Correction du TD 7

Exercice 1 (Lois conditionnelles I)

1. (a) Par définition la densité marginale f de X s'obtient à partir de la densité jointe de (X,Y) en intégrant par rapport à y. Donc pour tout $x \in \mathbb{R}$,

$$\begin{aligned} f(x) &= \int_{\mathbb{R}} \frac{1}{2\pi} \exp(-x^2 + xy - y^2/2)\,\mathrm{d}y = e^{-x^2} \int_{\mathbb{R}} \frac{1}{2\pi} e^{-(y-x)^2/2 + x^2/2}\,\mathrm{d}y \\ &= e^{-x^2/2} \int_{\mathbb{R}} \frac{1}{2\pi} e^{-(y-x)^2/2}\,\mathrm{d}y = e^{-x^2/2} \frac{1}{\sqrt{2\pi}}. \end{aligned}$$

Ainsi $\mathcal{L}(X) = \mathcal{N}(0,1)$. De même, la densité g de Y est donnée par, pour tout $y \in \mathbb{R}$,

$$\begin{aligned} g(y) &= \int_{\mathbb{R}} \frac{1}{2\pi} \exp(-x^2 + xy - y^2/2)\,\mathrm{d}x \\ &= \frac{e^{-y^2/2}}{2\pi} \int e^{-(x-\frac{y}{2})^2 + y^2/4}\,\mathrm{d}x \\ &= \frac{e^{-y^2/4}}{2\pi} \int e^{-(x-\frac{y}{2})^2}\,\mathrm{d}x = \frac{e^{-y^2/4}}{2\sqrt{\pi}}. \end{aligned}$$

On reconnaît la densité d'une loi $\mathcal{N}(0,2)$, qui est donc la loi de Y.

(b) Par définition la densité conditionnelle de Y sachant $X = x$ est, si $h(x,y)$ est la densité du couple (X,Y) (ici par rapport à la mesure de Lebesgue sur $\mathbb{R}^2$),

$$g_x(y) = \frac{h(x,y)}{\int_{\mathbb{R}} h(x,y)\,\mathrm{d}y} = \frac{h(x,y)}{f(x)} = \frac{\frac{1}{2\pi}\exp(-x^2 + xy - y^2/2)}{\frac{1}{\sqrt{2\pi}} e^{-x^2/2}},$$

d'après le calcul de $f(x)$ ci-dessus. On simplifie et on obtient

$$g_x(y) = \frac{1}{\sqrt{2\pi}} e^{-(y-x)^2/2}.$$

Ainsi $\mathcal{L}(Y \,|\, X = x) = \mathcal{N}(x,1)$. On aurait aussi pu obtenir ce résultat par la méthode du 'proportionnel à'. Utilisons cette méthode pour déterminer l'autre densité conditionnelle, $f_y(x)$. Comme on s'intéresse à une fonction de x, tout ce qui ne dépend pas de x peut aller dans la constante :

$$\begin{aligned} f_y(x) &\propto_x \exp(-x^2 + xy - y^2/2) \\ &\propto_x e^{-x^2 + xy} \\ &\propto_x e^{-(x-y/2)^2 + y^2/4} \\ &\propto_x e^{-(x-y/2)^2}. \end{aligned}$$

Ainsi $\mathcal{L}(X \,|\, Y = y) = \mathcal{N}(y/2, 1/2)$.

2. On rappelle que la densité d'un vecteur gaussien $\mathcal{N}(\mu, V)$ dans $\mathbb{R}^d$, avec $\mu \in \mathbb{R}^d$ et V une matrice symétrique définie positive de taille $d \times d$ est donnée par
$$z \mapsto \frac{1}{\sqrt{(2\pi)^d \det(V)}} \exp\left\{-\frac{1}{2}\,{}^t(z-\mu)V^{-1}(z-\mu)\right\}.$$
Ici, on nous dit que $Z = {}^t(X\ Y)$ (avec ${}^tu$ la transposée de u) a pour loi $\mathcal{N}(\mu, V)$. Il suffit d'identifier μ et V avec l'expression de $h(x, y)$ donnée par l'énoncé. Remarquons déjà que d'après la question précédente $\mathbb{E}(X) = \mathbb{E}(Y) = 0$ ce qui suggère de poser $\mu = 0$. Pour identifier V, on peut procéder par identification. Pour ${}^tz = {}^t(x\ y)$, et $M = \begin{bmatrix} a & b \\ b & c \end{bmatrix}$ une matrice 2×2 symétrique définie positive, on a
$$\frac{1}{2}\,{}^tzMz = ax^2/2 + bxy + cy^2/2.$$
En identifiant avec l'expression $f(x, y)$ il vient $a = 2$, $b = -1$, $c = 1$. Ceci nous donne la matrice $V^{-1} = \begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix}$. La formule d'inversion (transposée de la co-matrice) donne alors
$$V = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}.$$
On en conclut $\mathcal{L}(Z) = \mathcal{N}\left(0, \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}\right)$. Remarquons que l'on aurait aussi pu identifier V par
$$V = \begin{pmatrix} \mathrm{Var}(X) & \mathrm{Cov}(X, Y) \\ \mathrm{Cov}(X, Y) & \mathrm{Var}(Y) \end{pmatrix}.$$
On sait que $\mathrm{Var}(X) = 1$ et que $\mathrm{Var}(Y) = 2$. De plus, on a $\mathrm{Cov}(X, Y) = \mathbb{E}[XY] = \mathbb{E}\left[X\mathbb{E}[Y \,|\, X]\right] = \mathbb{E}[X^2] = 1$, car $\mathcal{L}(Y \,|\, X) = \mathcal{N}(X, 1)$. On trouve donc directement $V = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}$. On peut ensuite vérifier que la densité jointe de l'énoncé est bien celle d'un vecteur gaussien centré de matrice de covariance V.

Exercice 2 (Lois conditionnelles II) Une solution consiste à passer par les fonctions caractéristiques. Puisque X et Y sont indépendantes, on a en effet, pour tout réel u,
$$\Phi_S(u) = \Phi_X(u)\Phi_Y(u).$$
Il "suffit" donc de connaître les fonctions caractéristiques des lois de X et de Y et de reconnaître celle qui en découle pour S. Dans ce qui suit, nous faisons les calculs de deux manières : via les fonctions caractéristiques d'une part, et d'autre part en calculant directement le produit de convolution des lois de X et de Y.

1. Par les fonctions caractéristiques : pour tout réel u,
$$\Phi_S(u) = \Phi_X(u)\Phi_Y(u) = \exp(\lambda(e^{iu} - 1)) \exp(\mu(e^{iu} - 1)) = \exp((\lambda + \mu)(e^{iu} - 1))$$
donc $S \sim \mathcal{P}(\lambda + \mu)$.
Produit de convolution : pour tout $s \in \mathbb{N}$, on a
$$\begin{aligned}\mathbb{P}(S = s) &= \sum_{k=0}^{s} \mathbb{P}(X = k)\mathbb{P}(Y = s - k) \\ &= \sum_{k=0}^{s} \frac{e^{-(\lambda+\mu)}\lambda^k \mu^{s-k}}{k!(s-k)!} = \frac{e^{-(\lambda+\mu)}}{s!} \sum_{k=0}^{s} \binom{s}{k} \lambda^k \mu^{s-k} = \frac{e^{-(\lambda+\mu)}(\lambda + \mu)^s}{s!}.\end{aligned}$$

Ainsi $S \sim \mathcal{P}(\lambda+\mu)$.
Pour $x, s \in \mathbb{N}$, on a

$$\begin{aligned}\mathbb{P}(X=x \mid S=s) &= \frac{\mathbb{P}(X=x)\mathbb{P}(S=s \mid X=x)}{\mathbb{P}(S=s)} = \frac{\mathbb{P}(X=x)\mathbb{P}(Y=s-x)}{\mathbb{P}(S=s)} \\ &= \binom{s}{x}\frac{\lambda^x \mu^{s-x}}{(\lambda+\mu)^s}.\end{aligned}$$

Ainsi $\mathcal{L}(X \mid S) = \mathcal{B}\left(S, \frac{\lambda}{\lambda+\mu}\right)$ et $\mathbb{E}[X \mid S] = \frac{\lambda}{\lambda+\mu}S$.

2. Par les fonctions caractéristiques : pour tout réel u

$$\Phi_S(u) = \Phi_X(u)\Phi_Y(u) = (1-iu/\lambda)^{-r}(1-iu/\lambda)^{-t} = (1-iu/\lambda)^{-(r+t)}$$

donc $S \sim \Gamma(r+t, \lambda)$.
Produit de convolution : notons respectivement f_X, f_Y et f_S les densités en jeu. Puisque $S \geq 0$, le produit de convolution des densités s'écrit, pour tout $s \geq 0$,

$$f_S(s) = \int_0^s f_X(x)f_Y(s-x)\,\mathrm{d}x = \int_0^s \frac{\lambda^r x^{r-1}e^{-\lambda x}}{\Gamma(r)}\frac{\lambda^t(s-x)^{t-1}e^{-\lambda(s-x)}}{\Gamma(t)}\,\mathrm{d}x$$

donc, via le changement de variable $u = x/s$,

$$f_S(s) = \frac{\lambda^{r+t}s^{r+t-1}e^{-\lambda s}}{\Gamma(r)\Gamma(t)}\int_0^1 u^{r-1}(1-u)^{t-1}\,\mathrm{d}u = \frac{\lambda^{r+t}s^{r+t-1}e^{-\lambda s}}{\Gamma(r+t)}$$

car on a reconnu dans l'intégrale la densité non normalisée d'une loi $\mathrm{Beta}(r,t)$. Ainsi $S \sim \Gamma(r+t, \lambda)$.
Déterminons maintenant la densité de $S|X$. Soit φ une fonction borélienne bornée. Pour tout x,

$$\begin{aligned}\mathbb{E}\left[\varphi(S) \mid X=x\right] &= \mathbb{E}\left[\varphi(x+Y) \mid X=x\right] \\ &= \mathbb{E}\left[\varphi(x+Y)\right] \\ &= \int_{\mathbb{R}} \varphi(x+y)f_Y(y)\,dy \\ &= \int_{\mathbb{R}} \varphi(s)f_Y(s-x)\,ds\end{aligned}$$

où l'on a utilisé l'indépendance de X et Y. Ainsi, $S|X=x$ a pour densité $s \mapsto f_Y(s-x)$.
Autre méthode : pour déterminer la densité de $S|X$, on commence par celle du couple (X,S). Soit φ une fonction borélienne bornée de $\mathbb{R}^2$ dans $\mathbb{R}$, alors l'indépendance de X et Y permet d'écrire, via un changement de variable,

$$\mathbb{E}\left[\varphi(X,S)\right] = \mathbb{E}\left[\varphi(X, X+Y)\right] = \iint \varphi(x, x+y)f_X(x)f_Y(y)dxdy$$

et le changement de variable $s = x+y$ donne

$$\mathbb{E}\left[\varphi(X,S)\right] = \iint \varphi(x,s)f_X(x)f_Y(s-x)dxds,$$

ce qui montre que la densité jointe du couple (X,S) est $f_X(x)f_Y(s-x)$.
À présent, pour $s>0$, la densité conditionnelle $f_{X \mid S=s}$ de X sachant $S=s$ est donnée, pour $x>0$, par :

$$f_{X \mid S=s}(x) \propto_x f_X(x)f_Y(s-x) \propto_x x^{r-1}(s-x)^{t-1}\mathbb{1}_{0\leq x\leq s} \propto_x \left(\frac{x}{s}\right)^{r-1}\left(1-\frac{x}{s}\right)^{t-1}\mathbb{1}_{0\leq \frac{x}{s}\leq 1}.$$

Cela correspond à la densité d'une variable $s\,\mathrm{Beta}(r,t)$. Ainsi $X \mid S \sim S\,\mathrm{Beta}(r,t)$ et $\mathbb{E}[X \mid S] = \frac{r}{r+t}S$.

3. Par les fonctions caractéristiques : si $X \sim \mathcal{B}(n,p)$, alors $\Phi_X(u) = ((1-p)+pe^{iu})^n$ pour tout réel u, d'où
$$\Phi_S(u) = \Phi_X(u)\Phi_Y(u) = ((1-p)+pe^{iu})^n((1-p)+pe^{iu})^m = ((1-p)+pe^{iu})^{n+m}$$
donc $S \sim \mathcal{B}(n+m,p)$.
Produit de convolution : pour tout $s \in [\![0, n+m]\!]$, on a
$$\begin{aligned}\mathbb{P}(S=s) &= \sum_{k=0}^{s} \mathbb{P}(X=k)\mathbb{P}(Y=s-k) = \sum_{k=0}^{s} \binom{n}{k}\binom{m}{s-k} p^k(1-p)^{n-k}p^{s-k}(1-p)^{m-s+k} \\ &= \sum_{k=0}^{s} \binom{n}{k}\binom{m}{s-k} p^s(1-p)^{n+m-s} = \binom{n+m}{s} p^s(1-p)^{n+m-s}\end{aligned}$$
où l'on a utilisé l'identité de Vandermonde, à savoir que
$$\sum_{k=0}^{s} \binom{n}{k}\binom{m}{s-k} = \binom{n+m}{s},$$
laquelle peut se déduire des deux façons d'obtenir le coefficient de X^s dans le polynôme
$$(1+X)^{n+m} = (1+X)^n(1+X)^m.$$
Ainsi $S \sim \mathcal{B}(n+m,p)$.
Pour arriver au même résultat, une méthode plus rapide consiste à considérer des variables $(B_i)_{1\leq i\leq n+m}$ i.i.d. selon la loi de Bernoulli $\mathcal{B}(p)$ et à voir que, en loi, on a $X = B_1 + \cdots + B_n$ et $Y = B_{n+1} + \cdots + B_{n+m}$, de sorte que
$$S = X + Y = \sum_{i=1}^{n+m} B_i \sim \mathcal{B}(n+m,p).$$
Pour $x, s \in [\![0, n+m]\!]$, on a
$$\mathbb{P}(X = x \,|\, S = s) = \frac{\mathbb{P}(X=x)\mathbb{P}(Y=s-x)}{\mathbb{P}(S=s)} = \frac{\binom{n}{x}\binom{m}{s-x}}{\binom{n+m}{s}}.$$
On reconnaît la loi hypergéométrique de paramètres $(N,p,n) = (n+m, n/(n+m), s)$. On a donc par l'indication de l'énoncé : $\mathbb{E}[X \,|\, S] = \frac{n}{n+m}S$. Ce résultat peut se retrouver comme suit :
$$\mathbb{E}[X \,|\, S] = \sum_{x=0}^{S} \frac{x\binom{n}{x}\binom{m}{S-x}}{\binom{n+m}{S}} = \frac{1}{\binom{n+m}{S}} \sum_{x=0}^{S-1} n\binom{n-1}{x}\binom{m}{S-x-1} = \frac{n\binom{n+m-1}{S-1}}{\binom{n+m}{S}} = \frac{n}{n+m}S$$
où l'on a à nouveau appliqué l'identité de Vandermonde.

Exercice 3 (Tirages à pile ou face)

1. (a) La loi a posteriori s'obtient avec la formule de Bayes discrète
$$\begin{aligned}\mathbb{P}(\boldsymbol{\theta} = a \,|\, X = x) &= \frac{\mathbb{P}(\boldsymbol{\theta} = a, X = x)}{\mathbb{P}(X=x)} \\ &= \frac{\mathbb{P}(\boldsymbol{\theta} = a)\mathbb{P}(X = x \,|\, \boldsymbol{\theta} = a)}{\mathbb{P}(\boldsymbol{\theta} = a)\mathbb{P}(X = x \,|\, \boldsymbol{\theta} = a) + \mathbb{P}(\boldsymbol{\theta} = b)\mathbb{P}(X = x \,|\, \boldsymbol{\theta} = b)} \\ &= \frac{\frac{1}{2}a^x(1-a)^{1-x}}{\frac{1}{2}a^x(1-a)^{1-x} + \frac{1}{2}b^x(1-b)^{1-x}} \\ &= \left(\frac{a}{a+b}\right)^x \left(\frac{1-a}{2-a-b}\right)^{1-x},\end{aligned}$$

car $x=0$ ou $x=1$. On en déduit : $\mathbb{P}(\boldsymbol{\theta}=b \,|\, X=x) = 1-\mathbb{P}(\boldsymbol{\theta}=a \,|\, X=x)$.

(b) La loi jointe de $(X,\boldsymbol{\theta})$ est donnée par

$$\mathbb{P}(X=x,\boldsymbol{\theta}=y) = \frac{1}{2}\mathbb{1}_{y=a}a^x(1-a)^{1-x} + \frac{1}{2}\mathbb{1}_{y=b}b^x(1-b)^{1-x} = \frac{1}{2}y^x(1-y)^{1-x}$$

La loi de X est donnée par

$$\mathbb{P}(X=x) = \frac{1}{2}a^x(1-a)^{1-x} + \frac{1}{2}b^x(1-b)^{1-x} = \left(\frac{a+b}{2}\right)^x\left(1-\frac{a+b}{2}\right)^{1-x},$$

car $x=0$ ou $x=1$. La loi de X est donc une loi de Bernoulli de paramètre $\frac{a+b}{2}$.

(c) La loi a posteriori est donnée par les probabilités $\pi(\theta \,|\, \mathbf{X}) = \pi(\theta \,|\, X_1,\ldots,X_n)$ pour $\theta=a$ ou $\theta=b$. Par la formule de Bayes discrète

$$\begin{aligned}\pi(a \,|\, \mathbf{X}) &= \frac{\frac{1}{2}\prod_{i=1}^n a^{X_i}(1-a)^{1-X_i}}{\frac{1}{2}\prod_{i=1}^n a^{X_i}(1-a)^{1-X_i} + \frac{1}{2}\prod_{i=1}^n b^{X_i}(1-b)^{1-X_i}} \\ &= \frac{a^{n\overline{X}_n}(1-a)^{n-n\overline{X}_n}}{a^{n\overline{X}_n}(1-a)^{n-n\overline{X}_n} + b^{n\overline{X}_n}(1-b)^{n-n\overline{X}_n}}\end{aligned}$$

L'espérance a posteriori est $\mathbb{E}[\boldsymbol{\theta} \,|\, \mathbf{X}] = a\pi(a \,|\, \mathbf{X}) + b\pi(b \,|\, \mathbf{X})$, ce qui s'obtient à partir de l'expression précédente, en notant que $\pi(b \,|\, \mathbf{X}) = 1-\pi(a \,|\, \mathbf{X})$ (l'expression obtenue ne se simplifie pas particulièrement).

2. (a) Le cadre bayésien s'écrit

$$\begin{aligned}\boldsymbol{\theta} &\sim \Pi \\ X_1,\ldots,X_n \,|\, \boldsymbol{\theta} &\sim P_{\boldsymbol{\theta}}^{\otimes n},\end{aligned}$$

avec ici $P_\theta = \mathcal{B}(\theta)$ et $\Pi = \text{Unif}[0,1]$. Le paramètre $\boldsymbol{\theta}$ décrit le point de $[0,1]$ où s'est arrêtée la première boule. Les X_i sont des variables décrivant, sachant $\boldsymbol{\theta}$, si oui ou non au tirage i la deuxième boule s'arrête à gauche de $\boldsymbol{\theta}$. La loi a priori est uniforme, puisqu'on précise que les boules s'arrêtent de façon uniforme sur l'intervalle.

(b) Ici $p_\theta(X_1,\ldots,X_n) = \prod_{i=1}^n \theta^{X_i}(1-\theta)^{1-X_i}$ est la densité de $X_1,\ldots,X_n$ sachant $\boldsymbol{\theta}=\theta$, par rapport à la mesure $\mu^{\otimes n}$, où μ est la mesure de comptage sur $\{0,1\}$. Ainsi, comme $\pi(\theta) = \mathbb{1}_{[0,1]}(\theta)$ (loi uniforme sur $[0,1]$),

$$\pi(\theta \,|\, \mathbf{X}) \propto \mathbb{1}_{[0,1]}(\theta)\prod_{i=1}^n \theta^{X_i}(1-\theta)^{1-X_i} \propto \mathbb{1}_{[0,1]}(\theta)\theta^{n\overline{X}_n}(1-\theta)^{n-n\overline{X}_n}.$$

Il s'agit d'une loi Beta$(n\overline{X}_n+1, n-n\overline{X}_n+1)$.

Exercice 4 (Lois conjuguées)

1. On part de $\Pi = \text{Gamma}(a,b)$. Dans le modèle $P_\lambda^{(n)} = \mathcal{E}(\lambda)^{\otimes n}$, la loi a posteriori $\Pi[\cdot \,|\, \mathbf{X}]$ a pour densité

$$\begin{aligned}\pi(\lambda \,|\, \mathbf{X}) &\propto \pi(\lambda)\prod_{i=1}^n p_\lambda(X_i) \\ &\propto \lambda^{a-1}\,\mathrm{e}^{-b\lambda}\mathbb{1}_{\lambda>0}\prod_{i=1}^n \lambda e^{-\lambda X_i}\mathbb{1}_{X_i>0} \\ &\propto \lambda^{n+a-1}\,\mathrm{e}^{-(b+n\overline{X}_n)\lambda}\mathbb{1}_{\lambda>0}.\end{aligned}$$

Ainsi $\Pi[\cdot \mid \mathbf{X}] = \mathrm{Gamma}(a+n, b+n\overline{X}_n)$. La moyenne a posteriori vaut $\frac{a+n}{b+n\overline{X}_n}$ (la moyenne d'une $\mathrm{Gamma}(a,b)$ est a/b).

2. On procède de même. On part de $\Pi = \mathrm{Beta}(a,b)$. Dans le modèle de Bernoulli $\mathcal{B}(p)^{\otimes n}$, la loi a posteriori $\Pi[\cdot \mid \mathbf{X}]$ a pour densité

$$\begin{aligned}\pi(p \mid \mathbf{X}) &\propto \pi(p)\prod_{i=1}^n p_p(X_i)\\ &\propto p^{a-1}(1-p)^{b-1}\mathbb{1}_{0<p<1}\prod_{i=1}^n p^{X_i}(1-p)^{1-X_i}\\ &\propto p^{a+n\overline{X}_n-1}(1-p)^{b+n-n\overline{X}_n-1}\mathbb{1}_{0<p<1}\end{aligned}$$

Ainsi $\Pi[\cdot \mid \mathbf{X}] = \mathrm{Beta}(a+n\overline{X}_n, b+n-n\overline{X}_n)$. La moyenne a posteriori vaut $\frac{a+n\overline{X}_n}{a+b+n}$ (la moyenne d'une $\mathrm{Beta}(a,b)$ est $a/(a+b)$).

3. Soit $\Pi = \mathcal{N}(\mu,\sigma^2)$. Il suffit de montrer que, si on prend Π comme loi a priori sur $\boldsymbol{\theta}$ dans le modèle $\{\mathcal{N}(\theta,1)^{\otimes n},\ \theta\in\mathbb{R}\}$, la loi a posteriori $\Pi[\cdot \mid \mathbf{X}]$ est de la forme $\mathcal{N}(\mu',\sigma'^2)$. Ici $\mathbf{X} = (X_1,\dots,X_n)$, et $\mathbf{X} \mid \boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\theta},1)^{\otimes n}$. On a

$$\begin{aligned}\mathrm{d}\Pi(\theta) &= \pi(\theta)\,\mathrm{d}\theta, &\qquad \pi(\theta) &= \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}(\theta-\mu)^2}.\\ \mathrm{d}P_\theta(x) &= p_\theta(x)\,\mathrm{d}x, &\qquad p_\theta(x) &= \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(x-\theta)^2}.\end{aligned}$$

La densité a posteriori s'écrit

$$\begin{aligned}\pi(\theta \mid \mathbf{X}) &\propto \pi(\theta)\prod_{i=1}^n p_\theta(X_i)\\ &\propto \exp\left\{-\frac{1}{2\sigma^2}(\theta-\mu)^2 - \frac{1}{2}\sum_{i=1}^n (X_i-\theta)^2\right\}\\ &\propto \exp\left\{-\frac{1}{2\sigma^2}(\theta^2-2\mu\theta) - \frac{1}{2}(n\theta^2 - 2n\overline{X}_n\theta)\right\}\\ &\propto \exp\left\{-\frac{\sigma^{-2}+n}{2}\left(\theta^2 - 2\theta\frac{\mu\sigma^{-2}+n\overline{X}_n}{\sigma^{-2}+n}\right)\right\}\\ &\propto \exp\left\{-\frac{\sigma^{-2}+n}{2}\left(\theta - \frac{\mu\sigma^{-2}+n\overline{X}_n}{\sigma^{-2}+n}\right)^2\right\}.\end{aligned}$$

Ainsi

$$\Pi[\cdot \mid \mathbf{X}] = \mathcal{N}\left(\frac{\mu\sigma^{-2}+n\overline{X}_n}{\sigma^{-2}+n}, \frac{1}{\sigma^{-2}+n}\right). \tag{1}$$

Donc la famille considérée est conjuguée. La moyenne a posteriori vaut $\mathbb{E}[\boldsymbol{\theta} \mid \mathbf{X}] = \frac{\mu\sigma^{-2}+n\overline{X}_n}{\sigma^{-2}+n}$. On peut noter que cette moyenne a posteriori est une moyenne pondérée entre la moyenne μ de la loi a priori et la moyenne empirique $\overline{X}_n$ des observations :

$$\mathbb{E}[\boldsymbol{\theta} \mid \mathbf{X}] = \frac{\sigma^{-2}}{\sigma^{-2}+n}\mu + \frac{n}{\sigma^{-2}+n}\overline{X}_n.$$

En particulier, plus le nombre n d'observations est grand et moins la loi a priori affecte la loi a posteriori.

Exercice 5 (Empirical Bayes et lois exponentielles)

1. Elle consiste à construire, à partir des observations $\mathbf{X}$, un estimateur $\widehat{\lambda}$ du paramètre λ de la classe de lois a priori $\{\Pi_\lambda,\ \lambda \in \mathbb{R}\}$ considérée, et à suivre l'approche bayésienne avec pour loi a priori $\Pi_{\widehat{\lambda}}$.
2. Pour tout λ, on forme la densité marginale de $\mathbf{X}$ lorsque l'a priori est Π_λ, i.e. $f_\lambda(\mathbf{X}) = \int_\Theta p_\theta(\mathbf{X})\,\mathrm{d}\Pi_\lambda(\theta)$. Puis on construit $\widehat{\lambda}$ comme un élément qui maximise cette densité marginale, soit
$$\widehat{\lambda} \in \arg\max f_\lambda(\mathbf{X})\,.$$
3. La loi marginale de X_1 a pour densité
$$\begin{aligned} f_\lambda(x) &= \int_\Theta p_\theta(x)\,\mathrm{d}\Pi_\lambda(\theta) \\ &= \int_0^{+\infty} \theta e^{-\theta x}\mathbf{1}_{x\geq 0}\lambda e^{-\lambda\theta}\,\mathrm{d}\theta \\ &= \frac{\lambda}{\lambda+x}\mathbf{1}_{x\geq 0}\int_0^{+\infty} \theta(\lambda+x)e^{-(\lambda+x)\theta}\,\mathrm{d}\theta \\ &= \frac{\lambda}{\lambda+x}\mathbf{1}_{x\geq 0}\mathbb{E}[\mathcal{E}(\lambda+x)] \\ &= \frac{\lambda}{(\lambda+x)^2}\mathbf{1}_{x\geq 0}\,. \end{aligned}$$
4. Rappelons que si $T \sim \mathcal{E}(a)$, avec $a > 0$, alors pour tout $k \in \mathbb{N}^\star$, $\mathbb{E}[T^k] = \frac{k!}{a^k}$. Ainsi, la densité marginale de $(X_1,\ldots,X_n)$ s'écrit, calculée aux points observés,
$$\begin{aligned} \int_\Theta \prod_{i=1}^n p_\theta(X_i)\,\mathrm{d}\Pi_\lambda(\theta) &= \int_0^\infty \theta^n e^{-n\overline{X}_n\theta}\lambda e^{-\lambda\theta}\,\mathrm{d}\theta \\ &= \frac{\lambda}{\lambda+n\overline{X}_n}\int_0^\infty \theta^n(\lambda+n\overline{X}_n)e^{-(\lambda+n\overline{X}_n)\theta}\,\mathrm{d}\theta \\ &= \frac{\lambda}{\lambda+n\overline{X}_n}\mathbb{E}[(\mathcal{E}(\lambda+n\overline{X}_n))^n] \\ &= \frac{n!\lambda}{(\lambda+n\overline{X}_n)^{n+1}}. \end{aligned}$$
5. Il suffit donc de trouver le point où le maximum de cette densité est atteint, ce qui revient à déterminer le maximum de
$$\psi(\lambda) = \log(n!) + \log\lambda - (n+1)\log(\lambda+n\overline{X}_n).$$
On annule la dérivée et on trouve que $\widehat{\lambda} = \overline{X}_n$, après avoir vérifié qu'il s'agit bien d'un maximum (la dérivée est positive puis négative donc c'est bien le cas). La loi a posteriori finale suggérée par la méthode est donc $\Pi_{\widehat{\lambda}}[\cdot \,|\, \mathbf{X}]$. On calcule la densité a posteriori pour tout λ fixé
$$\pi_\lambda(\theta \,|\, \mathbf{X}) \propto \theta^n e^{-(\lambda+n\overline{X}_n)\theta}\mathbf{1}_{\theta\geq 0}.$$
Donc $\Pi_\lambda[\cdot \,|\, \mathbf{X}]$ est une loi $\Gamma(n+1,\lambda+n\overline{X}_n)$. Ainsi, la pseudo-loi a posteriori est une loi
$$\Gamma\left(n+1,(n+1)\overline{X}_n\right).$$
On remarque que la moyenne a posteriori est $1/\overline{X}_n$. Elle coïncide donc ici (ce sera aussi le cas dans l'exercice suivant) avec l'estimateur du maximum de vraisemblance pour θ dans ce modèle.

Exercice 6 (Empirical Bayes et lois normales)

1. Soient $\boldsymbol{\theta} \sim \mathcal{N}(\mu, 1)$ et $\mathbf{X}|\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\theta}\mathbb{1}, I_n)$. Il apparaît que $\mathbf{X} - \boldsymbol{\theta}\mathbb{1}$ et $\boldsymbol{\theta}$ sont indépendants puisque par définition $\mathbf{X} - \boldsymbol{\theta}\mathbb{1}|\boldsymbol{\theta} \sim \mathcal{N}(0, I_n)$, loi qui ne dépend pas de $\boldsymbol{\theta}$, donc $\mathbb{E}\left[\varphi(\mathbf{X} - \boldsymbol{\theta}\mathbb{1})|\boldsymbol{\theta}\right] = c$ constante, puis pour toute fonction $\varphi : \mathbb{R}^n \to \mathbb{R}$ mesurable bornée, on a

$$\mathbb{E}\left[\varphi(\mathbf{X} - \boldsymbol{\theta}\mathbb{1})\right] = \mathbb{E}[\mathbb{E}\left[\varphi(\mathbf{X} - \boldsymbol{\theta}\mathbb{1})|\boldsymbol{\theta}\right]] = \mathbb{E}[c] = c.$$

Par conséquent,

$$\begin{aligned}
\mathbb{E}\left[\varphi(\mathbf{X} - \boldsymbol{\theta}\mathbb{1})\psi(\boldsymbol{\theta})\right] &= \mathbb{E}\left[\psi(\boldsymbol{\theta})\mathbb{E}\left[\varphi(\mathbf{X} - \boldsymbol{\theta}\mathbb{1})|\boldsymbol{\theta}\right]\right] \\
&= \mathbb{E}\left[\psi(\boldsymbol{\theta})c\right] \\
&= \mathbb{E}\left[\psi(\boldsymbol{\theta})\right] c \\
&= \mathbb{E}\left[\psi(\boldsymbol{\theta})\right] \mathbb{E}\left[\varphi(\mathbf{X} - \boldsymbol{\theta}\mathbb{1})\right].
\end{aligned}$$

Ainsi $\mathbf{X} = \boldsymbol{\theta}\mathbb{1} + (\mathbf{X} - \boldsymbol{\theta}\mathbb{1}) \stackrel{\mathcal{L}}{=} \boldsymbol{\theta}\mathbb{1} + \mathbf{Z}$, où $\mathbf{Z} \sim \mathcal{N}(0, I_n)$, $\boldsymbol{\theta}\mathbb{1} \sim \mathcal{N}(\mu\mathbb{1}, J_n)$, où $J_n = \mathbb{1}^t\mathbb{1}$ est la matrice uniquement composée de 1, et $\mathbf{Z} \perp\!\!\!\perp \boldsymbol{\theta}$. Le vecteur $\mathbf{X}$ est donc gaussien de loi $\mathcal{N}(\mu\mathbb{1}, \Sigma)$ avec $\Sigma = I_n + J_n$. Puisque $J_n = nP$, où P est la matrice de projection sur $\text{Vect}(\mathbb{1})$, J_n admet n comme valeur propre simple (associée au vecteur propre $\mathbb{1}$) et 0 comme valeur propre de multiplicité $(n-1)$, donc Σ admet $(n+1)$ comme valeur propre simple (associée au vecteur propre $\mathbb{1}$) et 1 comme valeur propre de multiplicité $(n-1)$. En particulier Σ est inversible et $\Sigma^{-1}\mathbb{1} = \frac{1}{n+1}\mathbb{1}$.

Remarque : On aurait aussi pu simplement affirmer qu'en prenant $\boldsymbol{\theta} \sim \mathcal{N}(\mu, 1)$ indépendant des variables $(Y_1, \ldots, Y_n)$ elles-mêmes i.i.d. selon une loi $\mathcal{N}(0, 1)$, alors la construction bayésienne de l'énoncé revient à dire qu'en loi, on a $X_i = \boldsymbol{\theta} + Y_i$ pour tout i. Les $(n+1)$ variables $\boldsymbol{\theta}$ et $(Y_1, \ldots, Y_n)$ étant gaussiennes indépendantes, le vecteur $\mathbf{X}$ est gaussien. Par contruction les X_i sont de même loi et de moyenne μ. Par ailleurs, un calcul immédiat donne $\text{Var}(X_i) = 2$ pour tout i, et $\text{Cov}(X_i, X_j) = 1$ pour tout $i \neq j$, ce qui correspond bien à la matrice Σ.

Cherchons maintenant $\widehat{\mu}$, l'estimateur du maximum de vraisemblance marginal.

Première méthode.

Puisque Σ est inversible, $\mathbf{X}$ a une densité f_μ par rapport à la mesure de Lebesgue sur $\mathbb{R}^n$, définie pour tout $x \in \mathbb{R}^n$ par

$$f_\mu(x) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}{}^t(x-\mu\mathbb{1})\Sigma^{-1}(x-\mu\mathbb{1})}.$$

Maximiser en μ la vraisemblance $f_\mu(\mathbf{X})$ revient à minimiser

$${}^t(\mathbf{X} - \mu\mathbb{1})\Sigma^{-1}(\mathbf{X} - \mu\mathbb{1}),$$

ce qui équivaut à minimiser, en tenant compte du fait que $\Sigma^{-1}\mathbb{1} = \frac{1}{n+1}\mathbb{1}$,

$$({}^t\mathbb{1}\Sigma^{-1}\mathbb{1})\mu^2 - 2({}^t\mathbf{X}\Sigma^{-1}\mathbb{1})\mu = \frac{n}{n+1}\mu^2 - 2\frac{n}{n+1}\overline{X}_n\mu.$$

On obtient l'estimateur $\widehat{\mu} = \overline{X}_n$.

Deuxième méthode.

Pour calculer l'estimateur du maximum de vraisemblance marginale, on n'a en fait pas besoin de déterminer exactement la densité $f_\mu(X_1, \ldots, X_n)$, mais seulement à une constante indépendante de

μ près.

$$\begin{aligned}
f_\mu(X_1,\ldots,X_n) &= \int \prod_{i=1}^n \left(\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(X_i-\theta)^2}\right)\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(\theta-\mu)^2}\,\mathrm{d}\theta \\
&\propto_{\mu,\theta}\ e^{-\frac{\mu^2}{2}}\int e^{-\frac{n}{2}\theta^2+n\overline{X}_n\theta-\frac{1}{2}\theta^2+\theta\mu}\,\mathrm{d}\theta \\
&\propto_{\mu,\theta}\ e^{-\frac{\mu^2}{2}}\int e^{-\frac{n+1}{2}\left(\theta-\frac{n\overline{X}_n+\mu}{n+1}\right)^2}e^{\frac{(n\overline{X}_n+\mu)^2}{2(n+1)}}\,\mathrm{d}\theta \\
&\propto_\mu\ e^{-\frac{\mu^2}{2}+\frac{(n\overline{X}_n+\mu)^2}{2(n+1)}} \\
&\propto_\mu\ e^{-\frac{n}{2(n+1)}(\mu-\overline{X}_n)^2}.
\end{aligned}$$

Le maximum en μ de la densité marginale de $\mathbf{X}$ est donc atteint pour $\mu = \overline{X}_n$.

2. La loi a posteriori correspondant à la loi a priori $\Pi_\mu = \mathcal{N}(\mu, 1)$ est, par le calcul habituel (voir l'équation (1)), une loi $\mathcal{N}\left(\frac{n\overline{X}_n+\mu}{n+1}, \frac{1}{n+1}\right)$, on obtient comme pseudo-loi a posteriori la loi $\mathcal{N}\left(\overline{X}_n, \frac{1}{n+1}\right)$. On remarque que cette loi est centrée en le maximum de vraisemblance $\overline{X}_n$ pour ce modèle.

Exercice 7 (Laplace, mélange et bayésien hiérarchique)

1. La densité (marginale) f de X s'écrit, pour tout $x \in \mathbb{R}$,

$$f(x) = \int_0^\infty \frac{1}{\sqrt{2\pi\cdot 2t}}e^{-\frac{1}{4t}x^2}\lambda e^{-\lambda t}\,\mathrm{d}t.$$

2. La fonction génératrice $M_X(s) = \mathbb{E}[e^{sX}]$, pour tout s tel que $|s| < \sqrt{\lambda}$, s'écrit

$$\begin{aligned}
M_X(s) = \mathbb{E}[e^{sX}] &= \mathbb{E}\left[\mathbb{E}[e^{sX} \,\middle|\, T]\right] \\
&= \mathbb{E}\left[\int e^{sx}\frac{1}{\sqrt{4\pi T}}e^{-\frac{1}{4T}x^2}\,\mathrm{d}x\right] \\
&= \mathbb{E}\left[\int \frac{1}{\sqrt{4\pi T}}e^{-\frac{1}{4T}(x-2Ts)^2+Ts^2}\,\mathrm{d}x\right] \\
&= \mathbb{E}\left[e^{s^2T}\right] = \int_0^\infty e^{s^2t}\lambda e^{-\lambda t}\,\mathrm{d}t = \frac{\lambda}{\lambda-s^2}.
\end{aligned}$$

Remarque : dans le calcul précédent, on aurait pu aller plus vite via la fonction caractéristique d'une loi normale. En effet, si $Y \sim \mathcal{N}(m,\sigma^2)$, alors $\Phi_Y(t) = \exp(imt - \sigma^2t^2/2)$. Pour calculer $\mathbb{E}[e^{sX} \,|\, T]$, il suffit d'appliquer ce résultat avec $m = 0$, $\sigma^2 = 2T$ et en remplaçant t par $-is$.

3. Si Y suit une loi de Laplace de paramètre μ, alors pour tout $|s| < \mu$,

$$\begin{aligned}
M_Y(s) = \mathbb{E}[e^{sY}] &= \int e^{sy}\frac{\mu}{2}e^{-\mu|y|}\,\mathrm{d}y \\
&= \frac{\mu}{2}\left(\int_{-\infty}^0 e^{(s+\mu)y}\,\mathrm{d}y + \int_0^\infty e^{(s-\mu)y}\,\mathrm{d}y\right) \\
&= \frac{\mu}{2}\left(\frac{1}{s+\mu}+\frac{1}{\mu-s}\right) \\
&= \mu^2/(\mu^2-s^2).
\end{aligned}$$

4. D'après ce qui précède, comme la fonction génératrice identifie la loi, on obtient que X suit une loi de Laplace de paramètre $\sqrt{\lambda}$. La loi a priori correspondant à la hiérarchie $\theta \,\big|\, T \sim \mathcal{N}(0, 2T)$ et $T \sim \mathcal{E}(\alpha)$ est donc une loi Laplace($\sqrt{\alpha}$).

Exercice 8 ($\star$) (Famille de Dirichlet et modèle multinomial)

1. (a) Soit $Z = (Z_1, \ldots, Z_K) \sim \mathrm{Dir}(a)$, où $a = (a_1, \ldots, a_K)$.
Première méthode. Soit $\varphi : \mathbb{R} \to \mathbb{R}$ continue bornée. On commence par remarquer qu'intégrer sur le simplexe $\mathcal{S}_K$, c'est en fait intégrer sur $\{z \in [0,1]^{K-1}, \sum_{i=1}^{K-1} z_i \leq 1\}$ (et z_K est alors donné par $1 - z_1 - \cdots - z_{K-1}$) :

$$\mathbb{E}[\varphi(Z_1)] = \frac{1}{B(a)} \int_{[0,1]^{K-1}} \varphi(z_1) \prod_{i=1}^{K-1} z_i^{a_i - 1} \left(1 - \sum_{j=1}^{K-1} z_j\right)^{a_K - 1} \mathbb{1}_{\{\sum_{i=1}^{K-1} z_i \leq 1\}} \,\mathrm{d}z_1 \ldots \mathrm{d}z_{K-1} \,.$$

En intégrant d'abord sur z_1 puis sur le reste, il vient

$$\int_{[0,1]^{K-1}} \varphi(z_1) \prod_{i=1}^{K-1} z_i^{a_i - 1} \left(1 - \sum_{j=1}^{K-1} z_j\right)^{a_K - 1} \mathbb{1}_{\{\sum_{i=1}^{K-1} z_i \leq 1\}} \,\mathrm{d}z_1 \ldots \mathrm{d}z_{K-1}$$
$$= \int_0^1 \varphi(z_1) z_1^{a_1 - 1} \left(\int_{[0,1-z_1]^{K-2}} \prod_{i=2}^{K-1} z_i^{a_i - 1} \left(1 - z_1 - \sum_{j=2}^{K-1} z_j\right)^{a_K - 1} \mathbb{1}_{\{\sum_{i=2}^{K-1} z_i \leq 1 - z_1\}} \,\mathrm{d}z_2 \ldots \mathrm{d}z_{K-1} \right) \mathrm{d}z_1$$

et en effectuant, pour z_1 fixé, le changement de variable $u_j = \frac{z_j}{1 - z_1}$ pour $j = 2, \ldots, K-1$ (de jacobien $(1 - z_1)^{K-2}$), on a

$$\int_{[0,1-z_1]^{K-2}} \prod_{i=2}^{K-1} z_i^{a_i - 1} \left(1 - z_1 - \sum_{j=2}^{K-1} z_j\right)^{a_K - 1} \mathbb{1}_{\{\sum_{i=2}^{K-1} z_i \leq 1 - z_1\}} \,\mathrm{d}z_2 \ldots \mathrm{d}z_{K-1}$$
$$= (1 - z_1)^{\sum_{j=2}^{K} a_j - 1} \int_{[0,1]^{K-2}} \prod_{i=2}^{K-1} u_i^{a_i - 1} \left(1 - \sum_{j=2}^{K-1} u_j\right)^{a_K - 1} \mathbb{1}_{\{\sum_{i=2}^{K-1} u_i \leq 1\}} \,\mathrm{d}u_2 \ldots \mathrm{d}u_{K-1} \,.$$

Enfin, en remarquant que

$$\int_{[0,1]^{K-2}} \prod_{i=2}^{K-1} u_i^{a_i - 1} \left(1 - \sum_{j=2}^{K-1} u_j\right)^{a_K - 1} \mathbb{1}_{\{\sum_{i=2}^{K-1} u_i \leq 1\}} \,\mathrm{d}u_2 \ldots \mathrm{d}u_{K-1}$$
$$= B(a_2, \ldots, a_K) = \frac{\prod_{j=2}^{K} \Gamma(a_j)}{\Gamma(\sum_{j=2}^{K} a_j)} \,,$$

on obtient

$$\mathbb{E}[\varphi(Z_1)] = \frac{\Gamma(\sum_{j=1}^{K} a_j)}{\Gamma(a_1)\Gamma(\sum_{j=2}^{K} a_j)} \int_0^1 \varphi(z_1) z_1^{a_1 - 1} (1 - z_1)^{\sum_{j=2}^{K} a_j - 1} \,\mathrm{d}z_1 \,.$$

Autrement dit, $Z_1 \sim$ Beta$(a_1, \sum_{j=2}^{K} a_j)$.
Seconde méthode. On note $f(z_1)$ la densité de Z_1 et on utilise le symbole $\propto$, sous-entendu $\propto_{z_1}$. Puisque la densité jointe de $z = (z_1, \ldots, z_{K-1})$ par rapport à la mesure de Lebesgue sur $\mathbb{R}^{K-1}$ vaut

$$f(z_1, \ldots, z_{K-1}) = \frac{1}{B(a)} \prod_{i=1}^{K-1} z_i^{a_i - 1} \left(1 - \sum_{j=1}^{K-1} z_j\right)^{a_K - 1} \mathbb{1}_{\{z \in [0,1]^{K-1}, \sum_{i=1}^{K-1} z_i \leq 1\}},$$

ou encore

$$f(z_1,\ldots,z_{K-1}) \propto z_1^{a_1-1}\prod_{i=2}^{K-1} z_i^{a_i-1}\left(1-z_1-\sum_{j=2}^{K-1} z_j\right)^{a_K-1}\mathbb{1}_{\left\{z\in[0,1]^{K-1},\ \sum_{i=2}^{K-1} z_i\leq 1-z_1\right\}},$$

la marginalisation donne, pour tout $z_1 \in [0,1]$,

$$f(z_1) \propto z_1^{a_1-1}\int_{[0,1-z_1]^{K-2}}\prod_{i=2}^{K-1} z_i^{a_i-1}\left(1-z_1-\sum_{j=2}^{K-1} z_j\right)^{a_K-1}\mathbb{1}_{\left\{\sum_{i=2}^{K-1} z_i\leq 1-z_1\right\}}\,\mathrm{d}z_2\ldots\mathrm{d}z_{K-1}$$

En effectuant, pour z_1 fixé, le changement de variable $u_j = \frac{z_j}{1-z_1}$ pour $j = 2,\ldots,K-1$ (de jacobien $(1-z_1)^{K-2}$), on en déduit

$$f(z_1) \propto z_1^{a_1-1}(1-z_1)^{\sum_{j=2}^K a_j-1}\int_{[0,1]^{K-2}}\prod_{i=2}^{K-1} u_i^{a_i-1}\left(1-\sum_{j=2}^{K-1} u_j\right)^{a_K-1}\mathbb{1}_{\left\{\sum_{i=2}^{K-1} u_i\leq 1\right\}}\,\mathrm{d}u_2\ldots\mathrm{d}u_{K-1}$$

où le terme intégral a le bon goût de ne plus dépendre de z_1, donc

$$f(z_1) \propto z_1^{a_1-1}(1-z_1)^{\sum_{j=2}^K a_j-1}$$

ce qui prouve que $Z_1 \sim$ Beta$(a_1, \sum_{j=2}^K a_j)$.

(b) On se place dans le cadre bayésien :

$$p = (p_1,\ldots,p_K) \sim \Pi = \mathrm{Dir}(a), \qquad \text{avec } a = (a_1,\ldots,a_K) \in (\mathbb{R}_+^*)^K$$
$$\mathbf{X}\,\big|\,p \sim \left(\sum_{k=1}^K p_k\delta_k\right)^{\otimes n}.$$

Autrement dit, conditionnellement à p, les variables $X_1,\ldots,X_n$ sont i.i.d. à valeurs dans $\{1,\ldots,K\}$, de loi donnée par

$$\forall k \in \{1,\ldots,K\},\ \mathbb{P}(X_1 = k\,|\,p) = p_k.$$

La loi a posteriori s'écrit

$$\begin{aligned}\pi(p\,|\,\mathbf{X}) &\propto \prod_{k=1}^K p_k^{a_k-1}\mathbb{1}_{p\in\mathcal{S}_K}\prod_{i=1}^n\prod_{k=1}^K p_k^{\mathbb{1}_{X_i=k}}\\ &\propto \prod_{k=1}^K p_k^{a_k+N_k-1}\mathbb{1}_{p\in\mathcal{S}_K},\end{aligned}$$

où l'on a noté $N_k = \sum_{i=1}^n \mathbb{1}_{X_i=k}$ le nombre d'observations égales à k dans l'échantillon. Ainsi $\Pi[p\,|\,\mathbf{X}] = \mathrm{Dir}(a_1+N_1,\ldots,a_K+N_K)$. La famille des lois de Dirichlet est donc bien conjuguée pour ce modèle.

2. (a) Pour tout $j \in \{1,\ldots,K\}$, la variable N_j suit une loi binomiale $\mathcal{B}(n,p_j)$.

(b) Pour avoir $N_j = n_j$ pour tout j, il faut

i. choisir une partition de $\{1,\ldots,n\}$ en K sous-ensembles $U_1\cup\cdots\cup U_K$ de tailles respectives $n_1,\ldots,n_K$. Il y a, par définition, $\binom{n}{n_1,\ldots,n_K}$ façons de faire cela.

ii. pour une partition fixée, la probabilité que les observations $(X_i)_{i\in U_k}$ soient toutes égales à k vaut $p_k^{n_k}$, pour tout $1 \le k \le K$.

On note enfin qu'il faut avoir $\sum_{j=1}^K n_j = n$, puisque

$$\sum_j \sum_i \mathbb{1}_{X_i=j} = \sum_i \sum_j \mathbb{1}_{X_i=j} = \sum_i 1 = n.$$

(c) Soit

$$\begin{aligned} p = (p_1, \dots, p_K) &\sim \Pi = \mathrm{Dir}(a), \qquad \text{avec } a = (a_1, \dots, a_K) \in (\mathbb{R}_+^*)^K \\ N = (N_1, \dots, N_K) \,\big|\, p &\sim \mathrm{Mult}(n, p_1, \dots, p_K). \end{aligned}$$

La loi a posteriori s'écrit

$$\begin{aligned} \pi(p \,|\, N) &\propto \prod_{k=1}^K p_k^{a_k-1} \mathbb{1}_{p\in\mathcal{S}_K} \prod_{k=1}^K p_k^{N_k} \\ &\propto \prod_{k=1}^K p_k^{a_k+N_k-1} \mathbb{1}_{p\in\mathcal{S}_K}, \end{aligned}$$

Ainsi $\Pi[p \,|\, N] = \mathrm{Dir}(a_1 + N_1, \dots, a_K + N_K)$. La famille des lois de Dirichlet est donc bien conjuguée pour ce modèle.

Exercice 9 ($\star$)(A posteriori séquentiel et information)

1. La formule de Bayes donne les densités a posteriori

$$\pi(\theta \,|\, X_1) = \frac{p_\theta(X_1)\pi(\theta)}{\int p_\theta(X_1)\pi(\theta)\,\mathrm{d}\nu(\theta)}, \quad \pi(\theta \,|\, X_1, X_2) = \frac{p_\theta(X_1)p_\theta(X_2)\pi(\theta)}{\int p_\theta(X_1)p_\theta(X_2)\pi(\theta)\,\mathrm{d}\nu(\theta)}.$$

Posons $\widetilde{\Pi} = \Pi[\cdot \,|\, X_1]$ et $\widetilde{\pi} = \pi(\cdot \,|\, X_1)$ la densité de $\widetilde{\Pi}$ par rapport à ν. On considère le cadre

$$\begin{aligned} \theta &\sim \widetilde{\Pi} \\ X_2 \,|\, \theta &\sim P_\theta \end{aligned}$$

La formule de Bayes donne que la densité a posteriori $\widetilde{\pi}(\cdot \,|\, X_2)$ est

$$\widetilde{\pi}(\theta \,|\, X_2) = \frac{p_\theta(X_2)\widetilde{\pi}(\theta)}{\int p_\theta(X_2)\widetilde{\pi}(\theta)\,\mathrm{d}\nu(\theta)} \propto p_\theta(X_2)\pi(\theta \,|\, X_1) \propto p_\theta(X_2)p_\theta(X_1)\pi(\theta).$$

Cette dernière quantité est donc bien à constante près $\pi(\theta \,|\, X_1, X_2)$. On en conclut que $\widetilde{\Pi}[\cdot \,|\, X_2] = \Pi[\cdot \,|\, X_1, X_2]$.

2. Par récurrence, on a de même $\Pi[\cdot \,|\, X_1, \dots, X_n] = \widetilde{\Pi}_{n-1}[\cdot \,|\, X_n]$ où $\widetilde{\Pi}_{n-1} = \Pi[\cdot \,|\, X_1, \dots, X_{n-1}]$ et ainsi de suite $\widetilde{\Pi}_{n-1} = \widetilde{\Pi}_{n-2}[\cdot \,|\, X_{n-1}]$ etc.

3. L'ordre des X_i n'a pas d'importance, puisque le produit dans l'expression de la vraisemblance

$$\prod_{i=1}^n p_\theta(X_i)$$

est commutatif. En revanche, si la loi de $X_1, \ldots, X_n \,\big|\, \theta$ n'était pas produit mais une loi $P_\theta^{(n)}$ arbitraire, l'ordre des X_i pourrait compter au sens où la densité $p_\theta^{(n)}(x_1, \ldots, x_n)$ pourrait être différente de $p_\theta^{(n)}(x_{\pi(1)}, \ldots, x_{\pi(n)})$ pour une permutation π de $[\![1, n]\!]$.

Remarque. En cours, pour n observations nous avons écrit la formule de Bayes dans le cas où $P_\theta^{(n)} = P_\theta^{\otimes n}$. Elle s'écrit bien sûr plus généralement pour $P_\theta^{(n)}$ quelconque avec $dP_\theta^{(n)}(x_1, \ldots, x_n) = p_\theta^{(n)}(x_1, \ldots, x_n) d\mu(\theta)$ comme

$$\pi(\theta \,\big|\, X_1, \ldots, X_n) = \frac{p_\theta(X_1, \ldots, X_n)\pi(\theta)}{\int_\Theta p_\theta(X_1, \ldots, X_n)\pi(\theta)\,\mathrm{d}\nu(\theta)}.$$

Exercice 10 (Estimateur de Laplace-Bayes)

1. S'il y a r boules rouges dans l'urne, alors, pour $x \in \{0, \ldots, r\}$, on a

$$\mathbb{P}_r(X = x) = \frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}}.$$

En effet, le nombre total de tirages possibles est $\binom{N}{n}$, et le nombre de tirages qui donnent x boules rouges est $\binom{r}{x}\binom{N-r}{n-x}$ (on choisit x boules parmi les r boules rouges et $n - x$ parmi les $N - r$ boules noires). La loi conditionnelle de X sachant r, notée P_r, est la loi hypergéométrique de paramètre (N, n, r).

2. La loi conditionnelle de r sachant X est donnée par

$$\pi(r \,\big|\, X) = \frac{\frac{1}{N+1}\frac{\binom{r}{X}\binom{N-r}{n-X}}{\binom{N}{n}}}{\sum_{k=0}^{N}\frac{1}{N+1}\frac{\binom{k}{X}\binom{N-k}{n-X}}{\binom{N}{n}}} = \frac{\binom{r}{X}\binom{N-r}{n-X}}{\sum_{k=0}^{N}\binom{k}{X}\binom{N-k}{n-X}} = \frac{\binom{r}{X}\binom{N-r}{n-X}}{\binom{N+1}{n+1}}.$$

où l'on a utilisé la formule donnée en indication sous la forme : $\binom{N+1}{n+1} = \sum_{k=1}^{N+1}\binom{k-1}{X}\binom{N+1-k}{n-X}$. En effet, si on considère $N + 1$ boules numérotées et on s'intéresse au nombre de façons de tirer (sans prendre en compte l'ordre) $n + 1$ boules parmi ces $N + 1$ boules. On met les boules ainsi tirées dans l'ordre et on regarde la $X + 1$-ème boules ainsi obtenue. Si celle-ci vaut k, cela veut dire qu'il y a X boules prenant leurs valeurs dans $\{1, \ldots, k - 1\}$ (et donc $\binom{k-1}{X}$ façons de les choisir) et $N - X$ boules prenant leurs valeurs dans $\{k + 1, \ldots, N + 1$ (et donc $\binom{N+1-k}{N-X}$ façons de les choisir), ce qui nous fait $\binom{k-1}{X}\binom{N+1-k}{N-X}$ tirages tels que la $X + 1$-ème plus grande boule vaille k. On obtient ainsi (en sommant sur tous les k possibles)

$$\binom{N+1}{n+1} = \sum_{k=1}^{N+1}\binom{k-1}{X}\binom{N+1-k}{n-X}.$$

Une autre preuve, analytique, est la suivante : on commence par montrer par récurrence que pour tout $n \in \mathbb{N}^\star$:

$$\frac{1}{(1-X)^n} = \sum_{j=0}^{\infty}\binom{n+j-1}{j}X^j,$$

puis on développe de deux manières $\frac{1}{(1-X)^{n+1}} \times \frac{1}{(1-X)^{m+1}}$.

3. Notons E l'événement « la $(n+1)^{\text{ième}}$ boule tirée est rouge ». Par la formule des probabilités totales appliquée, on a pour tout $x \leq N$

$$\mathbb{P}(E \,|\, X = x) = \sum_{r=x}^{N} \pi(r \,|\, X = x)\mathbb{P}(E \,|\, X = x, R = r).$$

Or

$$\mathbb{P}(E \,|\, X = x, R = r) = \frac{r-x}{N-n}\cdot$$

Ainsi

$$\begin{aligned}
\mathbb{P}(E \,|\, X = x) &= \sum_{r=x}^{N} \frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N+1}{n+1}} \cdot \frac{r-x}{N-n} \\
&= \frac{(x+1)}{(N-n)\binom{N+1}{n+1}} \sum_{r=x}^{N} \binom{r}{x+1}\binom{N-r}{n-x} \\
&= \frac{x+1}{N-n} \cdot \frac{\binom{N+1}{n+2}}{\binom{N+1}{n+1}} \\
&= \frac{x+1}{n+2}\cdot
\end{aligned}$$

Puisque cette formule est valable pour tout $x \leq N$, on a donc

$$\mathbb{P}(E \,|\, X) = \frac{X+1}{n+2}\cdot$$

Exercice 11 (Identifiabilité)

1. Soit $(\varepsilon_i)_{i\geq 1}$ une suite de variables aléatoires i.i.d. de loi de Bernoulli $\mathcal{B}(p)$. Alors la variable Y a la même loi que

$$\sum_{i=1}^{X} \varepsilon_i.$$

On a que $\mathcal{L}(X) = \mathcal{P}(\lambda)$ et X et les ε_i sont indépendantes.

(a) Les variables aléatoires X et Y sont à valeurs dans $\mathbb{N}$, avec par construction $Y \leq X$. De plus, pour tous $0 \leq y \leq x$,

$$\begin{aligned}
\mathbb{P}(Y = y \,|\, X = x) &= \frac{\mathbb{P}(Y = y \text{ et } X = x)}{\mathbb{P}(X = x)} \\
&= \frac{\mathbb{P}\left(\sum_{i=1}^{x} \varepsilon_i = y \text{ et } X = x\right)}{\mathbb{P}(X = x)} \\
&= \frac{\mathbb{P}\left(\sum_{i=1}^{x} \varepsilon_i = y\right)\mathbb{P}(X = x)}{\mathbb{P}(X = x)} = \mathbb{P}\left(\sum_{i=1}^{x} \varepsilon_i = y\right),
\end{aligned}$$

où l'on a utilisé l'indépendance de X et des ε_i. Or $\sum_{i=1}^{x} \varepsilon_i$ suit une loi binomiale $\mathcal{B}(x,p)$, donc la loi $\mathcal{L}(Y \,|\, X = x)$ est une $\mathcal{B}(x,p)$, i.e.

$$\mathbb{P}(Y = y \,|\, X = x) = \mathbb{P}(\mathcal{B}(x,p) = y) = \binom{x}{y} p^y (1-p)^{x-y}.$$

(b) Le modèle statistique s'écrit $\mathcal{P} = \{P_{\lambda,p},\ (\lambda,p) \in \mathbb{R}_*^+ \times]0,1[\}$, où $P_{\lambda,p}$ est la loi du couple (X,Y). Pour déterminer cette loi, il suffit de se donner $\mathbb{P}(X = x, Y = y)$ pour tous $x, y \in \mathbb{N}$. On a

$$\begin{aligned}\mathbb{P}(X = x, Y = y) &= \mathbb{P}(X = x)\mathbb{P}(Y = y \,\big|\, X = x) \\ &= \mathbb{P}(\mathcal{P}(\lambda) = x)\mathbb{P}(\mathcal{B}(x,p) = y) \\ &= \binom{x}{y} p^y (1-p)^{x-y} e^{-\lambda} \frac{\lambda^x}{x!}.\end{aligned}$$

On peut aussi noter que se donner la loi jointe de (X,Y) est équivalent à se donner les lois $\mathcal{L}(Y \,\big|\, X)$ et $\mathcal{L}(X)$, ce qui revient à la première identité de la série d'égalités précédente.

Montrons que ce modèle est identifiable. Supposons $P_{\lambda_1,p_1} = P_{\lambda_2,p_2}$. Alors si X, Y suivent cette même loi, on a $\mathbb{E}_{\lambda_1,p_1}[X] = \mathbb{E}[\mathcal{P}(\lambda_1)] = \lambda_1$ et de même $\mathbb{E}_{\lambda_2,p_2}[X] = \mathbb{E}[\mathcal{P}(\lambda_2)] = \lambda_2$. Les deux lois étant les mêmes, les espérances sont égales soit $\lambda_1 = \lambda_2$. Puis notons que

$$\begin{aligned}\mathbb{E}_{\lambda_1,p_1}[Y] &= \mathbb{E}_{\lambda_1,p_1}\left[\mathbb{E}_{\lambda_1,p_1}[Y \,\big|\, X]\right] \\ &= \mathbb{E}_{\lambda_1,p_1}\left[\mathbb{E}_{\lambda_1,p_1}[\mathcal{B}(X,p_1) \,\big|\, X]\right] = \mathbb{E}_{\lambda_1,p_1}[Xp_1] = \lambda_1 p_1.\end{aligned}$$

En écrivant ceci sous P_{λ_2,p_2} et en utilisant le fait que les espérances sont égales on obtient $\lambda_1 p_1 = \lambda_2 p_2$. Donc $p_1 = p_2$ car $\lambda_1 = \lambda_2 > 0$. Le modèle est donc identifiable.

(c) Puisque $\mathbb{E}_{\lambda,p}[X] = \lambda$ et $\mathbb{E}_{\lambda,p}[Y] = \lambda p$, la méthode des moments incite à proposer

$$\begin{aligned}\widehat{\lambda}_n &= \frac{1}{n}\sum_{i=1}^n X_i = \overline{X}_n \\ \widehat{p}_n &= \frac{\overline{Y}_n}{\overline{X}_n} \mathbb{1}_{\overline{X}_n > 0}.\end{aligned}$$

En effet la LFGN implique que $\widehat{\lambda}_n$ converge p.s. vers $\mathbb{E}_{\lambda,p}[X] = \lambda > 0$ et que $\overline{Y}_n$ converge p.s. vers $\mathbb{E}_{\lambda,p}[Y] = \lambda p$. Ceci assure que $\widehat{p}_n$ converge p.s. vers $\lambda p/\lambda = p$.

Remarque : la méthode du maximum de vraisemblance mène aux mêmes estimateurs. En effet, la vraisemblance de l'échantillon $(X_1, Y_1), \ldots, (X_n, Y_n)$ en (λ, p) vaut

$$L_n(\lambda,p) = \prod_{i=1}^n \binom{X_i}{Y_i} p^{Y_i}(1-p)^{X_i - Y_i} e^{-\lambda}\frac{\lambda^{X_i}}{X_i!} = Cp^{n\overline{Y}_n}(1-p)^{n(\overline{X}_n - \overline{Y}_n)} e^{-n\lambda}\lambda^{n\overline{X}_n},$$

où C ne dépend ni de λ ni de p. La log-vraisemblance s'en déduit :

$$\ell_n(\lambda,p) = \log C + n\overline{Y}_n \log p + n(\overline{X}_n - \overline{Y}_n)\log(1-p) - n\lambda + n\overline{X}_n \log\lambda,$$

de dérivées partielles respectives

$$\frac{\partial \ell_n}{\partial \lambda}(\lambda,p) = n\left(\frac{\overline{X}_n}{\lambda} - 1\right),$$

et

$$\frac{\partial \ell_n}{\partial p}(\lambda,p) = n\frac{\overline{Y}_n - p\overline{X}_n}{p(1-p)}.$$

On peut noter que si $\overline{X}_n = 0$, alors $\overline{Y}_n = 0$ (par construction du modèle, puisqu'on a toujours $0 \le \overline{Y}_n \le \overline{X}_n$), auquel cas la log-vraisemblance vaut $\log C - n\lambda$, maximale en $(\widehat{\lambda}_n, \widehat{p}_n) = (0, p)$ pour tout $0 < p < 1$. Noter que, en toute rigueur, ce n'est pas un estimateur du maximum de vraisemblance car 0 n'est pas une valeur possible de λ pour une loi de Poisson. Si $\overline{X}_n > 0$ (ce qui arrive p.s. asymptotiquement par la LFGN), l'unique maximum est obtenu en annulant les dérivées partielles ci-dessus, et l'on retrouve le couple $(\widehat{\lambda}_n, \widehat{p}_n)$ de la méthode des moments.

2. (a) Pour tout $y \in \mathbb{N}$, on a

$$\begin{aligned}
\mathbb{P}(Y=y) &= \sum_{x=y}^{\infty} \mathbb{P}(X=x\,,\,Y=y) = \sum_{x=y}^{\infty} \mathbb{P}(X=x)\mathbb{P}(Y=y \,\big|\, X=x) \\
&= \sum_{x=y}^{\infty} \binom{x}{y} p^y (1-p)^{x-y} e^{-\lambda} \frac{\lambda^x}{x!} \\
&= e^{-\lambda} \frac{(\lambda p)^y}{y!} \sum_{x=y}^{\infty} \frac{(1-p)^{x-y}\lambda^{x-y}}{(x-y)!} \\
&= e^{-\lambda} \frac{(\lambda p)^y}{y!} \sum_{k=0}^{\infty} \frac{(1-p)^k \lambda^k}{k!} \\
&= e^{-\lambda} \frac{(\lambda p)^y}{y!} e^{\lambda(1-p)} = e^{-\lambda p} \frac{(\lambda p)^y}{y!} = \mathbb{P}(\mathcal{P}(\lambda p) = y).
\end{aligned}$$

Ainsi la loi de Y est une loi de Poisson $\mathcal{P}(\lambda p)$.

(b) Si le paramètre d'intérêt est toujours le couple (λ, p), le modèle statistique associé s'écrit $\mathcal{P} = \{Q_{\lambda,p},\ (\lambda,p) \in \mathbb{R}_*^+ \times]0,1[\}$, où $Q_{\lambda,p}$ est la loi de Y décrite ci-dessus, c'est-à-dire la loi de Poisson de paramètre λp. Ce modèle n'est pas identifiable. En effet, il suffit de choisir $\lambda_1, \lambda_2, p_1, p_2$ avec $\lambda_1 \neq \lambda_2$ mais $\lambda_1 p_1 = \lambda_2 p_2$ (par exemple $\lambda_1 = 2, p_1 = 1/4$ et $\lambda_2 = 1, p_1 = 1/2$), car alors la loi de Y sous chacune de ces lois est toujours $\mathcal{P}(\lambda_1 p_1) = \mathcal{P}(\lambda_2 p_2)$.

Exercice 12 (A priori de Jeffreys)

1. Dans un modèle régulier, pour $\Theta \subset \mathbb{R}$ ouvert, l'a priori de Jeffreys est la mesure sur Θ de densité $\pi(\theta)$ par rapport à $\nu = \mathrm{Leb}_{|\Theta}$ proportionnelle à $\sqrt{I(\theta)}$, où $I(\theta) = \mathbb{E}_\theta\left[\ell_\theta'(X)^2\right]$ (en prenant $X \sim P_\theta$) est l'information de Fisher en θ. Autrement dit,

$$\pi(\theta) = \frac{\sqrt{I(\theta)}}{\int_\Theta \sqrt{I(\theta)}\,\mathrm{d}\theta} \qquad \text{si l'a priori est propre, i.e. } \int_\Theta \sqrt{I(\theta)}\,\mathrm{d}\theta < +\infty.$$

$$\pi(\theta) = \sqrt{I(\theta)} \qquad \text{si l'a priori est impropre, i.e. } \int_\Theta \sqrt{I(\theta)}\,\mathrm{d}\theta = +\infty.$$

2. (a) Dans le modèle gaussien, pour tout $x \in \mathbb{R}$, la fonction $\theta \mapsto p_\theta(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\theta)^2}{2}}$ est $\mathcal{C}^\infty$ et l'on a, pour tout $\theta \in \mathbb{R}$, en prenant $X \sim \mathcal{N}(\theta, 1)$,

$$\ell_\theta(X) = -\frac{1}{2}\ln(2\pi) - \frac{1}{2}(X-\theta)^2\,,$$

et en dérivant par rapport à θ,

$$\ell_\theta'(X) = (X-\theta)\,.$$

Cette variable possède bien un moment d'ordre 2 :

$$\mathbb{E}_\theta\left[\ell_\theta'(X)^2\right] = \mathrm{Var}_\theta(X) = 1\,.$$

L'a priori de Jeffreys est impropre : il s'agit de la mesure de Lebesgue sur $\mathbb{R}$.

(b) Pour tout $x \in \{0,1\}$, la fonction $\theta \mapsto \theta^x(1-\theta)^{1-x}$ est $\mathcal{C}^\infty$, et pour tout $\theta \in]0,1[$, en prenant $X \sim \mathcal{B}(\theta)$,

$$\ell_\theta(X) = X\ln(\theta) + (1-X)\ln(1-\theta)\,,$$

et en dérivant par rapport à θ,

$$\ell'_\theta(X) = \frac{X}{\theta} - \frac{1-X}{1-\theta} = \frac{X-\theta}{\theta(1-\theta)}\,.$$

Cette variable possède bien un moment d'ordre 2 :

$$\mathbb{E}_\theta\left[\ell'_\theta(X)^2\right] = \frac{1}{\theta^2(1-\theta)^2}\,\mathrm{Var}_\theta(X) = \frac{1}{\theta(1-\theta)}\,.$$

L'a priori de Jeffreys est la loi de densité proportionnelle à $\theta \mapsto \theta^{-1/2}(1-\theta)^{-1/2}$ par rapport à la mesure de Lebesgue sur $]0,1[$. Il s'agit de la loi Beta$(1/2,1/2)$.

(c) Pour tout $x \in \mathbb{N}$, la fonction $\theta \mapsto \mathrm{e}^{-\theta}\frac{\theta^x}{x!}$ est $\mathcal{C}^\infty$, et pour tout $\theta \in \mathbb{R}_+^*$, en prenant $X \sim \mathcal{P}(\theta)$,

$$\ell_\theta(X) = -\theta + X\ln(\theta) - \ln(X!)\,,$$

et en dérivant par rapport à θ,

$$\ell'_\theta(X) = -1 + \frac{X}{\theta}\,.$$

Cette variable possède bien un moment d'ordre 2 :

$$\mathbb{E}_\theta\left[\ell'_\theta(X)^2\right] = \frac{1}{\theta^2}\,\mathrm{Var}_\theta(X) = \frac{1}{\theta}\,.$$

L'a priori de Jeffreys est l'a priori impropre de densité $\theta \mapsto \theta^{-1/2}$ par rapport à la mesure de Lebesgue sur $\mathbb{R}_+^*$.