fonction de densité s'étend de 0 à a. On calcule dans un premier temps l'EMV de a, que l'on notera $\hat{a}_{ML}$. Dans la Figure ??, les valeurs $x_1, x_2, \ldots, x_N$ $(N = 6)$ sont indiquées par un trait coupant l'axe des abscisses, et on suppose 3 valeurs a_1, a_2 et a_3 du paramètre a, avec $a_1 < \max_{1\leq n\leq N} x_n$, $a_2 = \max_{1\leq n\leq N} x_n$ et $a_3 > \max_{1\leq n\leq N} x_n$. Les log-vraisemblances correspondantes sont dans chaque cas

$$\ell(x_1, x_2, \ldots, x_N, a_1) = -\infty$$

puisque deux observations sont impossibles dès lors que $\hat{a} = a_1$, et

$$\ell(x_1, x_2, \ldots, x_N, a_i) = -N \ln a_i, \quad i = 2, 3.$$

On en déduit que l'EMV de a est égal à a_2, et donc

$$\hat{a}_{ML} = \max_{1\leq n\leq N} x_n.$$

Puisque $\mathbb{E}(X) = a/2$, l'estimateur par la méthode des moments est

$$\hat{a}_{mom} = \frac{2}{N}\sum_{n=1}^{N} x_n.$$

Si l'on souhaite comparer ces deux estimateurs, il est nécessaire de faire quelques cacul d'espérance et de variance. Pour les espérances, on a

$$\mathbb{E}(\hat{a}_{ML}) = \mathbb{E}\left(\max_{1\leq n\leq N} X_n\right) = a\frac{N}{N+1}$$

et

$$\mathbb{E}(\hat{a}_{mom}) = \mathbb{E}\left(\frac{2}{N}\sum_{n=1}^{N} X_n\right) = a.$$

Pour les variances, les calculs donnent

$$\mathbb{V}(\hat{a}_{ML}) = \mathbb{V}\left(\max_{1\leq n\leq N} X_n\right) = \frac{Na^2}{(N+1)^2(N+2)}$$

et

$$\mathbb{V}(\hat{a}_{mom}) = \mathbb{V}\left(\frac{2}{N}\sum_{n=1}^{N} X_n\right) = \frac{a^2}{3N^2}.$$

On en déduit que la variance de l'EMV est inférieure à celle de l'estimateur par la méthode des moments. Leur rapport est approximativement proportionnel à la taille de l'échantillon N. Néanmoins, contrairement à l'estimateur des moments, l'EMV est biaisé puisque son espérance n'est pas égale à a.

On remarquera que l'on peut tout à fait baser l'estimation de a sur des moments d'ordre ≥ 2. Pour le $k$ième moment d'une v.a. X distribuée uniformément, on a

$$\mathbb{E}(X^k) = \frac{a^{k+1}}{k+1},$$

ce qui donne l'estimateur du $k$ième moment de a :

$$\hat{a}_{mom,k} = \left[\frac{k+1}{N}\sum_{n=1}^{N} x_n^k\right]^{\frac{1}{k+1}}.$$

On pourra vérifier que la statistique ci-dessus converge vers l'EMV de a lorsque $k \to \infty$.

Distribution de Cauchy

On considère une v.a. X qui suit une loi de Cauchy, dont la fonction de densité est définie comme

$$f(x, a) = \frac{1}{\pi\big(1 + (x-a)^2\big)},$$

avec un paramètre inconnu de position, a, à estimer. La distribution de Cauchy diffère des autres distributions vues jusqu'alors dans la mesure où elle ne possède pas de moments

finis. Ceci résulte de ce que pour tout $k \geq 1$, $\mathbb{E}(|X^k|)$ devient une intégrale impropre, non convergente :

$$\int_{-\infty}^{+\infty} \frac{|x|}{\pi\big(1+(x-a)^2\big)} dx = \infty. \tag{20}$$

Par conséquent, les estimateurs des moments n'ont aucun sens.

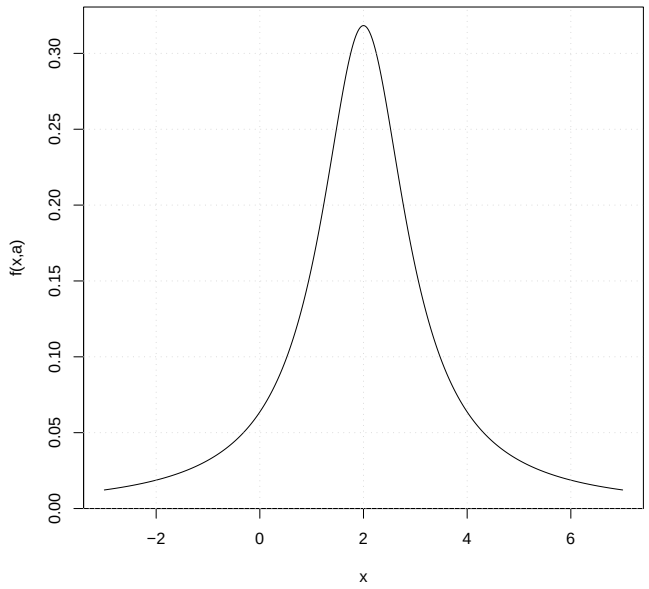


Figure 3 Graphe de la fonction de densité de probabilité de la distribution de Cauchy de paramètre $a = 2$.

La fonction de densité est représentée dans la **Figure 3**. À la lecture de ce graphique, on pourrait penser que si N réalisations, $x_1, x_2, \ldots, x_N$ d'une v.a. X tirée de cette loi étaient observées, la moyenne empirique $(1/N)\sum_{i=1}^{N} x_i$ semble un bon estimateur de a. Cela n'est toutefois pas le cas puisque, d'après **20**, la variance de $(1/N)\sum_{i=1}^{N} x_i$ est infinie pour tout N.

Si l'on dérive la log-vraisemblance par rapport à a et si l'on cherche à l'annuler, on a

$$\ell(x_1, x_2, \ldots, x_N, a) = \sum_{n=1}^{N} -\ln\pi - \ln\big(1+(x_n-a)^2\big)$$

ce qui entraîne la relation suivante pour l'EMV de a :

$$\sum_{n=1}^{N} \frac{x_n - \hat{a}}{1+(x_n-\hat{a})^2} = 0.$$

Sauf pour $N = 1$ et $N = 2$, cette équation ne peut être résolue que numériquement pour obtenir $\hat{a}$. Mais on peut montrer que, à partir de $N = 3$ (Hanson and Wolf, 1996), l'estimateur résultant est sans biais et de variance finie. Un autre estimateur de a, plus simple que l'EMV, est la médiane empirique (Hanson and Wolf, 1996), qui est également sans biais et de variance finie.

2.3 Estimateurs de variance minimale

On a vu que l'on peut construire des estimateurs de différentes manières (e.g. MV, méthode des moments). Une question que l'on peut se poser est naturellement : existe-t-il un estimateur de variance minimale, ou du moins de variance plus petite que celle de l'EMV ? On peut montrer que dans de nombreux cas, l'EMV est l'estimateur de variance minimale lorsque $n \to \infty$. Toutefois, pour des échantillons de taille finie, il existe souvent des statistiques possédant une variance plus petite. Par exemple, dans le cas de la distribution de Cauchy vue à la section précédente, on peut trouver numériquement un estimateur de variance minimale $\hat{a}$ (Hanson and Wolf, 1996). Nous présentons dans les lignes suivantes les outils techniques utiles pour dériver de tels estimateurs.

Information de Fisher

L'information de Fisher, $I(p)$, où p est le paramètre d'une distribution de probabilité $f(x, p)$ se définit comme

$$I(p) = \mathbb{E}\left(\left(\frac{\partial}{\partial p} \log f(x,p) \right)^2 \right) = -\mathbb{E}\left(\frac{\partial^2}{\partial p^2} \log f(x,p) \right). \tag{21}$$

Il s'agit donc de l'espérance de la dérivée de la log-vraisemblance, ou encore de l'espérance, au signe près, de la dérivée seconde (le hessien dans le cas où plusieurs paramètres entrent en jeu). De cette définition, on en déduit que l'information de Fisher est additive par rapport à des mesures indépendantes répétées, c'est-à-dire que

$$I_{X_1,X_2}(p) = I_{X_1}(p) + I_{X_2}(p) = 2I_{X_1}(p), \tag{22}$$

où les indices $\{1, 2\}$ dénotent des mesures différentes. L'**égalité 22** est valide pour deux séries i.i.d..

Théorème de Cramer-Rao

Le théorème, ou la borne, de Cramer-Rao dit que tout estimateur sans biais $\hat{a}$ d'un paramètre a doit vérifier

$$\mathbb{V}(\hat{p}) \geq \frac{1}{I(p)}. \tag{23}$$

Avec **23**, il est possible de calculer la limite inférieure de la variance de n'importe quel estimateur sans biais.

Concernant le paramètre de position d'une distribution normale, on obtient à partir de **21** l'information de Fisher :

$$I_{X_1,X_2,\ldots,X_N}(\mu) = \frac{N}{\sigma^2}.$$

Pour l'estimateur $\hat{\mu}$ donné en **19**, on peut calculer $\mathbb{V}(\hat{\mu}) = \sigma^2/N$ qui, en vertu du théorème de Cramer-Rao, prouve que $\hat{\mu}$ est de variance minimale et aucun estimateur de meilleure qualité ne peut être trouvé.

Considérons de nouveau la distribution de Cauchy et son paramètre de position, a. L'expression **21** nous permet de calculer l'information de Fisher correspondant aux observations $X_1, X_2, \ldots, X_N$,

$$I_{X_1,X_2,\ldots,X_N}(a) = \frac{N}{2}.$$

Ceci donne comme borne inférieure pour la variance de n'importe quel estimateur sans biais $\hat{a}$,

$$\mathbb{V}(\hat{a}) \geq \frac{2}{N}.$$

On peut montrer numériquement que l'estimateur de a présenté **page 21** n'atteint pas cette borne.

La borne de Cramer-Rao comme estimateur de la variance

Si l'on considère que dans de nombreuses applications, l'**expression 23** est assez précise, on l'utilise comme approximation pour les estimateurs de variance, soit

$$\mathbb{V}(\hat{p}) \simeq \frac{1}{I(p)}.$$

Les estimés des paramètres d'une distribution sont souvent obtenus en maximisant numériquement la fonction de vraisemblance de l'échantillon. Lorsqu'il n'existe aucune forme analytique, l'information de Fisher, $(\partial/\partial p) \log f(x,p)$, peut être obtenue numériquement par ré-échantillonnage. Par ré-échantillonnage on entend le moyennage de $\big((\partial/\partial p) \log f(x,p)\big)^2$ à partir de simulations numériques basées sur une variante de la méthode MCMC (voir plus loin).

Exhaustivité

Une statistique dite *exhaustive* possède la propriété de fournir la même information que l'échantillon complet. L'exhaustivité d'une statistique peut être vérifiée grâce au critère de factorisation de Fisher, selon lequel $t(x_1, x_2, \ldots, x_N)$ est une statistique exhaustive pour les observations $x_1, x_2, \ldots, x_N$ si

$$f(x_1, x_2, \ldots, x_N, p) = g(t,p)h(x_1, x_2, \ldots, x_N), \tag{24}$$

pour g et h deux fonctions données. En introduisant 24 dans 21, on peut vérifier que

$$I_{X_1,X_2,\ldots,X_N}(p) = I_{t(X_1,X_2,\ldots,X_N)}(p).$$

Théorème de Rao-Blackwell

Le théorème de Rao-Blackwell montre comment il est possible d'améliorer les estimateurs d'un paramètre à l'aide du principe d'exhaustivité. Si l'on note $\hat{p}$ un estimateur de p, étant données les observations $X_1, X_2, \ldots, X_N$, on peut définir un nouvel estimateur $\hat{p}^{new}$ comme l'espérance conditionnelle

$$\hat{p}^{new} = \mathbb{E}\left(\hat{p} \mid t(X_1, X_2, \ldots, X_N)\right)$$

où $t(X_1, X_2, \ldots, X_N)$ est une statistique suffisante pour p. Le théorème de Rao-Blackwell dit alors que

$$\mathbb{E}\left((\hat{p}^{new} - p)^2\right) \leq E\left((\hat{p} - p)^2\right).$$

Si l'on considère à nouveau la distribution uniforme, sachant que $t(x_1, x_2, \ldots, x_N) = \max(x_1, x_2, \ldots, x_N)$ est une statistique suffisante pour le paramètre a, on peut améliorer l'estimateur du moment $\hat{a}_{mom}$ en définissant l'estimateur RB suivant :

$$\hat{a}^{RB} = \mathbb{E}\left(\frac{2}{N}\sum_{n=1}^{N} X_n \mid \max(X_1, X_2, \ldots, X_N)\right).$$

2.4 Exemple d'application : construction de différentes statistiques de test

Pour illustrer les notions vues à la section précédente, on va s'intéresser à la mise en œuvre des trois statistiques de test les plus fréquemment rencontrées : le test de Wald, le test du score et le test du rapport de vraisemblance. Pour cela, on considère une expérience aléatoire dans laquelle on lance successivement n fois la même pièce de monnaie. On note $\theta = P(\text{« face »})$. On observe

$$\begin{cases} Y_i = 1 & \text{si le } i\text{e lancer amène « face »,} \\ Y_i = 0 & \text{sinon,} \end{cases}$$

de sorte que $Y_i \sim \mathcal{B}(1, \theta)$.

Dans un premier temps, on peut calculer l'information de Fisher et l'EMV $\hat{\theta}$ de θ. Pour cela, nous avons besoin de développer la fonction de vraisemblance pour une observation. On a donc :

$$\tilde{f}(y_i; \theta) = \theta^{y_i}(1 - \theta)^{1-y_i}$$

d'où

$$\log \tilde{f}(y_i;\theta) = y_i \log\theta + (1-y_i)\log(1-\theta)$$
$$\frac{\partial \log \tilde{f}(y_i;\theta)}{\partial\theta} = \frac{y_i}{\theta} - \frac{1-y_i}{1-\theta} = \frac{y_i-\theta}{\theta(1-\theta)}$$

On notera que le vecteur score est centré car $\mathbb{E}(Y_i) = 1\times\theta + 0\times(1-\theta) = \theta$. L'information de Fisher n'est autre que la variance du vecteur score :

$$\begin{aligned}
\tilde{I}(\theta) &= \mathbb{V}\left(\frac{\partial \log \tilde{f}(y_i;\theta)}{\partial\theta}\right) \\
&= \mathbb{V}\left(\frac{Y_i-\theta}{\theta(1-\theta)}\right) \\
&= \frac{1}{\big(\theta(1-\theta)\big)^2}\mathbb{V}(Y_i-\theta) \\
&= \frac{1}{\theta(1-\theta)} \quad \text{car } \mathbb{V}(Y_i) = \theta(1-\theta)
\end{aligned}$$

À présent, on peut s'intéresser à $\hat{\theta}$. Pour cela, il nous faut l'expression de la log vraisemblance sur l'ensemble des observations. Elle se déduit aisément de l'expression précédente :

$$\begin{aligned}
L_n(\theta) &= \sum_{i=1}^{n} \log \tilde{f}(y_i;\theta) \\
&= \sum_{i=1}^{n} y_i \log\theta + (1-y_i)\log(1-\theta)
\end{aligned}$$

On en déduit la dérivée première de la log vraisemblance :

$$\begin{aligned}
\frac{\partial L_n(\theta)}{\partial\theta} &= \sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log \tilde{f}(y_i;\theta) \\
&= \sum_{i=1}^{n} \frac{y_i-\theta}{\theta(1-\theta)} \\
&= \frac{\left(\sum_{i=1}^{n} y_i\right) - n\theta}{\theta(1-\theta)}
\end{aligned}$$

et on peut vérifier que

$$\left.\frac{\partial L_n(\theta)}{\partial\theta}\right|_{\hat{\theta}} = 0$$

avec $\hat{\theta} = \sum_{i=1}^{n} y_i/n$, qui est tout simplement la fréquence empirique des « pile ».

Si l'on pose l'hypothèse nulle $H_0 : \theta = \frac{1}{2}$, on peut calculer les trois statistiques de test considérées plus haut. Notons que cette hypothèse est équivalente à $H_0 : \theta - \frac{1}{2} = 0$, et l'on peut se donner une fonction $g(\theta) = \theta - \frac{1}{2}$, à valeurs dans $\mathbb{R}$.

Le test de Wald est construit comme suit :

$$\xi^W = n(\hat{\theta} - \frac{1}{2})\tilde{I}^{-1}(\hat{\theta} - \frac{1}{2}),$$

d'où l'on en tire, après simplification,

$$\xi^W = \left(\frac{\hat{\theta} - \frac{1}{2}}{\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}} \right)^2. \quad (\star)$$

Sous H_0, $\xi^W \to_\ell \chi^2(1)$, et la région critique (à 5 %) est de la forme (à une approximation près) :

$$W = \{\xi^W \geq 4\} = \{(\star) \geq 2\}$$

Pour le test du score, la statistique de test est de la forme :

$$\xi^S = \frac{1}{n} \frac{\partial}{\partial \theta'} L_n(\hat{\theta}^o) \hat{I}^{-1}(\hat{\theta}^o) \frac{\partial}{\partial \theta} L_n(\hat{\theta}^o).$$

Ici θ' signifie que l'on prend le vecteur transposé, pour d'évidentes contraintes de conformité dans le produit. On travaille également avec $\hat{\theta}^o = \frac{1}{2}$ (cf. H_0 précédente). On a donc

$$\frac{\partial}{\partial \theta} L_n = \sum_{i=1}^{n} \frac{y_i - \hat{\theta}^o}{\theta^o(1-\theta^o)}.$$

La statistique recherchée devient donc

$$\begin{aligned} \xi^S &= \frac{1}{n} \left(\frac{\left(\sum_i y_i\right) - n\hat{\theta}^o}{\hat{\theta}^o(1-\hat{\theta}^o)} \right)^2 \hat{\theta}^o(1-\hat{\theta}^o) \\ &= n \frac{(\sum_i y_i - \theta^o)^2}{\theta^o(1-\theta^o)} \\ &= \left(\frac{\hat{\theta} - \frac{1}{2}}{\sqrt{\frac{1/2(1-1/2)}{n}}} \right)^2. \end{aligned}$$

Sous H_0, on a comme précédemment $\xi^S \to_\ell \chi^2(1)$. À la différence du test de Wald, dans lequel l'évaluation se fait dans le modèle général, ici on se place directement sous H_0 pour l'estimation.

Enfin, le test du rapport de vraisemblance est peut-être plus simple à formuler puisque l'on a besoin que des log vraisemblances en $\hat{\theta}$ et $\hat{\theta}^o$: $\xi^R = 2\left(L_n(\hat{\theta}) - L_n(\hat{\theta}^o)\right)$. On a donc

$$\begin{aligned}\xi^R &= 2\left[\sum_i (y_i)\log\hat{\theta} + \big(n - \sum_i y_i\big)\log(1-\hat{\theta})\right. \\ &\qquad \left. - \big(\sum_i y_i\big)\log\frac{1}{2} + \big(n - \sum_i y_i\big)\log\big(1-\frac{1}{2}\big)\right] \\ &= 2\left[\big(\sum_i y_i\big)\log\left(\frac{\hat{\theta}}{1/2}\right) + \big(n - \sum_i y_i\big)\log\left(\frac{1-\hat{\theta}}{1-1/2}\right)\right].\end{aligned}$$

On a les mêmes propriétés de convergence asymptotique vers la loi du χ^2.

Une simple application numérique donne les résultats suivants :

$$\xi^W = 1.01, \quad \xi^S = 1.00, \quad \xi^R = 1.00.$$

Dans tous les cas, on ne rejette pas H_0.

3 La méthode Expectation-Maximization

Dans la plupart des cas exposés plus haut, les estimateurs MV des paramètres étudiés pouvaient être dérivés à partir d'une formule analytique. De même, on pouvait généralement montrer directement qu'il n'existait qu'un unique maximum de la fonction de vraisemblance considérée sur l'espace des paramètres. Cependant, dans de nombreux problèmes d'analyse de données, l'application du principe de maximum de vraisemblance conduit à des problèmes numériques d'une grande complexité. De plus, la fonction de vraisemblance étudiée possède souvent plusieurs extrêmums. Par conséquent, dans de nombreux cas, les EMV sont calculés grâce à des techniques d'optimisation numérique, statique, dynamique ou mêlant les deux approches.

Un cas particulièrement remarquable de calcul récursif des EMV est la méthode appelée *Expectation-Maximisation* (EM) (Dempster et al., 1977 and McLachan and Krishnan, 1997). Cette approche est privilégiée lorsque la difficulté pour obtenir des EMV provient de la présence de valeurs manquantes (encore appelées variables cachées ou latentes). Si les variables manquantes avaient été observées, l'estimation MV en aurait été largement simplifiée. Dans ce contexte, la méthode EM opère de manière récursive. Chaque récursion consiste en une étape E dans laquelle on calcule l'espérance conditionnelle par rapport aux données inconnues, étant données les variables observées, et une étape M dans laquelle on maximise par rapport aux paramètres. La construction de l'algorithme est telle que l'on peut garantir qu'à chaque itération la valeur de la fonction de vraisemblance augmente. En raison de sa simplicité et de sa robustesse, la méthode EM est très largement employée, et bien qu'elle converge de manière relativement lente, de nouvelles améliorations sont publiées régulièrement dans la littérature spécialisée.

3.1 Construction de l'algorithme

Le principe de l'algoritme EM repose sur une inégalité portant sur l'espérance conditionnelle de la log-vraisemblance de variables manquantes. Ci-après, on montre deux méthodes permettant d'établir cette inégalité : à partir de l'inégalité de Jensen, et en utilisant la mesure de distance de Kullback-Leibler. Avant cela, nous rappelons quelques éléments nécesaires à la compréhension de l'algorithme EM.

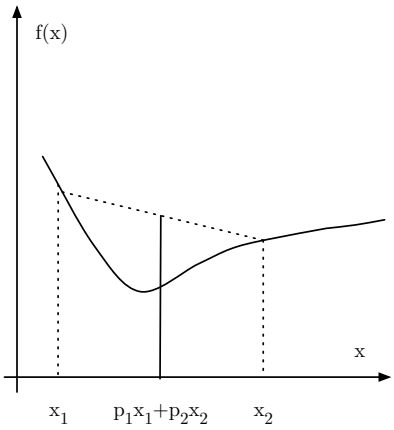


Figure 4

Illustration de la convexité de $f(x)$. La fonction $f(x)$ est dite convexe si, lorsque $p_1 \geq 0$, $p_2 \geq 0$ et $p_1 + p_2 = 1$, $f(p_1x_1 + p_2x_2) \leq p_1f(x_1) + p_2f(x_2)$.

Inégalité de Jensen

La définition de la convexité d'une fonction $g(x)$, comme illustré dans la **figure 4**, est la suivante :

$$g(p_1x_1 + p_2x_2) \leq p_1g(x_1) + p_2g(x_2), \quad p_1 \geq 0,\ p_2 \geq 0,\ p_1 + p_2 = 1.$$

Par induction, on peut montrer que cela implique une inégalité analogue pour tout $n \geq 2$,

$$g(p_1x_1 + p_2x_2 + \ldots + p_nx_n) \leq p_1g(x_1) + p_2g(x_2) + \ldots + p_ng(x_n), \tag{25}$$

$p_i \geq 0$, $i = 1, 2, \ldots, n$, $p_1 + p_2 + \ldots + p_n = 1$. On peut également passer d'un espace de paramètres $x \in \mathbb{R}$ à une dimension à un espace m-dimensionnel plus général, $x \in \mathbb{R}^m$, et cette inégalité reste vérifiée. Toute fonction convexe $g(x)$, $\mathbb{R}^m \to \mathbb{R}$ satisfait **25**, et l'**inégalité 25** est appelée l'*inégalité finie et discrète de Jensen*. On notera que l'on peut autoriser $n \to \infty$ et **25** est toujours vérifiée.

On peut également remplacer la distribution discrète de probabilité contenant les masses $p_1, p_2, \ldots, p_n$ apparaissant dans l'**expression 25** par une distribution continue $f(x)$, où $\int_{-\infty}^{+\infty} f(x)dx = 1$, et on a une inégalité analogue :

$$g\left(\int_{-\infty}^{+\infty} xf(x)\right) \leq \int_{-\infty}^{+\infty} g(x)f(x)dx,$$

qui peut également s'exprimer, l'aide de l'opérateur espérance, comme

$$g\big(\mathbb{E}(X)\big) \leq \mathbb{E}\big(g(X)\big).$$

Dans l'expression ci-dessus, X désigne une v.a. dont la densité de probabilité est $f(x)$. Les deux ingéalités précédentes restent valables pour toute fonction convexe $g(x)$, et on parle de l'*inégalité de Jensen*. Plus généralement, celle-ci peut être formulée comme suit :

$$g\left(\int_{-\infty}^{+\infty} h(x)f(x)\right) \leq \int_{-\infty}^{+\infty} g\big(h(x)\big)f(x)dx,$$

ou

$$g\left(\mathbb{E}\big(h(X)\big)\right) \leq \mathbb{E}\left(g\big(h(X)\big)\right),$$

où $g(x)$ est convexe et $h(x)$ désigne n'importe quelle fonction mesurable. Si l'on utilise la transformation $Y = h(X)$, on retrouve naturellement les deux expressions précédentes.

Distance de Kullback-Leibler

Considérons deux v.a. finies discrètes X et Y, chacune prenant les valeurs $1, 2, \ldots, n$ avec probabilités $p_1, p_2, \ldots, p_n$, $p_1+p_2+\ldots+p_n = 1$, pour X et $q_1, q_2, \ldots, q_n$, $q_1+q_2+\ldots+q_n = 1$, pour Y. La distance de Kullback-Leibler $K_{X,Y}$ entre les distributions de X et de Y se définit comme

$$K_{X,Y} = -\sum_{i=1}^{n} q_i \ln \frac{p_i}{q_i}.$$

On peut vérifier que $K_{X,Y} \geq 0$ et que

$$K_{X,Y} = 0 \Leftrightarrow p_i = q_i, \quad i = 1, 2, \ldots, n.$$

La distance de Kullback-Leibler est encore appelée *entropie* de la distribution $p_1, p_2, \ldots, p_n$ par rapport à celle de $q_1, q_2, \ldots, q_n$.

Pour des v.a. continues X et Y, de densités de probabilités $f_X(z)$ et $f_Y(z)$, leur distance de Kullback-Leibler est définie comme

$$K_{X,Y} = -\int_{-\infty}^{+\infty} f_Y(z) \ln \frac{f_X(z)}{f_Y(z)} dz,$$

et on a toujours

$$K_{X,Y} \geq 0 \tag{26}$$

et $K_{X,Y} = 0 \Leftrightarrow f_X(z) = f_Y(z)$ (à l'exception éventuellement d'un ensemble de mesures nulles).

Itérations EM

On supposera dans les paragraphes qui suivent que les observations disponibles peuvent être modélisées par une v.a. (ou un vecteur aléatoire) X et que l'objectif est d'estimer

un paramètre (ou un vecteur de paramètres) p. Par ailleurs, on supposera qu'il existe un certain nombre de valeurs manquantes X^m. En aggrégeant les deux séries de données (observées et manquantes), on obtient un vecteur d'*observations complètes*

$$X^c = (X^m, X).$$

On va essayer de montrer comment l'estimation de p à partir de la fonction de log-vraisemblance pour une observation x,

$$\ln\big(f(x,p)\big),$$

entraîne des problèmes computationnels alors que la maximisation de la log-vraisemblance des observations complètes, $\ln\big(f(x^c,p)\big)$, est relativement directe.

Dans un premier temps, on cherche à exprimer la distribution conditionnelle des observations manquantes étant donnés les observations disponibles et les paramètres d'intérêt, $f(x^m \mid x,p)$, à partir de la formule de Bayes :

$$f(x^m \mid x,p) = \frac{f(x^m,x,p)}{f(x,p)} = \frac{f(x^c,p)}{f(x,p)}.$$

Par simple substitution dans l'expression ci-dessus, on obtient

$$f(x,p) = \frac{f(x^c,p)}{f(x^m \mid x,p)}.$$

En passant au logarithme des deux côtés de l'équation, on a alors

$$\ln f(x,p) = \ln f(x^c,p) - \ln f(x^m \mid x,p). \tag{27}$$

On a besoin de fournir une première estimation *a priori* pour les paramètres, que l'on notera p^{old}, et on rappelle que x est connu et fixé. La distribution de x^m (inconnu) étant donné les observations disponibles x est $f(x^m \mid x,p^{old})$. On moyenne **27** sur la distribution des données inconnues, ou en d'autres termes, on calcule l'espérance des deux membres de l'**équation 27** par rapport à $f(x^m \mid x,p^{old})$. Puisque $\mathbb{E}\big(h(X) \mid X\big) = h(X)$ quelle que soit $h(X)$, on peut écrire

$$\ln f(x,p) = \mathbb{E}\big(\ln f(X^c,p) \mid x,p^{old}\big) - \mathbb{E}\big(\ln f(X^m,p) \mid x,p^{old}\big).$$

Si l'on introduuit la notation suivante :

$$Q(p,p^{old}) = \mathbb{E}\big(\ln f(X^c,p) \mid x,p^{old}\big) = \int f(x^m \mid x,p^{old}) \ln f(x^c,p) dx^m \tag{28}$$

et

$$H(p,p^{old}) = \mathbb{E}\big(\ln f(X^m,p) \mid x,p^{old}\big) = \int f(x^m \mid x,p^{old}) \ln f(x^m \mid x,p) dx^m,$$

on a alors

$$\ln f(x,p) = Q(p,p^{old}) - H(p,p^{old}). \tag{29}$$

On en déduit que

$$H(p^{old}, p^{old}) - H(p, p^{old}) = -\int f(x^m \mid x, p^{old}) \ln \frac{f(x^m \mid x, p)}{f(x^m \mid x, p^{old})} dx^m.$$

On peut appliquer l'inégalité de Jensen au membre droit de l'**équation 29**, en prenant comme fonction convexe $g(x^m) = -\ln(x^m)$ ainsi que $h(x^m) = f(x^m \mid x, p)/f(x^m \mid x, p^{old})$ ou l'**inégalité 26** pour la distance de Kullback-Leibler. Dans les deux cas, on arrive à la conclusion que

$$H(p^{old}, p^{old}) - H(p, p^{old}) \geq 0. \tag{30}$$

Si l'on est capable un nouvel estimé p^{new} qui vérifie $Q(p^{new}, p^{old}) > Q(p^{old}, p^{old})$, alors de **29** et **30** on peut conclure que

$$\ln f(x, p^{new}) > \ln f(x, p^{old}),$$

et donc on a réussi à augmenter la log-vraisemblance. Typiquement, p^{new} sera choisi en maximisant $Q(p, p^{old})$ par rapport à p.

En résumant l'ensemble de la démarche exposée ci-dessus, on en arrive à définir la construction de l'algorithme EM comme suit :

- **Étape E.** Calculer $Q(p, p^{old})$ comme défini en **28**.
- **Étape M.** Calculer $p^{new} = \arg\max_p Q(p, p^{old})$.

En répétant les étapes E et M, en mettant à jour à chaque fois $p^{old} = p^{new}$, on augmente, itérativement, la valeur de la log-vraisemblance $\ln f(x, p^{old})$. Dans la plupart des cas, cette approche itérative donne un maximum global unique. Toutefois, les itérations EM peuvent également se terminer sur des maxima locaux, voire ne pas converger du tout.

Avant de fournir des exemples d'applications concrets de l'algorithme sur des distributions simples ou un peu plus complexes, une illustration de l'algorithme EM appliqué à un lancer de pièces est proposée dans la **figure 5**. L'exemple est tiré de REF.

3.2 Exemples d'application de l'algorithme EM

Les exemples suivants ont pour but d'illustrer le principe de base de l'algorithme EM exposé à la section précédente, ainsi que d'étudier sa convergence.

Distribution exponentielle avec données censurées

Les données censurées se rencontrent fréquemment en épidémiologie, et plus particulièrement dans les études de survie (Cox and Oakes, 1984). On les retrouve également avec des instruments de mesure pour lesquels la gamme de mesure observables est trop limitée. Ici, on considèrera une v.a. T exponentielle. L'objectif est d'estimer le paramètre a à

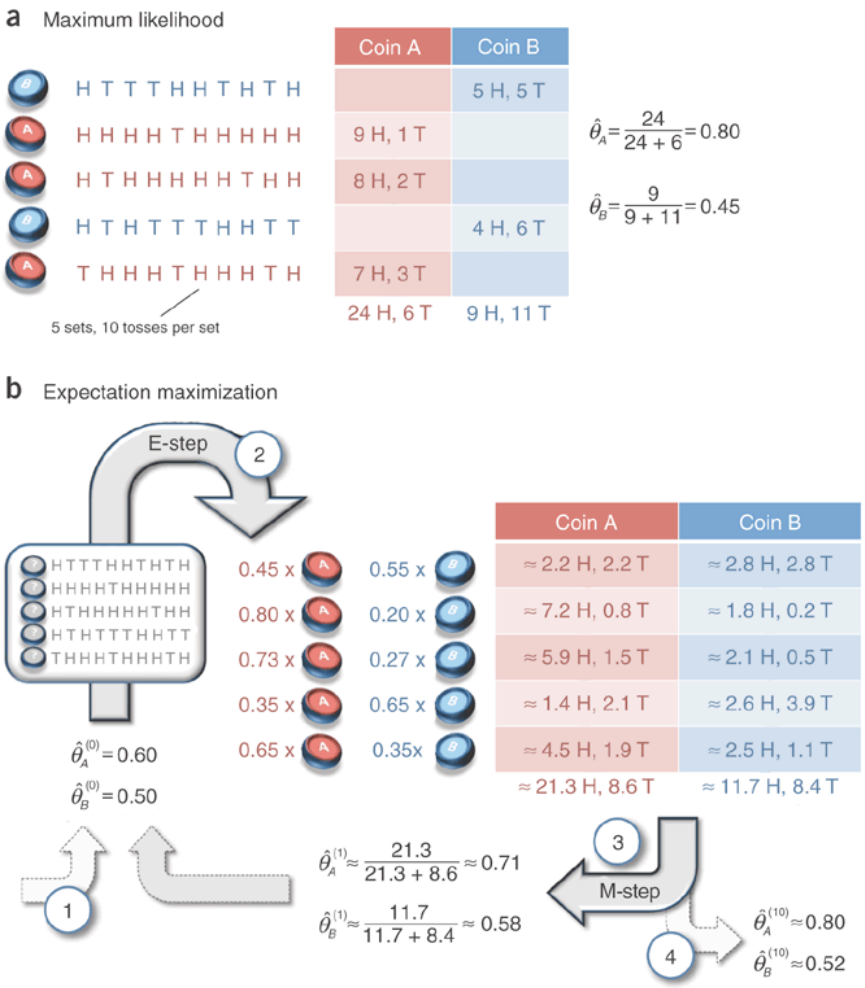


Figure 5 Principe de l'algorithme EM. Tiré de ? ?

partir d'un ensemble de N observations, en tenant compte du fait qu'il existe un mécanisme de censure de seuil constant C : si une mesure T est plus grande que C alors on ne connaît pas sa vraie valeur, mais on sait seulement que le seuil a été dépassé. Supposons que les observations $t_1, t_2, \ldots, t_k$ n'ont pas excédé le seuil et que $t_{k+1}, \ldots, t_N$ sont au-dessus du seuil. Les informations disponibles sont donc $t_1, t_2, \ldots, t_k$ et $[t_{k+1}, \ldots, t_N \geq C]$. L'information complète est constituée par le vecteur $t^c = t_1, t_2, \ldots, t_k, t_{k+1}, \ldots, t_N$. Afin d'initier l'algorithme EM, on démarre avec des valeurs *a priori* pour le paramètre, a^{old}. L'expression de $Q(a, a^{old})$ avec $f(t, a)$ définie en 14 devient alors

$$
\begin{aligned}
Q(a, a^{old}) &= \mathbb{E}\left(\ln f(T^c, a) \mid t_1, t_2, \ldots, t_k, [t_{k+1}, \ldots, t_N \geq C], a^{old}\right) \\
&= \sum_{i=1}^{k} \ln\left(a \exp(-a t_i)\right) + \sum_{i=k+1}^{N} \mathbb{E}\left(\ln\left(a \exp(-a t_i)\right) \mid t_i \geq C, a^{old}\right)
\end{aligned}
$$

$$= N \ln a - a \sum_{i=1}^{k} t_i - a(N-k) \frac{\int_C^{+\infty} t a^{old} \exp(a^{old} t) dt}{\int_C^{+\infty} a^{old} \exp(a^{old} t) dt}$$
$$= N \ln a - a \left[\sum_{i=1}^{k} t_i + (N-k) \left(C + \frac{1}{a^{old}} \right) \right].$$

Dans la transformation ci-dessus, on a utilisé le fait que

$$\mathbb{E}(-a t_i \mid t_i \geq C, a^{old}) = -a \frac{\int_C^{+\infty} t a^{old} \exp(a^{old} t) dt}{\int_C^{+\infty} a^{old} \exp(a^{old} t) dt} = -a \left(C + \frac{1}{a^{old}} \right).$$

D'après la relation précédente, la valeur a^{new} qui maximise $Q(a, a^{old})$ par rapport à a est

$$a^{new} = \frac{N}{\sum_{i=1}^{k} t_i + (N-k)(C + 1/a^{old})}. \tag{31}$$

On peut indexer les itérations des estimations EM du paramètre a par une suite $1, 2, \ldots, m, \ldots$, et on peut ainsi écrire $a_m = a^{old}$ et $a_{m+1} = a^{new}$. En passant à la limite dans la **relation 31**, on obtient

$$\hat{a} = \lim_{m \to \infty} a_m = \frac{k}{\sum_{i=1}^{k} t_i + (N-k)C}.$$

La limite $\lim_{m \to \infty} a_m$ peut être calculée analytiquement et on obtient ainsi l'EMV de a. On pourrait également arriver au même résultat en écrivant directement la fonction de log-vraisemblance d'une distribution exponentielle incluant des données censurées.

Modèle de mélange

Les mélanges de lois sont souvent utilisés pour étudier la structure des données expérimentales (McLachan and Peel, 2000). Les mélanges de distributions prennent la forme :

$$f^{mix}(x, \alpha_1, \ldots, \alpha_K, p_1, \ldots, p_K) = \sum_{k=1}^{K} \alpha_k f_k(x, p_k), \tag{32}$$

où $\alpha_1, \ldots, \alpha_K, p_1, \ldots, p_K$ sont les paramètres de la loi composée. Les poids (probabilités) $\alpha_1, \ldots, \alpha_K$ sont non négatifs et somment à 1, i.e.

$$\sum_{k=1}^{K} \alpha_k = 1, \tag{33}$$

et les $f_k(x, p_k)$ sont des fonctions de densité de probabilités. On dira qu'une v.a. X suit ce type de distribution si elle est générée de la manière suivante :

1. on génère un nombre entier k dans l'intervalle $1, \ldots, K$ avec probabilité $\alpha_1, \ldots, \alpha_K$,
2. on génère un nombre (ou un vecteur) x à partir de la distribution $f_k(x, p_k)$.

La plupart du temps, les $f_k(x, p_k)$ sont des distributions du même type, par exemple gaussienne ou poissonienne, avec des paramètres différents, mais il est également possible de mélanger des distributions de différent type. On appelle $f_k(x, p_k)$, $k = 1, 2, \ldots, K$ la composante de la distribution de mélange.

Supposons qu'un échantillon aléatoire de taille N soit tiré de la distribution **32**. Le calcul des EMV des paramètres $\alpha_1, \ldots, \alpha_K, p_1, \ldots, p_K$ pose typiquement des problèmes d'optimisation numérique. Toutefois, l'algorithme EM fournit une approche naturelle à ce problème. En effet, on suppose que l'information complète est donnée par $x^c = k_1, k_2, \ldots, k_N, x_1, x_2, \ldots, x_N$; en d'autres termes, on fait l'hypothèse que l'on connaît l'indice k_n de la composante $f_k(x, p_k)$ qui a permis de générer l'observation x_n. Avec cette information complète, le problème de l'estimation par MV peut être divisé en différents sous-problèmes :

a. estimation des paramètres $p_1, \ldots, p_M$ des composantes de distribution,
b. estimation par MV des poids $\alpha_1, \ldots, \alpha_K$.

La dernière étape de calcul peut être effectuée à partir des indices k_n. Grâce à cette décomposition, la log-vraisemblance des données complètes est de la forme :

$$\ln\big(f(x^c, p)\big) = \sum_{n=1}^{N} \ln \alpha_{k_n} + \sum_{n=1}^{N} \ln f_{k_n}(x_n, p_{k_n}),$$

avec $x^c = k_1, k_2, \ldots, k_N, x_1, x_2, \ldots, x_N$ et $p = \alpha_1, \ldots, \alpha_K, p_1, \ldots, p_K$.

Étape E. On postule des valeurs *a priori* pour les paramètres $p^{old} = \alpha_1^{old}, \ldots, \alpha_K^{old}$, $p_1^{old}, \ldots, p_K^{old}$ et on écrit l'expression de $Q(p, p^{old})$ en considérant l'information disponible $x = x_1, x_2, \ldots, x_N$ et l'information manquante $x^m = k_1, k_2, \ldots, k_N$:

$$\begin{aligned}
Q(p, p^{old}) &= \mathbb{E}\big(\ln f(X^c, p) \mid x, p^{old}\big) \\
&= \mathbb{E}\left(\sum_{n=1}^{N} \ln \alpha_{k_n} \mid x, p^{old}\right) + \mathbb{E}\left[\sum_{n=1}^{N} \ln f_{k_n}(x_n, p_{k_n}) \mid x, p^{old}\right] \\
&= \sum_{n=1}^{N} \mathbb{E}(\ln \alpha_{k_n} \mid x, p^{old}) + \sum_{n=1}^{N} \mathbb{E}\big(\ln f_{k_n}(x_n, p_{k_n}) \mid x, p^{old}\big) \\
&= \sum_{n=1}^{N} \sum_{k=1}^{K} p(k \mid x_n, p^{old}) \ln \alpha_{k_n} + \sum_{n=1}^{N} \sum_{k=1}^{K} p(k \mid x_n, p^{old}) \ln f_k(x_n, p_k) \qquad (34)
\end{aligned}$$

.

La distribution $p(k \mid x_n, p^{old})$ des données manquantes conditionnellement aux données observées et les paramètres spécifiés *a priori* sont obtenus en utilisant la formule de Bayes :

$$p(k \mid x_n, p^{old}) = \frac{\alpha_k^{old} f_k(x_n, p^{old})}{\sum_{\kappa=1}^{K} \alpha_\kappa^{old} f_\kappa(x_n, p^{old})}. \qquad (35)$$

Étape M. L'expression de $Q(p, p^{old})$ peut être directement optimisée par rapport aux poids $\alpha_1, \ldots, \alpha_K$. En considérant la contrainte **33** et en utilisant un schéma similare à celui exposé pour la distribution multinomiale, on obtient ainsi :

$$\alpha_k^{new} = \frac{\sum_{n=1}^{N} p(k \mid x_n, p^{old})}{N}.$$

Les itérations ci-dessus pour les poids restent valides quelle que soit la forme des composantes de la distribution de mélange. Afin de dériver les itérations EM pour les estimations des paramètres des composantes de la distribution $p_1^{new}, \ldots, p_K^{new}$, on peut se pencher sur deux cas particuliers.

Distribution de mélange de Poisson

Supposons que la $k$ième composante de la distribution de mélange dans la $n$ième expérience, $f_k(x_n, p_k)$, soit une distribution de Poisson avec un paramètre d'intensité $p_k = \lambda_k$:

$$f_k(x_n, \lambda_k) = \exp(-\lambda_k)\frac{\lambda_k^{x_n}}{x_n!}. \tag{36}$$

À présent, $p(k \mid x_n, p^{old})$ est donné par **35** en remplaçant la fonction de répartition $f_k(x_n, p^{old})$ par une distribution de Poisson avec comme paramètre fourni *a priori* λ_k^{old}, $k = 1, 2, \ldots, K$:

$$p(k \mid x_n, \lambda^{old}) = \frac{\alpha_k^{old} \exp(-\lambda_k^{old})(\lambda_k^{old})^{x_n}}{\sum_{\kappa=1}^{K}\left[\alpha_\kappa^{old} \exp(-\lambda_\kappa^{old})(\lambda_\kappa^{old})^{x_n}\right]}$$

Dans l'expression ci-dessus, $\lambda^{old} = \lambda_1^{old}, \ldots, \lambda_K^{old}$. En remplaçant **36** dans **34** et en maximisant par rapport à λ_k, on obtient la valeur mise à jour λ_k^{new} :

$$\lambda_k^{new} = \frac{\sum_{n=1}^{N} x_n p(k \mid x_n, \lambda^{old})}{\sum_{n=1}^{N} p(k \mid x_n, \lambda^{old})}, \quad k = 1, 2, \ldots, K.$$

Distribution de mélange gaussienne

Ici, toutes les composantes de distribution sont gaussienne de paramètres $\mu_k, \sigma_k, k = 1, 2, \ldots, K$. Pour la $n$ième observation, on a

$$f_k(x_n, \mu_k, \sigma_k) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x_n - \mu_k)^2}{2\sigma_k^2}\right]. \tag{37}$$

En supposant initialement $\mu_k^{old}, \sigma_k^{old}, k = 1, 2, \ldots, K$, l'expression pour les données manquantes conditionnellement aux données observées et aux paramètres initiaux est de la forme :

$$p(k \mid x_n, p^{old}) = \frac{\alpha_k^{old} \exp\left[-(x_n - \mu_k^{old})^2/\left(2(\sigma_k^{old})^2\right)\right]}{\sum_{\kappa=1}^{K} \alpha_\kappa^{old} \exp\left[-(x_n - \mu_\kappa^{old})^2/\left(2(\sigma_\kappa^{old})^2\right)\right]}.$$

Dans l'expression ci-dessus, on a utilisé les notations $p^{old} = \alpha_1^{old}, \ldots, \alpha_K^{old}, \mu_1^{old}, \ldots, \mu_K^{old}, \sigma_1^{old}, \ldots, \sigma_K^{old}$ pour désigner le vecteur composé de l'ensemble des paramètres estimés. Lorsque **37** est substitué dans **34**, la maximisation par rapport à μ_k, σ_k donne la règle de mise à jour suivante pour la moyenne et le paramètre de dispersion :

$$\mu_k^{new} = \frac{\sum_{n=1}^{N} x_n p(k \mid x_n, p^{old})}{\sum_{n=1}^{N} p(k \mid x_n, p^{old})}, \quad k = 1, 2, \ldots, K,$$

et

$$(\sigma_k^{new})^2 = \frac{\sum_{n=1}^{N} (x_n - \mu_k^{new})^2 p(k \mid x_n, p^{old})}{\sum_{n=1}^{N} p(k \mid x_n, p^{old})}, \quad k = 1, 2, \ldots, K.$$

4 Tests statistiques

5 Chaînes de Markov

Avant de présenter le formalisme des chaînes de Markov, on rappellera brièvement quelques notions sur les processus stochastiques.

Un processus stochastique est en fait une famille de fonctions d'une variable t, $\{X(t, \omega),\ t \in T,\ \omega \in \Omega\}$ (t sera généralement assimilé au temps), paramétrée par des réalisations aléatoires ω. Quel que soit ω, $X(\cdot, \omega)$ est une fonction ; quel que soit un instant fixé t, $X(t, \cdot)$ est une variable aléatoire.

Les processus de Markov constituent l'une des classes les plus utilisées et les mieux connues des processus stochastiques (Feller, 1968, Iosifescu, 1980 and Gikhman and Skorokhod, 1996). Un processus de Markov est un cas particulier d'un processus stochastique dans la mesure où il s'agit d'un processus à mémoire limitée. Ceci signifie que pour un processus $X(t, \omega)$ qui s'est déroulé dans le passé ($t \leq t_0$), le futur $\{X(t, \omega),\ t > t_0\}$ est caractérisé par le présent uniquement, i.e. $X(t_0, \omega)$. Cette propriété est connue sous le nom de *propriété de Markov*.

Une chaîne de Markov est un processus markovien pour lequel $X(t, \omega) \in S$, où S est un ensemble discret. Habituellement, l'espace d'état S est un sous-ensemble de $\mathbb{N}$. En d'autres termes, une chaîne de Markov (CM) présente des transitions aléatoires entre différents états discrets. La théorie présentée dans cette section se concentre sur le cas d'un nombre de fini d'états N, indexés $1, 2, \ldots, N$. De même, nous n'aborderons que le cas des intervalles de temps discrets, $0, 1, 2, \ldots, k, \ldots$. Toutefois, on évoquera quelques généralités sur le cas continu. La plupart du temps, on notera la CM $X(t, \omega)$ $X_k(\omega)$ ou simplement X_k.

Comme on vient de l'énoncer, la propriété fondamentale d'une CM est que les états futurs ne sont déterminés que par le présent, X_k, ce que l'on peut exprimer de la manière suivante :

$$P(X_{k+1} = j \mid X_k = i, X_{k-1} = i_1, X_{k-2} = i_2, \ldots) = P(X_{k+1} = j \mid X_k = i). \tag{38}$$

La probabilité conditionnelle $P(X_{k+1} = j \mid X_k = i)$ est appelée la *probabilité de transition* de $X_k = i$ à $X_{k+1} = j$, et on la dénote p_{ij}, avec donc

$$p_{ij} = P(X_{k+1} = j \mid X_k = i). \tag{39}$$

Une propriété importante des CM considérée ici est l'homogénéité dans le temps, qui signifie que les probabilités de transition p_{ij} ne dépendent pas du temps.

La **propriété 38** entraîne les conséquences les plus fondamentales pour l'analyse des CM et permet de dériver des relations de récurrence pour les probabilités liées à X_k. En particulier, la probabilité d'observer la séquence d'états $i_0, i_1, \ldots, i_K$ est donnée par le produit des probabilités de transition

$$P(i_0, i_1, \ldots, i_K) = \pi_{i_0} p_{i_0 i_1} \cdots p_{i_{K-1} i_K}, \tag{40}$$

où $\pi_{i_0} = P(X_0 = i_0)$. L'équation ci-dessus peut être retrouvée en utilisant la règle de la chaîne et la **propriété 38**.

5.1 Matrice des probabilités de transition et graphe des transitions d'état

Les probabilités de transition p_{ij} données en **39** peuvent se représenter sous la forme d'une matrice $N \times N$, notée P, que l'on appelle la matrice des probabilités de transition de la chaîne considérée :

$$P = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1N} \\ p_{21} & p_{22} & \cdots & p_{2N} \\ \cdots & \cdots & \cdots & \cdots \\ p_{N1} & p_{N2} & \cdots & p_{NN} \end{bmatrix}.$$

Les transitions d'état et leurs probabilités associées peuvent également se représenter sous la forme d'un graphe de transition d'états, comme celui illustré dans la **figure 6**. Dans ce schéma, les cercles représentent les états et les arcs représentent les transitions d'état. Chaque arc est associé à une probabilité de transition, et un arc n'est représenté que si $p_{ij} \neq 0$. Les représentations graphiques et matricielles sont strictement équivalentes. La matrice des probabilités de transition illustré dans la **figure 6** est :

$$P = \begin{bmatrix} 0.5 & 0.5 & 0 & 0 \\ 0 & 0 & 0.8 & 0.2 \\ 0 & 0 & 0.1 & 0.9 \\ 0 & 0 & 0 & 1 \end{bmatrix}. \tag{41}$$

On pourra vérifier que les probabilités de transition de l'état i à tous les autres états somment à 1, i.e.

$$\sum_{j=1}^{N} p_{ij} = 1. \tag{42}$$

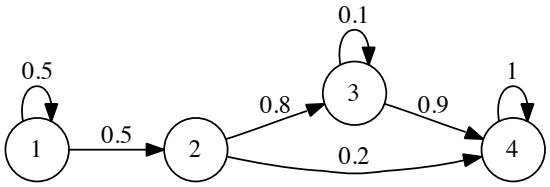


Figure 6 Graphe de transitions d'états pour la chaîne de Markov décrite en 41.

Une matrice P qui possède la **propriété 42** est appelée une matrice stochastique. La propriété du graphe de transition correspondant est alors : « Les poids associés aux arcs de transition sortant d'un état i somment à 1 ».

Il arrive fréquemment que la matrice des probabilités de transition soit une matrice creuse, dans laquelle de nombreuses transitions possèdent des probabilités nulles, auquel cas le graphe devient une représentation plus économique et surtout plus lisible.

5.2 Évolution temporelle des distributions de probabilités d'états

Une fois que l'on a spécifié la matrice ou le graphe des probabilités de transition ainsi que la distribution initiale des états, il est possible de calculer l'évolution de cette distribution de probabilité avec le temps. Si l'on considère qu'au temps 0, la distribution de probabilité des états est

$$P(X_0 = i) = \pi_i(0),$$

on a bien $\sum_{i=1}^{N} \pi_i(0) = 1$. En utilisant la loi des probabilités totales (1), on peut calculer la distribution de probabilité des états au temps suivant :

$$P(X_1 = j) = \pi_j(1) = \sum_{i=1}^{N} \pi_i(0) p_{ij}. \tag{43}$$

Si l'on introduit une notation vectorielle pour les probabilités de transition à l'état k,

$$\pi(k) = [\pi_1(k), \pi_2(k), \ldots, \pi_N(k)],$$

on peut alors représenter 43 à l'aide du produit matriciel par

$$\pi(1) = \pi(0)P. \tag{44}$$

En appliquant récursivement 44, on a finalement

$$\pi(k) = \pi(0)P^k.$$

5.3 Classification des états

La classification des états d'une CM et plus généralement des CM est importante pour bien comprendre la théorie et les applications des CM. Nous présentons ci-dessous cette classification, illustrée par quelques propriétés des graphes de transition d'état.

Irréductibilité

Une CM est dite irréductible si et seulement si le graphe des transitions d'états possède la propriété que tous les états peuvent être atteints depuis n'importe quel état. La CM dont le graphe associé est représenté dans la **figure 7** (gauche) est en ce sens irréductible. Si une CM n'est pas irréductible, comme c'est le cas dans la **figure 7** (droite), alors en renumérotant ses états, sa matrice de probabilités de transition peut être réarrangée sous la forme d'une matrice en blocs

$$P = \begin{bmatrix} Q & 0 \\ U & V \end{bmatrix},$$

où le bloc supérieur droit ne contient que des 0 et Q est une matrice carrée correspondant à une sous-chaîne de Markov irréductible.

La matrice de probabilités de transition P d'une CM irréductible possède la propriété que $P^k > 0$ (toutes les entrées sont strictement positives) pour un k.

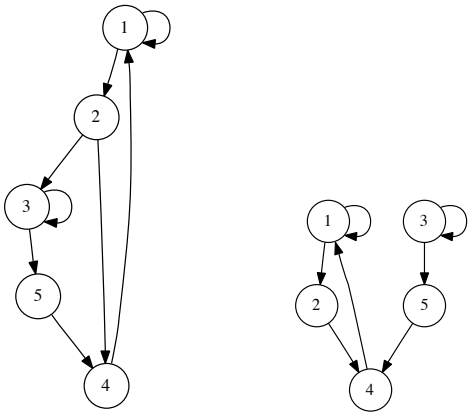


Figure 7 *Gauche.* un graphe des transitions d'état d'une chaîne de Markov irréductible. *Droite.* La chaîne de Markov représentée par ce graphe n'est pas irréductible. Dans les deux graphes, les arcs représentent des transitions avec $p_{ij} \neq 0$.

États persistents et transients

Un état i est persistent si une CM partant de l'état i retourne à ce même état avec la probabilité 1. En d'autres termes, parmi la séquence infinie des états d'une CM démarrant en i, l'état i apparaît un nombre infini de fois. Un état qui n'est pas persistent est dit transient. Il n'apparaît qu'un nombre fini de fois. Dans la CM illustrée dans la **figure 7** (gauche), tous les états sont persistents, alors que pour celle illustrée dans la même figure à droite, les états 3 et 5 sont transients et les états 1, 2 et 4 sont persistents.

On définit

$$f_i^{(k)} = \Pr(\text{La chaîne débutant en } i \text{ retourne en } i \text{ pour la première fois après } k \text{ étapes}),$$

avec, par convention, $f_i^{(0)} = 0$, et

$$f_i = \sum_{k=1}^{\infty} f_i^{(k)}.$$

Puisque les événements $f_i^{(k)}$ sont exclusifs, la somme de leurs probabilités ne peut excéder 1, i.e. $f_i \leq 1$. À partir de f_i, on peut donner une autre condition pour la caractérisation des états transients et persistents : un état i est transient si $f_i < 1$ et persistent si $f_i = 1$.

Les probabilités f_i peuvent être calculées à partir des entrées des matrices $P, P^2, \ldots, P^k, \ldots$. On définit

$$p_{ii}^{(k)} = \Pr(\text{La chaîne débutant en } i \text{ retourne en } i \text{ après } k \text{ étapes})$$

et on fixe par convention $p_{ii}^{(0)} = 1$. Les événements ci-dessus ne sont pas exclusifs. On peut voir également que $p_{ii}^{(k)}$ est la (i,i)ème entrée de la matrice P^k. À partir de la loi des probabilités totales **1**, on a

$$p_{ii}^{(k)} = f_i^{(1)} p_{ii}^{(k-1)} + f_i^{(2)} p_{ii}^{(k-2)} + \ldots + f_i^{(k)} p_{ii}^{(0)}.$$

En écrivant l'expression ci-dessus pour $k = 1, 2, \ldots$, on obtient un système d'équations linéaires qui permet de trouver $f_i^{(k)}$.

En utilisant les probabilités $p_{ii}^{(k)}$, on peut donc ajouter une condition supplémentaire. Si $\sum_{k=0}^{\infty} p_{ii}^{(k)} < \infty$, alors l'état i est transient. Cette dichotomie peut se démontrer en utilisant la méthode des fonctions génératrices à l'équation précédente.

Si l'état i est persistent, on peut se demander quel est le temps d'attente espéré μ_i pour la récurrence de i. À partir de la définition de $f_i^{(k)}$, μ_i peut être estimé comme

$$\mu_i = \sum_{k=1}^{\infty} k f_i^{(k)}.$$

États périodiques

Dans la figure 8, on peut voir un exemple d'un graphe de transition correspondant à des états péridioques. Les états 1, 2 et 3 sont périodiques de période 3. Généralement, un état i d'une CM est périodique si $p_{ii}^{(k)} \neq 0$ seuement pour $k = \nu t$, $t = 0, 1, \ldots$ et $\nu > 1$ entier. Le plus grand ν vérifiant la relation précédente est appelé la période de l'état i. On rencontre rarement des phénomènes de périodicité dans les applications des CM. Il s'agit plutôt d'une possibilité théorique, dont on doit tenr compte dans les définitions et les démonstrations de théorème. Un état i est apériodique si aucun $\nu > 1$ ne peut être trouvé.

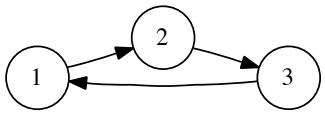


Figure 8 Graphe de transitions d'états correspondant à des états périodiques.

5.4 Ergodicité

Un état i est dit ergodique s'il est apériodique et persistent. Une CM est dite ergodique si tous ses états sont ergodiques. Pour les CM avec un nombre fini d'états, l'ergodicité est induite par l'irréductibilité et l'apériodicité.

5.5 Distribution stationnaire

La distribution stationnaire (ou *invariante*) d'une CM est définie comme les π_S (un vecteur ligne) tels que

$$\pi_S = \pi_S P,$$

lorsqu'ils existent. En général, les π_S ne sont pas nécessairement uniques. Par exemple, si

$$P = \begin{bmatrix} P_1 & 0 \\ 0 & P_2 \end{bmatrix}$$

et $\pi_{S1} = \pi_{S1} P_1$, $\pi_{S2} = \pi_{S2} P_2$, alors pour tout $\alpha \in [0, 1]$, $\pi_S = \alpha[0\ \pi_{S1}] + (1-\alpha)[\pi_{S2}\ 0]$ est une distribution stationnaire. Les distributions stationnaires sont liées aux distributions limites, définies par $\pi(\infty) = \lim_{k\to\infty} \pi(k)$. Si la distribution existe à la limite, alors elle est stationnaire. Si une CM est ergodique, alors la limite de $\pi(k)$ existe et ne dépend pas de la distribution initiale $\pi(0)$, i.e.

$$\lim_{k\to\infty} \pi(k) = \pi_S.$$

Dans ce cas, la distribution stationnaire est unique. De plus, la limite de P^k existe également et

$$\lim_{k\to\infty} P^k = \mathbb{I}\pi_S. \tag{45}$$

Dans l'expression ci-dessus, $\mathbb{I}$ est un vecteur colonne de taille N ne contenant que des 1. La $i$ème colonne de la matrice limite définie en 45 consiste en éléments tous identiques et égaux à π_{Si}, le $i$ème élement du vecteur π_S. On peut également démontrer que

$$\pi_{Si} = \frac{1}{\mu_i}.$$

On appelle stationaire une CM si sa distribution initiale est une distribution stationnaire :

$$\pi(0) = \pi_S. \tag{46}$$

Dans une telle chaîne, par définition de π_S, $\pi(k) = \pi_S$ pour chaque k. En d'autres termes, la CM évolue selon sa distribution stationnaire.

5.6 Chaînes de Markov réversible

Dans cette section, on considère une CM en ordre inverse, $\{X_k, X_{k-1}, X_{k-2}, \ldots\}$. On peut montrer que le processus $X_k, X_{k-1}, X_{k-2}, \ldots$ possède lui aussi la propriété de Markov. En utilisant la règle de Bayes, on peut calculer la probabilité de transition $i \to j$ en temps inverse,

$$\begin{aligned} p_{ij}^{rev} &= P(X_{k-1} = j \mid X_k = i) \\ &= \frac{P(X_{k-1} = j)P(X_k = i \mid X_{k-1} = j)}{P(X_k = i)} = \frac{\pi_j(k-1)p_{ji}}{\pi_i(k)}. \end{aligned} \tag{47}$$

Il y a toutefois une inconsistence dans la notation ci-dessus car p_{ij}^{rev} dépend de l'instant k. Par simplicité de notation, on supprime l'indice k. On retiendra qu'une CM en temps inversé devient inhomogène.

Dans la plupart des applications, il est important d'analyser la CM inversée sous l'hypothèse additionnelle de stationnarité 46. Dans ce cas la CM en temps inversé devient homogène. On a $P(X_{k-1} = j) = \pi_{Sj}$ et $P(X_k = i) = \pi_{Si}$, et 47 devient

$$p_{ij}^{rev} = \frac{\pi_{Sj}p_{ji}}{\pi_{Si}}. \tag{48}$$

Une CM est dite inversible si elle satisfait la relation

$$p_{ij}^{rev} = p_{ij}. \tag{49}$$

Il est intéressant de remarquer que l'inversibilité implique la stationnarité de la CM directe et inversée. En effet, si

$$p_{ij} = \frac{\pi_j(k-1)p_{ji}}{\pi_i(k)}$$

pour tout i, j, alors si on pose $i = j$ on a $\pi_i(k-1)/\pi_i(k) = 1$.

De la **définition 49**, on voit que lorsque l'on observe les états d'une CM inversée, on ne peut dire si elle progresse de manière directe ou inversée. En combinant **48** et **49**, on obtient la condition suivante pour l'inversibilité d'une CM :

$$p_{ij}\pi_{Si} = \pi_{Sj}p_{ji}. \tag{50}$$

On appelle également cette condition la condition d'équilibre local, en raison de l'interprétation que l'on en fait. Supposons que l'on enregistre les événements d'une CM. Le nombre moyen de transitions $i \to j$, pour chaque événement enregistré, est $p_{ij}\pi_{Si}$. De manière analogue, pour les transitions $j \to i$, le nombre moyen de transitions est $\pi_{Sj}p_{ji}$. La **condition 50** stipule que ces deux quantités sont égales.

5.7 Chaînes de Markov à temps continu

Dans les sections précédentes, on a considéré que les transitions entre états ne pouvaient intervenir qu'à des instants discrets $0, 1, 2, \ldots, k, \ldots$. À présent, on considère que les transitions entre les états discrets $1, 2, \ldots, N$ peuvent survenir à n'importe quel instant $t \in \mathbb{R}$. On dénote par $X(t)$ le processus stochastique qui en résulte et on introduit la matrice de transition $P(t-s)$, de taille $N \times N$ et dont les éléments sont donnés par

$$p_{ij}(t-s) = P(X(t) = j \mid X(s) = i).$$

La propriété de Markov de $X(t)$ est équivalente à l'équation de Chapman-Kolmogorov

$$p_{ij}(s+t) = \sum_{n=1}^{N} p_{in}(s)p_{nj}(t).$$

En utilisant la notation matricielle $P(t)$, on peut réécrire cette expression comme

$$P(s+t) = P(s)P(t). \tag{51}$$

Dans ce cas, $s \geq 0$, $t \geq 0$ et

$$P(O) = \mathbb{I}, \tag{52}$$

où $\mathbb{I}$ désigne la matrice identité. $P(t)$ est différentiable, et en calculant sa dérivée, à partir de **51** on a

$$\frac{d}{dt}P(t) = P'(t) = QP(t), \tag{53}$$

où la matrice Q, appelée matrice d'intensité de la CM en temps continu $X(t)$, est donnée en prenant la limite de la dérivée en 0,

$$Q = \lim_{t \to 0^+} \frac{dP(t)}{dt}.$$

La construction de processus markovien $X(t)$ utilisés dans les applications pratiques, comme par exemple les modèles de substitution de nucléotides, se fait en définissant tout d'abord la matrice d'intensité Q. Il s'agit de l'approche la plus naturelle. Une fois données la matrice Q, la matrice de transitions $P(t)$ s'obtient en résolvant **53** en prenant **52** comme conditions initiales. La solution est

$$P(t) = \exp(Qt) = \sum_{m=1}^{\infty} \frac{(Qt)^m}{m!}.$$

Pour chaque $t \geq 0$, $P(t)$ est une matrice stochastique, et si l'on se donne une distribution de probabilités initiale $\pi(0)$ pour les états $1, 2, \ldots, N$, on peut calculer la distribution au temps t, à partir de la relation

$$P\big(X(t + \Delta t) = j \mid X(t) = i\big) = q_{ij}\Delta t + o(\Delta t).$$

Les éléments diagonaux de la matrice d'intensité Q sont définis comme suit :

$$q_{ii} = -\sum_{j \neq i} q_{ij}.$$

6 Méthodes de Monte Carlo par Chaînes de Markov (MCMC)

Les méthodes de Monte Carlo, reposant sur les générateurs de nombres aléatoires, permettent de réaliser une grande variété de tâches, incluant les simulations stochastiques, le calcul d'intégrales dans des dimensions élevées, ou l'optimisation de fonctions et de fonctionnelles. L'approche de Monte Carlo par Chaînes de Markov (MCMC) utilise les CM pour réaliser ce genre de tâches. Un outil important dans les méthodes MCMC est l'algorithme de Metropolis-Hastings (Metropolis et al., 1953 and Hastings, 1970). Celui-ci a été initialement conçu pour calculer des intégrales en plusieurs dimensions en physique moléculaire, mais a depuis trouvé de nombreux autres domaines d'application.

La méthode de Metropolis-Hastings permet de proposer une solution au problème suivant : construire une CM ergodique avec les états $1, 2, \ldots, N$ et une distribution stationnaire prédéfinie par un vecteur π_S. Par construire une CM, on entend définir ses probabilités de transition d'état. Il existe clairement une infinité de CM avec une distribution stationnaire π_S. Si l'on connaît les probabilités de transition d'état, on peut calculer la distribution stationnaire π_S, mais il n'existe pas de formule explicite pour la relation inverse. La méthode de Metropolis-Hastings offre une solution à ce problème en partant d'une CM ergodique avec les états $1, 2, \ldots, N$ et en modifiant ses probabilités de transition de sorte que la condition d'équilibre local **50** soit renforcée. Par conséquent, la CM modifiée devient inversible et possède bien la distribution stationnaire désirée π_S.

En utilisant cette idée, supposons que l'on a définit une CM irréductible et apériodique avec les états $1, 2, \ldots, N$ et les probabilités de transition q_{ij}. L'étape suivante consiste

à modifier ces probabilités en les multipliant par des facteurs a_{ij}, ce qui amène à une nouvelle CM de probabilités de transition

$$p_{ij} = a_{ij} q_{ij}. \tag{54}$$

On cherche les a_{ij} tels que les probabilités de transition p_{ij} satisfassent la condition d'équilibre local **50**. En substituant **54** dans **50**, on obtient

$$a_{ij} q_{ij} \pi_{Si} = a_{ji} q_{ji} \pi_{Sj}.$$

On a ici deux variables et une seule équation, donc une infinité de solutions possibles. La solution la plus simple consiste à supposer que l'un des facteurs a_{ij} et a_{ji} est égal à 1. Il existe deux possibilités. Néanmoins, on doit tenir compte de la condition que les facteurs multiplicatifs doivent satisfaire $a_{ij} \leq 1$ pour tout i, j. Cette condition découle du fait que la mise à l'échelle **54** ne doit pas produire des probabilités en dehors de l'intervalle $]0, 1]$. Cela amène finalement à la solution

$$a_{ij} = \min\left(1, \frac{q_{ji}\pi_{Sj}}{q_{ij}\pi_{Si}}\right). \tag{55}$$

L'**équation 54** avec les a_{ij} définis ci-dessus permet de calculer les probabilités de transition p_{ij} pour tout $i \neq j$. Pour les probabilités p_{ii}, on utilise la formule

$$p_{ii} = 1 - \sum_{j \neq i} p_{ij},$$

qui résulte de la **propriété 42**.

Comme on le voit dans la **règle 55**, l'expression des a_{ij} ne dépend pas de la valeur absolue des π_{Si} mais seulement de leur rapport. Cela signifie qu'il suffit de connaître π_S à une constante près. Il s'agit d'un résultat important qui permet de simuler des distributions pour lesquelles il est difficile de trouver une constante de normalisation.

6.1 Règle d'acceptation-rejet

La méthode de Metropolis-Hastings permettant de modifier les probabilités de transition (**54** et suivantes) peut être formulée sous la forme de la rège d'acceptation-rejet, très utilisée dans les applications pratiques. Supposons que l'on ait défini une CM irréductible et apériodique avec les états $1, 2, \ldots, N$ et les probabilités de transition q_{ij}, et par ailleurs que l'on dispose d'un programme permettant de simuler les transitions entre les états. La modification des probabilités de transition q_{ij} décrite dans les paragraphes précédents revient à ajouter la règle d'acceptation-rejet suivante au programme de simulation des transitions d'états. Lorsqu'une transition $i \to j$ est rencontrée, on calcule a_{ij} selon **55**. Si $a_{ij} = 1$, on change rien (on passe à l'état j). Si $a_{ij} < 1$, alors, avec probabilité a_{ij} on passe à l'état j et avec probabilité $1 - a_{ij}$, on supprime la transition $i \to j$ (on reste dans l'état i).

6.2 Applications de l'algorithme de Metropolis-Hastings

En utilisant l'algorithme de Metropolis-Hastings, on peut effectuer de l'échantillonnage aléatoire dans n'importe quelle distribution. Cela se révèle très utile, par exemple pour estimer la forme ou les paramètres de distributions *a posteriori* compliquées. Une autre application importante de l'algorithme de Metropolis-Hastings est l'optimisation stochastique. Un exemple de ce type de problématique est la recherche de l'arbre le plus probable en fonction des données. Pour chaque arbre, on calcule la probabilité correspondante (vraisemblance), mais en raison du nombre important d'arbres possibles, on ne peut pas tous les évaluer et sélectionner celui qui possède la probabilité la plus élevée. Au contraire, on peut construire une CM telle que les différents arbres corerspondent à ses états. En appliquant l'algorithme de Metropolis-Hastings, on visite (échantillonne) les arbres avec une fréquence correspondant à leurs probabilités. Les arbres avec une probabilité élevée sont ainsi visités plus fréquemment, alors que ceux possédant une probabilité plus faible ne seront vraisembablement pas visités du tout. Par la suite, on peut limiter la recherche des arbres les plus vraisemblables à ceux visités lors de la procédure d'échantillonnage par l'algorithme de Metropolis-Hastings.

6.3 Recuit simulé et MC3

Est-il également possible d'utiliser le principe de l'algorithme de Metropolis-Hastings pour optimiser n'importe quelle fonction $f(x)$, sur l'espace des arguments ? Le challenge est alors que $f(x)$ peut prendre des valeurs à la fois positives et négatives et ne possède donc pas d'interprétation probabiliste.

Considérons la transformation

$$p(x) = \exp\left(\frac{f(x)}{T}\right), \tag{56}$$

reposant sur l'idée de la distribution d'énergie de Boltzmann. La fonction $p(x)$ est toujours strictement positive et prend son maximum à la même valeur x_{max} que $f(x)$. Cette fonction ne correspond pas nécessairement à une distribution de probabilité puisque son intégrale ne vaut généralement pas 1. Toutefois, seules la positivité stricte est importante dans notre cas puisque, comme on l'a déjà mentionné, les relations **54** et suivantes ne dépendent que du rapport des éléments du vecteur π_S. Il est donc possible de construire un algorithme de recherche du maximum de $p(x)$ à l'aide de la technique de Metropolis-Hastings. Si l'espace des arguments x est continu, on le discrétisera avant d'appliquer l'algorithme.

L'**équation 56** contient un paramètre libre T. Par analogie avec la distribution d'énergie de Boltzmann, ce paramètre peut être interprété comme la « température ». Le changement de sa valeur influence les propriétés de l'algorithme d'échantillonnage. L'augmentation de la température entraîne une recherche plus intensive dans l'espace des arguments, puisque les transitions d'un $p(x)$ élevé à un $p(x)$ plus bas deviennent plus probables. La diminution de la température revient au contraire à rendre les transitions moins probables. Dans la méthode du *recuit simulé* (Kirkpatrick et al., 1983), la température est modifiée

en fonction d'un certain échéancier tout en se promenant dans l'espace des arguments. Les algorithmes de recuit simulé débute la recherche avec une température élevée, pui graduellement la température est diminuée lorsque les itérations approchent du voisinage du maximum.

Une autre idée assez intéressante, dénommée MC3, consiste à effectuer la recherche dans l'espace des arguments en utilisant plusieurs (généralement 3) échantillonneurs de Metropolis-Hastings à différentes températures (Madigan and York, 1995). Les algorithmes opèrent donc en parallèle et peuvent échanger leurs états en fonction de la valeur des vraisemblances.

7 Chaînes de Markov cachées

Dans la section précédente, lorsque nous avons présenté les propriétés des CM, on a implicitement considéré que les états étaient observables. Cependant, cette hypothèse n'est souvent pas satisfaite dans les applications des modèles de CM. Les chaînes de Markov cachées (CMC) sont alors fréquemment utilisées dans ce contexte (Durbin et al., 1999, Rabiner, 1989 and Koski and Koskinen, 2001). Un modèle de Markov caché est une CM dont les états ne sont pas observables. Seule une séquence de symboles émis par les états est enregistrée.

Plus spécifiquement, considérons une CM avec les états $1, 2, \ldots, N$ sur un intervalle de temps discret $0, 1, 2, \ldots, k, k+1, \ldots, K$. Par ailleurs, on considère M symboles possible dénotés $o_1, o_2, \ldots, o_m, o_{m+1}, \ldots, o_M$ et que l'on appelera des émissions. Chaque état possède une distribution de probabilité d'émissions

$$b_{im} = \Pr(\text{l'état } i \text{ émet } o_m). \tag{57}$$

7.1 Probabilité d'occurence d'une séquence de symboles

De 40 et 57, on conclut que la probabilité d'occurence des états $i_0, i_1, \ldots, i_K$ et des symboles $o_{j_0}, o_{j_1}, \ldots, o_{j_K}$ est

$$P(i_0, o_{j_0}, i_1, o_{j_1}, \ldots, i_K, o_{j_K}) = \pi_{i_0} b_{i_0 j_0} p_{i_0 i_1} b_{i_1 j_1} \cdots p_{i_{K-1} i_K} b_{i_K j_K}. \tag{58}$$

La probabilité d'enregistrer une séquence de symboles $o_{j_0}, o_{j_1}, \ldots, o_{j_K}$ est obtenue en sommant 58 sur l'ensemble des séquences $i_0, i_1, \ldots, i_K$ possibles, ce qui donne

$$P(o_{j_0}, o_{j_1}, \ldots, o_{j_K}) = \sum_{i_0=1}^{N} \pi_{i_0} b_{i_0 j_0} \sum_{i_1=1}^{N} p_{i_0 i_1} b_{i_1 j_1} \cdots \sum_{i_K=1}^{N} p_{i_{K-1} i_K} b_{i_K j_K}. \tag{59}$$

En pratique, lorsque l'on utilise l'expression ci-dessus, on arrange la sommation de manière récursive. Il y a alors deux possibilités auxquelles correspondent deux algorithmes différents : l'agorithme dit « backward » et l'algorithme « forward ».

7.2 Algorithme « backward »

On peut organiser le calcul récursif de 59 en partant de la dernière somme. On dénote celle-ci par

$$B_{K-1}(i_{K-1}) = \sum_{i_K=1}^{N} p_{i_{K-1}i_K} b_{i_K j_K}$$

et on voit que pour $B_k(i_k)$, défini comme

$$B_k(i_k) = \sum_{i_{k+1}=1}^{N} p_{i_k i_{k+1}} b_{i_{k+1} j_{k+1}} \cdots \sum_{i_K=1}^{N} p_{i_{K-1} i_K} b_{i_K j_K},$$

il existe une relation de récurrence

$$B_k(i_k) = \sum_{i_{k+1}=1}^{N} p_{i_k i_{k+1}} b_{i_{k+1} j_{k+1}} B_{k+1}(i_{k+1}),$$

valide pour $k = 0, 1, \ldots, K-2$. Finalement,

$$P(o_{j_0}, o_{j_1}, \ldots, o_{j_K}) = \sum_{i_0=1}^{N} \pi_{i_0} b_{i_0 j_0} B_0(i_0).$$

La récurrence définie ci-dessus implique de stocker des tableaux de taille N et des opérations de sommation sur un index.

7.3 Algorithme « forward »

Une autre possibilité consiste à partir de la première somme de l'**expression 59**. En définissant

$$F_k(i_k) = \sum_{i_0=1}^{N} \pi_{i_0} b_{i_0 j_0} \cdots \sum_{i_{k-1}=1}^{N} p_{i_{k-2} i_{k-1}} b_{i_{k-1} j_{k-1}} p_{i_{k-1} i_k},$$

on constate que $F_k(i_k)$, $k = 1, \ldots, K-1$ peut être calculé en utilisant la récursion suivante :

$$F_{k+1}(i_{k+1}) = \sum_{i_k=1}^{N} F_k(i_k) b_{i_k j_k} p_{i_k i_{k+1}}.$$

À présent, $P(o_{j_0}, o_{j_1}, \ldots, o_{j_K})$ est donné par

$$P(o_{j_0}, o_{j_1}, \ldots, o_{j_K}) = \sum_{i_k=1}^{N} F_K(i_K) b_{i_K j_K}.$$

Comme pour l'algorithme « backward », cet algorithme nécessite le stockage de tableaux de dimension N et des sommations sur un index.

7.4 Algorithme de Viterbi

L'algorithme de Viterbi permet de résoudre le problème suivant : étant donnée une séquence de symboles $o_{j_0}, o_{j_1}, \ldots, o_{j_K}$, trouver la séquence la plus probable pour les états $i_0, i_1, \ldots, i_K$. En d'autres termes, on cherche à calculer la séquence des états qui maximise la probabilité conditionnelle

$$P(i_0, i_1, \ldots, i_K \mid o_{j_0}, o_{j_1}, \ldots, o_{j_K}) = \frac{P(i_0, o_{j_0}, i_1, o_{j_1}, \ldots, i_K, o_{j_K})}{P(o_{j_0}, o_{j_1}, \ldots, o_{j_K})}.$$

Puisque $P(o_{j_0}, o_{j_1}, \ldots, o_{j_K})$ n'est qu'un facteur de normalisation dans ce cas, maximiser la probabilité conditionnelle revient à maximiser la probabilité conjointe **58** sur l'ensemble des séquences des états $i_0, i_1, \ldots, i_K$. En prenant le logarithme (naturel) de chaque membre de **58** et en définissant

$$L(i_0, i_1, \ldots, i_K) = \ln P(i_0, o_{j_0}, i_1, o_{j_1}, \ldots, i_K, o_{j_K}),$$

on obtient

$$L(i_0, i_1, \ldots, i_K) = \ln \pi_0 + \sum_{k=0}^{K-1} (\ln b_{i_k j_k} + \ln p_{i_k i_{k+1}})$$

et le problème de maximisation devient

$$\max_{i_0, i_1, \ldots, i_K} L(i_0, i_1, \ldots, i_K).$$

Ce problème de maximisation peut être résolu en utilisant les techniques de programmation dynamique puisque les décisions devant être prises à chaque étape intervienne séquentiellement et il est possible de définir des scores partiels pour chaque étape de ce processus, plus précisément

$$L_0(i_0, i_1, \ldots, i_K) = L(i_0, i_1, \ldots, i_K) = \ln \pi_0 + \sum_{k=0}^{K-1} (\ln b_{i_k j_k} + \ln p_{i_k i_{k+1}})$$

et

$$L_m(i_m, i_{m+1}, \ldots, i_K) = \sum_{k=m}^{K-1} (\ln b_{i_k j_k} + \ln p_{i_k i_{k+1}}).$$

À partir de ces deux expressions, on peut dériver une équation de Bellman pour la mise à jour des matrices des scores partiels oprimaux,

$$\hat{L}_{K-1}(i_{K-1}) = \max_{i_K} (\ln b_{i_{K-1} j_{K-1}} + \ln p_{i_{K-1} i_K})$$

et

$$\hat{L}_m(i_m) = \max_{i_{m+1}} \big(\ln b_{i_m j_m} + \ln p_{i_m i_{m+1}} + \hat{L}_{m+1}(i_{m+1}) \big).$$

En résolvant la récursion de Bellman ci-dessus, on peut calculer la solution à ce problème de maximisation.

7.5 Algorithme de Baum–Welch

Un autre problème souvent rencontré dans le domaine des CMC consiste en l'estimation des probabilités de transition d'une CM, lorsque l'on connaît une séquence de symboles $o_{j_0}, o_{j_1}, \ldots, o_{j_K}$. La solution par maximum de vraisemblance consiste à maximiser la probabilité donnée en **59** sur les entrées p_{ij} de la matrice des probabilités de transition de la CM considérée. Cependant, comme il s'agit d'un problème d'optimisation de grande dimension, il est nécessaire d'adopter une approche spécifique. L'une de ces approches consiste à utiliser l'algorithme de Baum–Welch qui repose sur l'idée des itérations EM présentées plus haut (**section 3**). Les paramètres à estimer sont les probabilités initiales des états, π_i, et les probabilités de transition, p_{ij}. Les variables observées sont les symboles $o_{j_0}, o_{j_1}, \ldots, o_{j_K}$. Les variables cachées sont les états $i_0, i_1, \ldots, i_K$. Avec ces hypothèses et en notant le vecteur incluant l'ensemble des paramètres estimés p, on peut spécifier $Q(p, p^{old})$ défini en **28** comme suit :

$$Q(p, p^{old}) = \sum_{i_0=1}^{N} \ldots \sum_{i_K=1}^{N} \left[\ln \pi_{i_0} + \sum_{k=0}^{K-1} \left(\ln b_{i_k j_k} + \ln p_{i_k i_{k+1}} \right) \right] \times \pi_{i_0}^{old} b_{i_0 j_0} p_{i_0 i_1}^{old} b_{i_1 j_1} \ldots p_{i_{K-1} i_K}^{old} b_{i_K j_K}.$$

L'expression ci-dessus constitue l'étape E. L'étape M consiste en la maximisation de $Q(p, p^{old})$ sur les paramètres π_i, p_{ij}, $i, j = 1, \ldots, N$. Ce faisant, on néglige bien sûr quelques détails de calcul. Le lecteur intéressé pourra se référer à (Koski and Koskinen, 2001).

8 Exercices

Les exercices qui suivent sont tirés de (Härdle and Hlávka, 2007). Certains dépassent le cadre des notions évoquées dans ce document, en particulier en ce qui concerne les applications du modèle linéaire. Les **solutions** aux exercices proposées sont disponibles sur le site hébergeant le présent doucument.

- It is well known that for two normal random variables, zero covariance implies independence. Why does this not apply to the following situation : $X \sim \mathcal{N}(0,1)$, $\mathrm{Cov}(X, X^2) = \mathbb{E}X^3 - \mathbb{E}X\mathbb{E}X^2 = 0 - 0 = 0$ but obviously X^2 is totally independent on X ?
- Trouver les valeurs $\hat{\alpha}$ et $\hat{\beta}$ qui minimisent la somme des carrés

$$\sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2.$$

- Soit $\mathcal{X}_* = \mathcal{H}\mathcal{X}\mathcal{D}^{-1/2}$, avec $\mathcal{X}$ une matrice $(n \times p)$, $\mathcal{H}$ une matrice de centrage et $\mathcal{D}^{-1/2} = \mathrm{diag}(s_{11}^{-1/2}, \ldots, s_{pp}^{-1/2})$. Montrer que $\mathcal{X}_*$ est une matrice standardisée, où $\bar{x}_* = 0_p$ et $S_{\mathcal{X}_*} = \mathcal{R}_{\mathcal{X}}$, la matrice de corrélation de $\mathcal{X}$.

- Un modèle linéaire peut s'exprimer sous la forme

$$Y = \mathcal{X}\beta + \varepsilon,$$

où $\mathcal{X}$ est de plein rang et ε symbolise les erreurs aléatoires. Montrer que la solution des moindres carrés

$$\hat{\beta} = \arg\min_{\beta}(Y - \mathcal{X}\beta)^T(Y - \mathcal{X}\beta) = \arg\min_{\beta}\varepsilon^T\varepsilon,$$

peut s'exprimer sous la forme $\hat{\beta} = (\mathcal{X}^T\mathcal{X})^{-1}\mathcal{X}^TY$. (voir aussi exercice 8)

- Supposons un vecteur aléatoire Y de distribution $Y \sim \mathcal{N}_p(0, \mathcal{I})$. Le transformer pour créer le vecteur $X \sim \mathcal{N}(\mu, \Sigma)$ avec $\mu = (3, 2)^T$ et $\Sigma = \begin{pmatrix} 1 & -1.5 \\ -1.5 & 4 \end{pmatrix}$. Comment peut-on implémenter la formule résultante sur un ordinateur ?

- Montrer que si $X \sim \mathcal{N}_p(0, \Sigma)$, alors la variable $U = (X - \mu)^T\Sigma^{-1}(X - \mu)$ suit une loi χ^2_p.

- Supposons que X soit de moyenne nulle et de covariance $\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$. Soit $Y = X_1 + X_2$. Écrire Y comme une transformation linéaire, c'est-à-dire trouver la matrice de transformation $\mathcal{A}$. Calculer ensuite $\mathbb{V}(Y)$.

- Calculer la moyenne et la variance de l'estimateur $\hat{\beta} = (\mathcal{X}^T\mathcal{X})^{-1}\mathcal{X}^TY$ dans le modèle linéaire $Y = \mathcal{X}\beta + \varepsilon$, où $\mathbb{E}(\varepsilon) = 0_n$ et $\mathbb{V}(\varepsilon) = \sigma^2\mathcal{I}_n$.

- Calculer les moments conditionnels $\mathbb{E}(X_2 \mid x_1)$ et $\mathbb{E}(X_1 \mid x_2)$ pour la fonction de densité bi-dimensionnelle suivante :

$$f(x_1, x_2) = \begin{cases} \frac{1}{2}x_1 + \frac{3}{2}x_2 & 0 \leq x_1, x_2 \leq 1 \\ 0 & \text{sinon} \end{cases}$$

- Montrer que $\mathbb{E}(X_2) = \mathbb{E}\{\mathbb{E}(X_2 \mid X_1)\}$, où $\mathbb{E}(X_2 \mid X_1)$ désigne l'espérance conditionnelle de X_2 connaissant X_1.

- Trouver la fonction de densité de probabilité associée au vecteur aléatoire $Y = \mathcal{A}X$ où $\mathcal{A} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$, sachant que X possède la fonction de densité définie à l'exercice 9.

- Montrer que la fonction

$$f_Y(y) = \begin{cases} \frac{1}{2}y_1 + \frac{1}{4}y_2 & 0 \leq y_1 \leq 2, \quad |y_2| \leq 1 - |1 - y_1| \\ 0 & \text{sinon} \end{cases}$$

est bien une densité de probabilité.

- Déterminer la distribution du vecteur aléatoire $Y = \mathcal{A}X$ avec $\mathcal{A} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$, où $X = (X_1, X_2)^T$ possède une distribution bi-normale.

Bibliographie

Billingsley, P. (1995). *Probability and Measure*. Wiley.

Feller, W. (1968). *An Introduction to Probability Theory and its Applications*, volume 1 and 2. Wiley.

Fisz, M. (1963). *Probability Theory and Mathematical Statistics*. Wiley.

Johnson, N. L., Kotz, S. and Balakrishnan, N. (1994). *Continuous Univariate Distributions*. Wiley.

Kendall, M. G., Stuart, A., Ord, J. K., Arnold, S. and O'Hagan, A. et al. (1991, 1999, 2004). *Kendall's Advanced Theory of Statistics*, volume 1, 2A, 2B. Oxford University Press.

Ditkin, V. A. and Prudnikov, A. P. (1965). *Integral Transforms and Operational Calculus*. Pergamon Press.

Wilf, H. S. (1990). *Generating Functionology*. Academic Press.

Kingman, J. F. C. (1993). *Poisson Processes*. Oxford University Press.

Hanson, K. M. and Wolf, D. R. (1996). In Heidbreder, G. R., editor, *Maximum Entropy and Bayesian Methods*, chapter Estimators for the Cauchy distribution, pages 255-263. Kluwer.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm (with discussion). *J. R. Statist. Soc., Ser. B*, 39:1-38. http://www.aliquote.org/pub/EM.pdf.

McLachan, G. J. and Krishnan, T. (1997). *The EM Algorithm and Extensions*. Wiley.

Cox, D. R. and Oakes, D. (1984). *Analysis of Survival Data*. Chapman and Hall, London.

McLachan, G. J. and Peel, W. (2000). *Finite Mixture Distributions*. Wiley.

Iosifescu, M. (1980). *Finite Markov Processes and Their Applications*. Wiley.

Gikhman, I. I. and Skorokhod, A. V. (1996). *Introduction to the Theory of Random Processes*. Dover.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N. and Teller, A. H. (1953). Equations of state calculations by fast computing machines. *J. Chem. Phys.*, 21:1087-1092. http://www.aliquote.org/pub/metropolis-et-al-1953.pdf.

Hastings, W. K. (1970). Monte carlo sampling method using markoc chains and their applications. *Biometrica*, 57:1317-1340. http://www.aliquote.org/pub/Hastings1970.pdf.

Kirkpatrick, S., Gelatt, C. D. and Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, 220:671-680. http://www.aliquote.org/pub/kirkpatrick83SA.pdf.

Madigan, D. and York, J. (1995). Bayesian graphical models for discrete data. *Internat. Statist. Rev.*, 63:215-232. http://www.aliquote.org/pub/10.1.1.9.1911.pdf.

Durbin, R., Eddy, S. R., Krogh, A. and Mitchison, G. (1999). *Biological Sequence Analysis : Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press.

Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proc IEEE*, 77:257-286. http://www.aliquote.org/pub/rabiner.pdf.

Koski, T. and Koskinen, T. (2001). *Hidden Markov Models for Bioinformatics*. Kluwer Academic.

Härdle, W. and Hlávka, Z. (2007). *Multivariate Statistics : Exercices and Solutions*. Springer.