Master 1 – Mathématiques et Applications

Statistique computationnelle

Tabea Rebafka

MU4MA074 – Partie II

Sorbonne Université

2022

Table des matières

Chapitre 1

Introduction à la statistique

Dans ce chapitre nous donnons une introduction rapide à la statistique en présentant l'approche générale et des notions et méthodes fondamentales de la statistique. Pour une introduction plus complète nous référons au polycopié de Guyader (2017) ou aux livres de statistique de Lejeune (2004) ou de Rivoirard and Stoltz (2012). Une présentation des outils de base de la statistique descriptive se trouve dans le polycopié de Rebafka (2017). Une bonne référence pour le logiciel `R` est le livre de Lafaye de Micheaux et al. (2010)

L'objet principal de la *statistique* sont des données ou observations. Les *données* sont issues de domaines très variés comme la médecine, l'économie, la sociologie, l'ingénierie, l'astrophysique, l'internet etc. L'objectif des statisticien·nes est d'extraire des informations utiles des données, les analyser et interpréter pour des objectifs concrets comme le contrôle de qualité, l'aide à la décision ou la prédiction.

Depuis quelques années, le terme *statisticien·ne* est remplacé par celui du *data scientist*. La *science des données* est bien l'analyse et l'extraction utile de données ainsi que la prédiction. La science des données se confond également avec l'*apprentissage statistique* (le *machine learning* en anglais) et l'*intelligence artificielle*. Dans toutes ces disciplines, très à la mode, dont les frontières sont difficiles à définir, une grande partie des méthodes et approches reposent sur des techniques statistiques.

L'approche statistique consiste à se donner un cadre mathématique, dans lequel la variabilité dans les données est expliquée par l'aléa. On adopte donc une *modélisation probabiliste* des données. On souligne qu'il n'est pas indispensable que le phénomène observé soit vraiment de nature aléatoire, c'est-à-dire les données soient issue d'une expérience où intervient le hasard. La modélisation probabiliste n'est que le moyen pour prendre en compte la variabilité dans les données, et on doit toujours justifier et critiquer le choix d'un modèle. Par ailleurs, il est clair que tout modèle est faux, car il ne peut être qu'une approximation de la réalité. Néanmoins, on espère que le modèle choisi est approprié pour apporter des réponses en vue des objectifs concrets de l'application.

Plus précisément, notons $\mathbf{x} = (x_1, \dots, x_n)$ les observations ou le jeu de données, qui est typiquement une série de valeurs numériques. En statistique, on considère $\mathbf{x}$ comme la réalisation d'une variable aléatoire $\mathbf{X}$ de loi $\mathbb{P}$, qui est une loi de probabilité inconnue.

On peut décrire la démarche statistique par trois étapes.

1. *Introduction d'un modèle statistique.* Grâce à la connaissance *a priori* du phénomène observé et à l'aide des outils de la statistique descriptive (histogrammes, QQ-plot,...), le statisticien choisit ce qu'on appelle un *modèle statistique* : la donnée d'une famille de lois de probabilité $\mathcal{P}$ considérée comme une famille de lois candidates pour $\mathbb{P}$. Voir le polycopié de Rebafka (2017) pour une présentation des outils classiques de la statistique descriptive.

TABLE 1.1 – Nombre de lynx attrapés par an au Canada entre 1821 et 1934. Source : `R` package `datasets`.

1821	1822	1823	1824	1825	1826	1827	1828	1829	1830	1831	1832	1833	1834	1835
269	321	585	871	1475	2821	3928	5943	4950	2577	523	98	184	279	409
1836	1837	1838	1839	1840	1841	1842	1843	1844	1845	1846	1847	1848	1849	1850
2285	2685	3409	1824	409	151	45	68	213	546	1033	2129	2536	957	361
1851	1852	1853	1854	1855	1856	1857	1858	1859	1860	1861	1862	1863	1864	1865
377	225	360	731	1638	2725	2871	2119	684	299	236	245	552	1623	3311
1866	1867	1868	1869	1870	1871	1872	1873	1874	1875	1876	1877	1878	1879	1880
6721	4254	687	255	473	358	784	1594	1676	2251	1426	756	299	201	229
1881	1882	1883	1884	1885	1886	1887	1888	1889	1890	1891	1892	1893	1894	1895
469	736	2042	2811	4431	2511	389	73	39	49	59	188	377	1292	4031
1896	1897	1898	1899	1900	1901	1902	1903	1904	1905	1906	1907	1908	1909	1910
3495	587	105	153	387	758	1307	3465	6991	6313	3794	1836	345	382	808
1911	1912	1913	1914	1915	1916	1917	1918	1919	1920	1921	1922	1923	1924	1925
1388	2713	3800	3091	2985	3790	674	81	80	108	229	399	1132	2432	3574
1926	1927	1928	1929	1930	1931	1932	1933	1934						
2935	1537	529	485	662	1000	1590	2657	3396						

Il est courant, et utile, de paramétrer cette famille de lois : on écrit $\mathcal{P} = \{\mathbb{P}_\theta, \theta \in \Theta\}$. Il existe alors un paramètre $\theta_0 \in \Theta$ tel que $\mathbb{P} = \mathbb{P}_{\theta_0}$. On dit que θ_0 est la *vraie valeur du paramètre* θ.

2. *Estimation de paramètre.* En utilisant les données $\mathbf{x}$, on cherche à déterminer la loi $\mathbb{P}_{\theta_0}$ des données, ce qui revient à estimer le paramètre θ_0. Différentes méthodes d'estimation de paramètre existent dans la littérature, comme l'approche de maximum de vraisemblance, la méthode des moments ou la méthode de substitution. On note typiquement $\hat{\theta}$ un estimateur de θ_0.

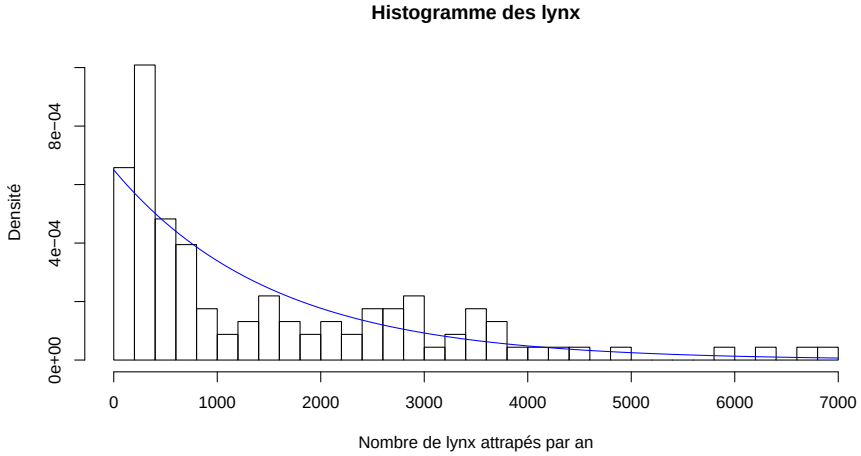
   Nous insistons sur le fait que le paramètre θ_0 est inconnu alors que l'ensemble Θ des paramètres possibles est connu, car la famille de lois $\{\mathbb{P}_\theta, \theta \in \Theta\}$ est connue. La démarche statistique consiste à extraire de l'information sur le paramètre θ_0 en s'appuyant sur l'observation $\mathbf{x}$; on parle d'inférence et de *statistique inférentielle.*

3. *Interprétation.* Enfin, il est question d'interpréter les résultats et de faire des conclusions du fait qu'on estime la loi $\mathbb{P}$ des données par une loi $\mathbb{P}_{\hat{\theta}} \in \mathcal{P}$. On peut quantifier l'incertitude de l'estimation de θ par des intervalles de confiance, effectuer des tests statistiques pour répondre à des questions d'intérêt pratique ou faire de la prédiction.

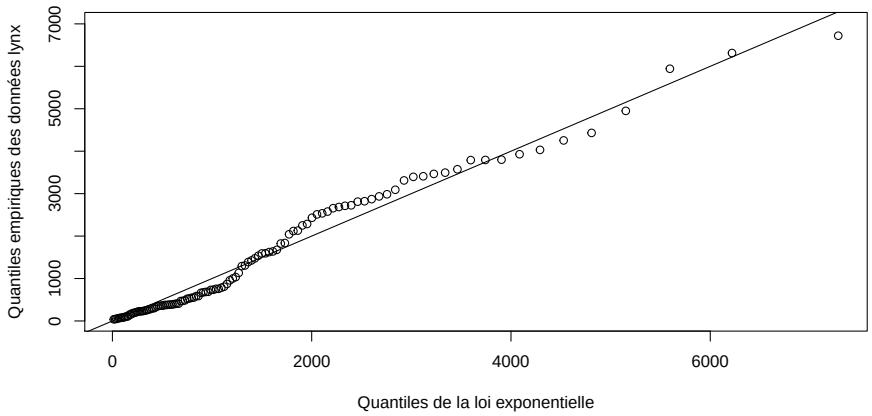
EXEMPLE. DONNÉES LYNX.

Le package `datasets` de `R` contient un jeu de données sur le nombre de lynx attrapés par an au Canada entre 1821 et 1934. La Table 1.1 montre le tableau des données, et son histogramme est donné dans la Figure 1.1 a). Quel modèle statistique choisir pour ces données ? La forme de l'histogramme a une tendance décroissante, les valeurs des observations sont toutes positives, cela nous fait penser à une loi exponentielle. Supposons donc que les observations $(x_1, \dots, x_n)$ avec $n = 114$ sont la réalisations de variables aléatoires $X_1, \dots, X_n$ indépendantes de loi exponentielle $\mathcal{E}(\theta)$ avec $\theta > 0$ inconnu.

Comment estimer θ ? Une approche courante, dite méthode des moments, pour des observations i.i.d. consiste à utiliser les moments de la loi des observations pour identifier la valeur du paramètre. Plus précisément, ici on a $\mathbb{E}_\theta[X] = \frac{1}{\theta}$ pour $X \sim \mathcal{E}(\theta)$ et pour tout $\theta > 0$. Par ailleurs, par la loi des grands nombres, on a $\bar{X}_n := \frac{1}{n}\sum_{i=1}^n X_i \approx \mathbb{E}_\theta[X_1]$ pour n assez grand. La méthode des moments consiste à poser l'équation $\bar{X}_n = \mathbb{E}_\theta[X_1]$, dont la



a)



b)

FIGURE 1.1 – Histogramme a) et QQ-plot b) pour les données des nombres de lynx attrapés par an au Canada de 1821 à 1934. Comparaison à la loi exponentielle $\mathcal{E}(\hat{\theta})$ avec $\hat{\theta} = 0.0006501876$.

solution est $\theta = \frac{1}{\bar{X}_n}$. Alors, un estimateur de θ est donné par $\hat{\theta} = 1/\bar{X}_n$. Pour les lynx on trouve $\hat{\theta} = 0.0006501876$. Donc, on modélise le nombre de lynx attrapé par an par la loi exponentielle $\mathcal{E}(0.0006501876)$. La densité de cette loi est superposé à l'histogramme de la Figure 1.1. On voit que l'approximation de l'histogramme par cette densité n'est pas mal.

Afin d'avoir plus de certitude sur le choix de la loi exponentielle $\mathcal{E}(0.0006501876)$ pour les données des lynx, on peut tracer un diagramme quantile-quantile ou *QQ-plot*. Un QQ-plot compare la loi empirique d'un jeu de données $\mathbf{x} = (x_1, \ldots, x_n)$ à une loi théorique F_0. C'est le nuage des points de coordonnées $(\hat{q}^{\mathbf{x}}_{j/n}, q^{F_0}_{j/n})$ pour $j = 1, \ldots, n$, où $q^{F_0}_{j/n} = F_0^{-1}(j/n)$ désigne le quantile (théorique) d'ordre j/n de la loi F_0 et $\hat{q}^{\mathbf{x}}_{j/n}$ le quantile empirique d'ordre j/n associé à $\mathbf{x}$. Plus précisément, on a $\hat{q}^{\mathbf{x}}_{j/n} = x_{(j)}$ est la j-ième statistique d'ordre associé à $\mathbf{x}$ (la j-ième plus petite valeur parmi $(x_1, \ldots, x_n)$). L'interprétation d'un QQ-plot est simple : si et seulement si les points du QQ-plot s'alignent sur la première bissectrice, la loi de l'échantillon $\mathbf{x}$ observé est la loi F_0 (pour plus de détails sur le QQ-plot voir Rebafka (2017)). Dans notre exemple, d'après la Figure 1.1 b), on considère que l'alignement des point sur la première bissectrice est plutôt bon (mais pas parfait) et on conclut que la loi exponentielle $\mathcal{E}(0.0006501876)$ modélise assez bien les données des lynx.

On peut pousser cette analyse plus loin en cherchant à quantifier l'incertitude de l'estimation du paramètre θ de la loi exponentielle par un intervalle de confiance. Ou encore on peut mettre en question le caractère i.i.d. des observations. En effet, les données forment une série temporelle et donc il peut y avoir des effets de temps qui font que la loi n'est pas la même durant toute la période d'observation. Une idée simple serait de couper l'échantillon en deux parties, ajuster une loi exponentielle à chaque période et ensuite effectuer un test statistique pour savoir si les deux paramètres exponentiels sont identiques.

Dans ce chapitre nous présentons des notions et méthodes fondamentales de la statistique.

1.1 Estimation ponctuelle

Tout au long de ce chapitre, nous considérons un modèle statistique $\mathcal{P} = \{\mathbb{P}_\theta, \theta \in \Theta\}$. Si $\Theta \subset \mathbb{R}^d$, autrement dit, le paramètre θ est un vecteur de dimension d, le modèle est dit *paramétrique*, à l'opposé des modèles *nonparamétriques* où la dimension de θ n'est pas finie. Dans le cas nonparamétrique θ est typiquement une fonction et Θ est un grand ensemble de fonctions. Un exemple est l'estimation de la fonction de répartition F sans aucune contrainte sur le forme de la loi, et dans ce cas, un estimateur nonparamétrique est la fonction de répartition empirique $\hat{F}$. Dans ce cours, nous considérons surtout le problème d'estimation de paramètre θ de dimension finie (ou d'une quantité $q(\theta)$ de dimension finie).

On appelle *estimateur* de θ toute fonction mesurable $\hat{\theta} = \hat{\theta}(\mathbf{X})$ à valeurs dans Θ définie sur les données $\mathbf{X}$.

1.1.1 Maximum de vraisemblance

Quelques cas particuliers de la méthode du maximum de vraisemblance sont connus depuis le XVIIIème siècle, mais sa définition générale et l'argumentation de son rôle fondamental en statistique sont dues à Ronald Fisher (1922).

Intuition

Pour comprendre l'intuition de la méthode du maximum de vraisemblance (MV) considérons le problème d'une pièce de monnaie et la question si cette pièce est équilibrée ou pas. Autrement dit, on veut savoir, en jouant à pile ou face, quelle est la probabilité d'obtenir 'pile'. On modélise les sorties d'une suite de n lancées $\mathbf{x} = (x_1, \dots, x_n)$ par une expérience Bernoulli, i.e. $\mathbf{x}$ est la réalisation d'un vecteur aléatoire $\mathbf{X} = (X_1, \dots, X_n)$ de v.a. X_i i.i.d. de loi Bernoulli de paramètre $p \in (0, 1)$ (et on identifie l'événement 'pile' avec 1, et 'face' avec 0). Si la pièce de monnaie est équilibrée p vaut $1/2$.

L'approche de MV consiste à étudier la fonction de vraisemblance $\mathcal{L}(\mathbf{x}; p)$ définie par

$$p \mapsto \mathcal{L}(\mathbf{x}; p) = \mathbb{P}_p(X_1 = x_1, \dots, X_n = x_n) = \prod_{i=1}^{n} \mathbb{P}_p(X_i = x_i),$$

où la notation $\mathbb{P}_p$ signifie que les v.a. X_i suivent la loi Bernoulli de paramètre p.

Notons que $\mathcal{L}(\mathbf{x}; p)$ correspond à la probabilité d'obtenir la suite $\mathbf{x}$ lorsque la vraie valeur du paramètre de la loi de Bernoulli est p. Or, la méthode de MV cherche la valeur de p qui maximise cette probabilité. Autrement dit, on cherche le paramètre qui rend cette suite d'observations le plus vraisemblable.

Soyons encore plus concret. Supposons que l'on observe $\mathbf{x} = (1, 1, 1, 1, 0, 1, 1, 1)$. Clairement, avec une pièce équilibrée, il est très peu probable d'obtenir la suite observée (même s'il n'est pas impossible). En revanche, avec une pièce fortement déséquilibrée, il est bien plus vraisemblable d'observer cette suite de valeur. En fait, pour cet exemple, on a $\mathcal{L}(\mathbf{x}; p) = p^7(1-p)$ et on montre facilement que cette fonction est maximale en $p = 7/8$. Donc, c'est avec une pièce de monnaie dont la probabilité d'observer pile est de $7/8$, qu'on a la plus de chance d'obtenir la suite $\mathbf{x} = (1, 1, 1, 1, 0, 1, 1, 1)$. La méthode de MV propose de considérer $\hat{p}^{MV} = 7/8$ comme estimateur de p.

La méthode en général

Considérons un modèle statistique $\{\mathbb{P}_\theta, \theta \in \Theta\}$, où $\Theta \subset \mathbb{R}^d$ et un échantillon $\mathbf{x} = (x_1, \dots, x_n)$ de loi $\mathbb{P}_{\theta_0}$. Supposons que le modèle soit dominée par une mesure μ, et notons $p_\theta = \frac{\mathrm{d}\mathbb{P}_\theta}{\mathrm{d}\mu}$ la densité de $\mathbb{P}_\theta$ par rapport à μ. Le plus souvent, la mesure dominante μ est soit la mesure de Lebesgue et on note $f_\theta = p_\theta = \frac{\mathrm{d}\mathbb{P}_\theta}{\mathrm{d}\mu}$, soit une mesure de comptage auquel cas on a $p_\theta(\mathbf{x}) = \mathbb{P}_\theta(\mathbf{X} = \mathbf{x})$.

Définissons la **fonction de vraisemblance** de $\mathbf{x}$ par

$$\theta \mapsto \mathcal{L}(\mathbf{x}; \theta) = p_\theta(\mathbf{x}).$$

Si $\mathbf{x}$ est un échantillon i.i.d. de densité p_{θ_0}, alors $\mathcal{L}(\mathbf{x}; \theta) = \prod_{i=1}^n p_\theta(x_i)$. Dans le cas i.i.d. continu, on a $\mathcal{L}(\mathbf{x}; \theta) = \prod_{i=1}^n f_\theta(x_i)$, et dans le cas i.i.d. discret, on a $\mathcal{L}(\mathbf{x}; \theta) = \prod_{i=1}^n \mathbb{P}_\theta(X_i = x_i)$.

On appelle **estimateur du maximum de vraisemblance (EMV)** du paramètre θ_0 dans le modèle statistique $\{P_\theta, \theta \in \Theta\}$ toute statistique $\hat{\theta}^{MV} \in \Theta$ telle que

$$\mathcal{L}(\mathbf{x}; \hat{\theta}^{MV}) = \max_{\theta \in \Theta} \mathcal{L}(\mathbf{x}; \theta). \tag{1.1}$$

Autrement dit, $\hat{\theta}^{MV}$ est tel que

$$\hat{\theta}^{MV} = \arg\max_{\theta \in \Theta} \mathcal{L}(\mathbf{x}; \theta).$$

L'EMV peut ne pas exister, car le problème de maximisation (1.1) n'admet pas de solution (dans Θ). Ou encore, l'EMV existe, mais il n'est pas unique si (1.1) admet plusieurs solutions.

Si le support des densités $x \mapsto f_\theta(x)$ ne dépend pas de θ (c'est-à-dire l'ensemble $\{x : f_\theta(x) > 0\}$ est le même pour tout $\theta \in \Theta$), on définit la **fonction de log-vraisemblance** $\ell(\theta)$ par

$$\ell(\theta) = \log(\mathcal{L}(\mathbf{x}; \theta)).$$

Remarquons que

$$\hat{\theta}^{MV} = \arg\max_{\theta \in \Theta} \ell(\theta),$$

car la fonction $t \mapsto \log(t)$ est strictement croissante. Par conséquent, au lieu de maximiser la fonction de vraisemblance $\mathcal{L}(\mathbf{x}; \theta)$, on peut aussi bien maximiser la fonction de log-vraisemblance $\ell(\theta)$ pour trouver l'EMV, ce qui s'avère en général beaucoup plus facile.

Si le maximum de $\mathcal{L}(\mathbf{x}; \theta)$ (ou de $\ell(\theta)$) n'est pas atteint sur la frontière de Θ et si l'application $\theta \mapsto \mathcal{L}(\mathbf{x}; \theta)$ est différentiable, une condition nécessaire de maximum est l'annulation du gradient :

$$\nabla_\theta \mathcal{L}(\mathbf{x}; \theta)|_{\theta = \hat{\theta}^{MV}} = 0, \tag{1.2}$$

ce qui représente un système de d équations, car $\theta \in \mathbb{R}^d$. De façon similaire, une condition nécessaire de maximum de la fonction de log-vraisemblance est

$$\nabla \ell(\theta) = 0. \tag{1.3}$$

On appelle (1.3) l'**équation de vraisemblance** si $\theta \in \mathbb{R}$ et **système des équations de vraisemblance** si $\theta \in \mathbb{R}^d, d > 1$.

On appelle **racine de l'équation de vraisemblance** (REV) dans le modèle $\{\mathbb{P}_\theta, \theta \in \Theta\}$, avec $\Theta \in \mathbb{R}^d$, toute statistique $\hat{\theta}^{RV}$ à valeurs dans Θ solution du système de d équations (1.3). Autrement dit,

$$\nabla \ell(\hat{\theta}^{RV}) = 0.$$

Notons qu'en résolvant le système (1.3) on obtient tous les maxima et tous les minima locaux de $\ell(\cdot)$, ainsi que ses points d'inflexion. Il est clair que la REV peut ne pas exister et, si elle existe, elle n'est pas toujours unique.

Pour que tous les EMV soient des REV et vice versa, il faut essentiellement que la fonction $\ell(\cdot)$ atteint son minimum global pour tous les θ tels que $\nabla \ell(\theta) = 0$. Cette condition est très restrictive : on ne peut effectivement la vérifier que si la fonction ℓ est convexe et son minimum global n'est pas atteint sur la frontière de Θ. L'équivalence des EMV et des REV n'a donc lieu que dans une situation très particulière. Il s'agit essentiellement de deux estimateurs différents, sauf cas exceptionnel.

Exemple : Loi exponentielle

Soit $\mathbf{x} = (x_1, \ldots, x_n)$ un échantillon i.i.d. de loi exponentielle $\mathcal{E}(\theta)$ de densité de probabilité $f_\theta(x) = \theta \mathrm{e}^{-\theta x}$ pour $x > 0$ et avec $\theta > 0$ inconnu. La fonction de vraisemblance s'écrit

$$\mathcal{L}(\mathbf{x}; \theta) = \prod_{i=1}^n f_\theta(x_i) = \prod_{i=1}^n \theta \mathrm{e}^{-\theta x_i)} = \theta^n \exp\left\{ -\sum_{i=1}^n x_i \right\}.$$

Pour la fonction de log-vraisemblance on obtient

$$\ell(\theta) = \log(L(\mathbf{x}; \theta)) = n \log \theta - \theta \sum_{i=1}^n x_i.$$

Les dérivées sont

$$\ell'(\theta) = \frac{n}{\theta} - \sum_{i=1}^{n} x_i, \qquad \ell''(\theta) = -\frac{n}{\theta^2}.$$

On observe que $\ell''(\theta) < 0$ pour tout $\theta > 0$. Donc, $\ell(\theta)$ est strictement concave. Pour le point critique on obtient

$$\begin{aligned} \ell'(\theta) = 0 &\iff \frac{n}{\theta} - \sum_{i=1}^{n} x_i = 0 \\ &\iff \theta = \frac{1}{\bar{x}_n}. \end{aligned}$$

Nous avons donc montré que $\ell(\theta)$ atteint son maximum en $\theta = \frac{1}{\bar{x}_n}$. Donc l'EMV de θ est donné par $\hat{\theta} = \frac{1}{\bar{x}_n}$, et cet estimateur est unique.

1.1.2 Méthode de substitution ou de plug-in

Dans ce paragraphe, nous supposons que l'on observe un échantillon i.i.d. $\mathbf{x} = (x_1, \ldots, x_n)$, c'est-à-dire les x_i sont des réalisations indépendantes d'une variable aléatoire X de loi $\mathbb{P}_\theta$ appartenant à un modèle statistique $\{\mathbb{P}_\theta, \theta \in \Theta\}$. Notons que nous n'avons pas fait cette hypothèse pour l'EMV.

La méthode de substitution est une approche très générale pour estimer le paramètre θ ou plus généralement une caractéristique $q(\theta)$ de la loi $\mathbb{P}_\theta$. L'idée principale est d'approcher des quantités théoriques de la loi $\mathbb{P}_\theta$ par leurs équivalents empiriques observés sur les données.

En fait, nous l'avons déjà utilisée dans l'exemple sur les données des lynx pour estimer le paramètre de la loi exponentielle en utilisant son espérance. En effet, en général le paramètre θ peut s'écrire comme une fonctionnelle de sa fonction de répartition F_θ :

$$\theta = T(F_\theta).$$

Par exemple, si $X \sim \mathcal{E}(\theta)$, alors $\theta = 1/\mathbb{E}_\theta[X] = 1/\int x \mathrm{d}F(x) = T(F)$.

Or, on obtient un estimateur $\hat{\theta}$ de $\theta = T(F)$ en remplaçant la loi F par une loi approchée $\hat{F}$ calculée sur les données $\mathbf{x}$ par

$$\hat{\theta} = T(\hat{F}).$$

Pour $\hat{F}$ on utilise la loi empirique, c'est-à-dire la fonction de répartition empirique associée à $\mathbf{x}$.

L'idée de construction de cet estimateur est appelée **méthode de substitution** ou **principe du plug-in** : on substitue $\hat{F}$ à F.

La **fonction de répartition empirique** $\hat{F}$ (ou $\hat{F}_n$) associée à l'échantillon $\mathbf{x} = (x_1, \ldots, x_n)$ est définie par

$$\hat{F}(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{x_i \leq t\} = \frac{\#\{i : x_i \leq t\}}{n}, \quad t \in \mathbb{R}. \tag{1.4}$$

Il est simple de voir que la fonction de répartition empirique $\hat{F}$ a les propriétés suivantes : $\hat{F}$ est une fonction définie sur tout $\mathbb{R}$, elle est croissante et continue à droite. Elle prend ses valeurs dans $[0, 1]$ et vérifie

$$\hat{F}(t) = 0, \forall t < \min\{x_1, \ldots, x_n\} \qquad \text{et} \qquad \hat{F}(t) = 1, \forall t > \max\{x_1, \ldots, x_n\}.$$

Plus précisément, $\hat{F}$ est une fonction en escalier avec des sauts en x_i. En fait, $\hat{F}$ est la fonction de répartition d'une loi discrète. Plus précisément, soit X une variable aléatoire de loi $\hat{F}$, alors pour $i = 1, \dots, n$ X vérifie

$$\begin{aligned}\mathbb{P}(X = x_i) &= \int_{\{x_i\}} \mathrm{d}\hat{F}(u) = \hat{F}(x_i) - \lim_{u \to x_i -} \hat{F}(u) \\ &= \frac{\#\{k : x_k \leq x_i\}}{n} - \frac{\#\{k : x_k < x_i\}}{n} = \frac{\#\{k : x_k = x_i\}}{n}.\end{aligned}$$

Si les valeurs des observations sont deux à deux distinctes ($x_i \neq x_j$ pour tout $i \neq j$), alors

$$\mathbb{P}(X = x_i) = \frac{1}{n},$$

et $\hat{F}$ est la loi uniforme (discrète) sur $\{x_1, \dots, x_n\}$.

Lorsque les x_i sont des réalisations i.i.d. de loi F, la fonction de répartition empirique $\hat{F}_n$ donne une approximation de F. On appelle $\hat{F}$ la **loi empirique** des observations $\mathbf{x}$.

Théorème 1. *Soient $X_1, X_2, \dots$ une suite de variables aléatoires i.i.d. de loi F et $\hat{F}_n$ la fonction de répartition empirique associée à $(X_1, \dots, X_n)$.*

(i) $n\hat{F}_n(t) \sim \mathrm{Bin}(n, F(t))$.

(ii) $\hat{F}_n(t) \longrightarrow F(t)$ *p.s. lorsque* $n \to \infty$ *pour tout* $t \in \mathbb{R}$.

(iii) $\sqrt{n}(\hat{F}_n(t) - F(t)) \xrightarrow{\mathcal{L}} \mathcal{N}(0, F(t)(1 - F(t)))$ *lorsque* $n \to \infty$.

(iv) **(Théroème de Glivenko-Cantelli)** $\hat{F}_n$ *converge uniformément presque sûrement vers F, c'est-à-dire*

$$\|\hat{F}_n - F\|_\infty := \sup\left\{|\hat{F}_n(t) - F(t)|, t \in \mathbb{R}\right\} \longrightarrow 0 \text{ p.s.}, \quad n \to \infty.$$

Démonstration. *(i)* Notons $Y_i = \mathbb{1}\{X_i \leq t\}$. Les Y_i sont i.i.d., car les X_i le sont. Comme Y_i prend ses valeurs dans $\{0, 1\}$, Y_i suit la loi de Bernoulli de paramètre $\mathbb{P}(Y_i = 1) = \mathbb{P}(X_i \leq t) = F(t)$. D'où $n\hat{F}_n(t) = \sum_{i=1}^n Y_i \sim \mathrm{Bin}(n, F(t))$.

(ii) Par la loi forte des grands nombres, $\hat{F}_n(t) = \frac{1}{n}\sum_{i=1}^n Y_i \longrightarrow \mathbb{E}[Y_1] = F(t)$ *p.s.* lorsque $n \to \infty$.

(iii) D'après le théorème central limite, quand $n \to \infty$,

$$\sqrt{n}(\hat{F}_n(t) - F(t)) = \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n Y_i - \mathbb{E}[Y_1]\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \mathbf{Var}(Y_1)) = \mathcal{N}(0, F(t)(1-F(t))).$$

(iv) Montrons d'abord le théorème de Glivenko-Cantelli dans le cas particulier où F est continue. Soit $0 < \varepsilon < 1$. Il existe une partition $-\infty = t_0 < t_1 < \cdots < t_k = \infty$ telle que $F(t_j) - F(t_{j-1}) < \varepsilon$ pour tout $j = 1, \dots, k$. Par *(ii)* on obtient la convergence uniforme sur un nombre fini de points, à savoir

$$\sup\left\{|\hat{F}_n(t) - F(t)|, t \in \{t_1, \dots, t_{k-1}\}\right\} \longrightarrow 0 \text{ p.s.}, \quad n \to \infty.$$

Or, pour tout $t \in [t_{j-1}, t_j]$, on a

$$\begin{aligned}\hat{F}_n(t) - F(t) &\leq \hat{F}_n(t_j) - F(t_j) + \varepsilon, \\ \hat{F}_n(t) - F(t) &\geq \hat{F}_n(t_{j-1}) - F(t_{j-1}) - \varepsilon.\end{aligned}$$

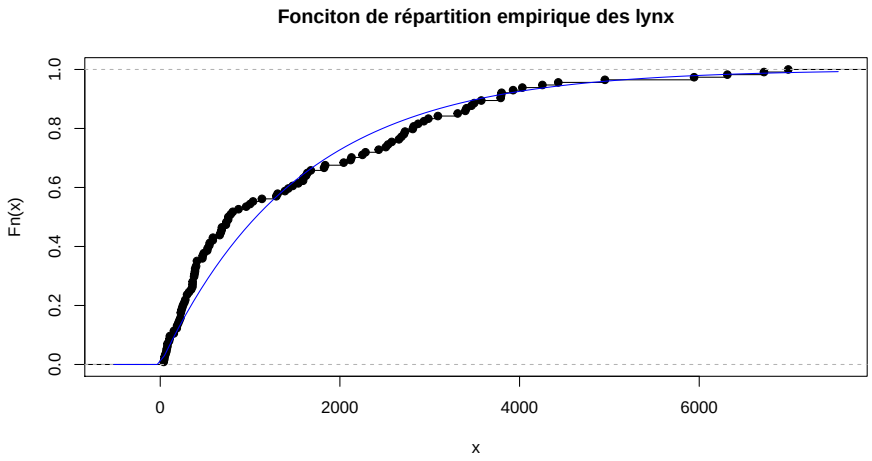


FIGURE 1.2 – Fonction de répartition empirique des données des lynx en comparaison avec la fonction de répartition de la loi exponentielle $\mathcal{E}(\hat{\theta})$ avec $\hat{\theta} = 0.0006501876$.

Donc,

$$\begin{aligned}\|\hat{F}_n - F\|_\infty &= \sup\left\{|\hat{F}_n(t) - F(t)|, t \in \mathbb{R}\right\} \\ &\leq \sup\left\{|\hat{F}_n(t) - F(t)|, t \in \{t_1, \ldots, t_{k-1}\}\right\} + \varepsilon\end{aligned}$$

On en déduit que $\limsup_{n\to\infty} \|\hat{F}_n - F\|_\infty \leq \varepsilon$ *p.s.* Ceci est vrai pour tout $\varepsilon > 0$, ce qui implique le résultat.

Pour le cas général où F n'est pas nécessairement continue, on fixe à nouveau $\varepsilon \in (0, 1)$, et on note $s_1 < s_2 < \cdots < s_K$ les points où F fait un saut de taille plus grande que ε, c'est-à-dire tels que $F(s_i) - F(s_i-) > \varepsilon$ (il n'y en a qu'un nombre fini $K < 1/\varepsilon$ puisque F est croissante et à valeur dans $[0, 1]$). On décompose ensuite chacun des intervalles $[s_i, s_{i+1}[$ en $s_i = t_{i,1} < \ldots t_{i,n_i} < s_{i+1}$ de telle sorte que $F(t_{i,j+1}) - F(t_j) < \varepsilon$ pour tout $j \in \{1, \ldots, n_i - 1\}$ et que $F(s_{i+1}-) - F(t_{i,n_i}) < \varepsilon$. On renumerote $\tau_1 < \ldots \tau_M$ l'ensemble des $(t_{i,j})_{i,j}$. Le reste de la démonstration est ensuite similaire : la convergence ponctuelle en les τ_i implique que $\limsup_{n\to\infty} \|\hat{F}_n - F\|_\infty \leq \varepsilon$ *p.s.*

□

En effet, sous des hypothèses assez générales sur la fonctionnelle T,

$$T(\hat{F}_n) \longrightarrow T(F_\theta) \text{ } p.s., \quad \text{quand } n \to \infty,$$

ce qui justifie l'application de la méthode de substitution.

La Figure 1.2 montre la fonction de répartition empirique des données des lynx en comparaison avec la fonction de répartition de la loi exponentielle $\mathcal{E}(\hat{\theta})$ avec $\hat{\theta} = 0.0006501876$. L'ajustement des deux courbes indique une grande similarité des deux lois.

EXEMPLE : MÉTHODE DE SUBSTITUTION.

Soit $\mathbf{x} = (x_1, \ldots, x_n)$ la réalisation d'un vecteur aléatoire $\mathbf{X} = (X_1, \ldots, X_n)$ composé de v.a. i.i.d. X_i de loi F intégrable. L'objectif est l'estimation de l'espérance $\theta := \mathbb{E}_F[X]$ à

partir de l'observation $\mathbf{x}$. On écrit

$$\theta = \mathbb{E}_F[X] = \int x \mathrm{d}F(x).$$

Notons $v_1, \ldots, v_m$ les différentes valeurs prises par $x_1, \ldots, x_n$ (c'est-à-dire $v_i \neq v_j$ pour tout $i \neq j$). Notons $\hat{F}$ la fonction de répartition empirique associée à $x_1, \ldots, x_n$. Par la méthode de substitution, on a

$$\begin{aligned} \hat{\theta} := \mathbb{E}_{\hat{F}}[X] &= \int x \mathrm{d}\hat{F}(x) = \sum_{j=1}^{m} \mathbb{P}_{\hat{F}}(X = v_j) v_j \\ &= \sum_{j=1}^{m} \frac{\#\{i : x_i = v_j\}}{n} v_j \\ &= \frac{1}{n} \sum_{i=1}^{n} x_i = \bar{x}_n. \end{aligned} \tag{1.5}$$

1.2 Propriétés d'un estimateur

Il existe différentes méthodes d'estimation et donc il est important de savoir comment comparer deux ou plusieurs estimateurs d'un paramètre θ dans un modèle $\{\mathbb{P}_\theta, \theta \in \Theta\}$ donné. Pour cela il faut savoir caractériser un estimateur. Dans cette partie, nous présenterons deux propriétés d'estimateur : une propriété minimale pour un bon estimateur (la consistance) et un critère pour comparer des estimateurs à n fini (le risque quadratique).

Un estimateur $\hat{\theta}$ est une fonction associant une valeur $\hat{\theta}(\mathbf{x})$ à une observation $\mathbf{x}$ que l'on espère proche de la vraie valeur θ_0. Pour évaluer si l'approximation est bonne, on ne veut pas se limiter à étudier un cas particulier d'un échantillon $\mathbf{x}$ fixé. En revanche, on s'intéresse à la règle générale à partir de laquelle est définie la statistique $\hat{\theta} = \hat{\theta}(\mathbf{x})$ pour une réalisation $\mathbf{x}$ *quelconque* de la loi $\mathbb{P}_{\theta_0}$. Donc, au lieu de vérifier si $\hat{\theta}(\mathbf{x})$ est près de θ_0 pour un jeu de données $\mathbf{x}$ fixé, on considère la *variable aléatoire* $\hat{\theta}(\mathbf{X})$ où $\mathbf{X}$ suit la loi $\mathbb{P}_{\theta_0}$ et on étudie la distance entre $\hat{\theta}(\mathbf{X})$ et θ_0. Pour distinguer $\hat{\theta}(\mathbf{x})$ de $\hat{\theta}(\mathbf{X})$, on appelle parfois $\hat{\theta}(\mathbf{X})$ l'**estimateur** et $\hat{\theta}(\mathbf{x})$ une **estimation**. Par ailleurs, la vraie valeur θ_0 étant inconnue, on s'intéresse au comportement de $\hat{\theta}(\mathbf{X})$ par rapport à θ_0 quelque soit la loi $\mathbb{P}_{\theta_0}$ de $\mathbf{X}$ *pour tout* $\theta_0 \in \Theta$.

Dans cette partie nous supposons que l'observation $\mathbf{x} = (x_1, \ldots, x_n)$ est la réalisation d'un vecteur aléatoire $\mathbf{X} = (X_1, \ldots, X_n)$ de loi $\mathbb{P}_{\theta_0}$ appartenant à une famille paramétrique de lois $\{\mathbb{P}_\theta, \theta \in \Theta\}$ avec $\Theta \subset \mathbb{R}^d$ et $d < \infty$. On va essentiellement traiter le cas d'un échantillon i.i.d., c'est-à-dire quand les X_i sont des variables aléatoires indépendantes et de même loi.

1.2.1 Consistance

Typiquement, un estimateur $\hat{\theta}$ est bien défini quelque soit la taille d'échantillon n (voir p. ex. la moyenne empirique $\bar{x}_n$). Pour mettre en avant la dépendance de $\hat{\theta}$ de n, on note $\hat{\theta}_n$.

Intuitivement, il devrait être plus facile d'estimer le paramètre θ_0 si on dispose d'un grand échantillon $\mathbf{x}_n = (x_1, \ldots, x_n)$ avec n grand que si n est petit, car chaque observation x_i apporte de l'information sur la loi $\mathbb{P}_{\theta_0}$ et donc sur le paramètre à estimer θ_0. Or, lorsque l'échantillon $\mathbf{x}_n$ croît indéfiniment, c'est-à-dire quand n tend vers l'infini, on attend d'un

estimateur "raisonnable" à ce que $\hat{\theta}_n(\mathbf{x}_n)$ converge vers θ_0. Cette réflexion mène à la notion de la consistance d'un estimateur, propriété minimale que l'on exigera de tout estimateur.

Quand on étudie les propriétés asymptotiques d'un estimateur $\hat{\theta} = \hat{\theta}_n$ lorsque n tend vers l'infini, le terme *estimateur* désignera aussi, pour abréger, une suite d'estimateurs $(\hat{\theta}_n(\mathbf{X}_n))_{n\geq 1}$ ou bien la règle à partir de laquelle est définie la statistique $\hat{\theta}_n(\mathbf{X}_n)$ pour tout n donné.

Un estimateur $\hat{\theta}_n = \hat{\theta}_n(\mathbf{X}_n)$ de θ est dit **convergent** ou **consistant** si

$$\hat{\theta}_n \xrightarrow{P} \theta, \quad \text{pour tout } \theta \in \Theta.$$

Plus précisément, cette notation veut dire que quelque soit $\theta \in \Theta$, pour $\mathbf{X}_n$ de loi $\mathbb{P}_\theta$, l'estimateur $\hat{\theta}_n(\mathbf{X}_n)$ converge en probabilité vers θ.

Dans cette définition, la convergence doit avoir lieu pour tout $\theta \in \Theta$, ce qui garantit qu'elle a lieu pour la vraie valeur inconnue θ_0 des observations $\mathbf{x}_n$. La consistance est une propriété liée au modèle statistique : un estimateur $\hat{\theta}_n$ peut être consistant pour un modèle et non-consistant pour un autre.

Si l'on a la convergence presque sûre : $\hat{\theta}_n \to \theta$ *p.s.* au lieu de la convergence en probabilité, on dit que l'estimateur $\hat{\theta}_n$ est **fortement consistant**.

La consistance est une propriété assez faible. Cette notion n'est pas assez informative pour nous guider dans le choix d'estimateurs. Néanmoins, elle n'est pas complètement inutile, car elle permet de rétrécir l'ensemble d'estimateurs que l'on doit étudier. En effet, les estimateurs non consistants doivent être exclus de toute considération.

Exemple d'un estimateur non-consistant

Soient $X_i, i = 1, 2, \ldots$ des v.a. i.i.d. de loi normale $\mathcal{N}(\mu, 1)$. Considérons l'estimateur $\tilde{\mu}_n := (X_{n-1} + X_n)/2$ de μ. Il est clair que $\tilde{\mu}_n$ ne converge pas en probabilité vers μ lorsque n tend vers l'infini. La raison pour la non-consistance est que, essentiellement, $\tilde{\mu}_n$ n'exploite pas toute l'information disponible dans les observations $X_1, \ldots, X_n$, contrairement à la moyenne empirique $\hat{\mu}_n := \bar{X}_n$ p.ex. qui est convergeant d'après la loi des grands nombres.

1.2.2 Risque quadratique ou erreur quadratique moyenne

Afin de comparer les estimateurs dans un modèle statistique pour une taille d'échantillon n finie, on utilise souvent le risque quadratique.

On appelle **risque quadratique** ou **erreur quadratique moyenne** de l'estimateur $\hat{\theta}$ au point $\theta \in \Theta$ la quantité

$$R(\theta, \hat{\theta}) = \mathbb{E}_\theta \left[\|\hat{\theta} - \theta\|^2 \right] = \mathbb{E}_\theta \left[\|\hat{\theta}(\mathbf{X}) - \theta\|^2 \right] = \int \|\hat{\theta}(\mathbf{x}) - \theta\|^2 \mathrm{d}F_\theta(\mathbf{x}).$$

Le risque quadratique est bien défini pour tout estimateur $\hat{\theta}$. Il peut, en particulier, prendre la valeur $R(\theta, \hat{\theta}) = +\infty$. Le risque permet de mesurer la distance entre l'estimateur $\hat{\theta}$ et la valeur θ_0.

Théorème 2. *Le risque quadratique admet la décomposition suivante :*

$$R(\theta, \hat{\theta}) = \left(\|\mathbb{E}_\theta[\hat{\theta}] - \theta\| \right)^2 + \mathbb{E}_\theta \left[\|\hat{\theta} - \mathbb{E}_\theta[\hat{\theta}]\|^2 \right].$$

Démonstration. On a

$$\begin{aligned} R(\theta,\hat{\theta}) &= \mathbb{E}_\theta\left[\|\hat{\theta}-\mathbb{E}_\theta[\hat{\theta}]+\mathbb{E}_\theta[\hat{\theta}]-\theta\|^2\right] \\ &= \mathbb{E}_\theta\left[\|\hat{\theta}-\mathbb{E}_\theta[\hat{\theta}]\|^2\right] + \|\mathbb{E}_\theta[\hat{\theta}]-\theta\|^2 + 2\left\langle \mathbb{E}_\theta[\hat{\theta}]-\mathbb{E}_\theta[\hat{\theta}], \mathbb{E}_\theta[\hat{\theta}]-\theta\right\rangle \\ &= \underbrace{\|\mathbb{E}_\theta[\hat{\theta}]-\theta\|^2}_{=:b^2(\theta,\hat{\theta})} + \underbrace{\mathbb{E}_\theta\left[\|\hat{\theta}-\mathbb{E}_\theta[\hat{\theta}]\|^2\right]}_{=:\sigma^2(\theta,\hat{\theta})}. \end{aligned}$$

□

Le terme $b^2(\theta,\hat{\theta})$ représente la partie déterministe de l'erreur d'estimation, alors que $\sigma^2(\theta,\hat{\theta})$ mesure la contribution de sa partie stochastique.

Si $\Theta \subset \mathbb{R}$, on appelle $b(\theta,\hat{\theta})$ le **biais** de l'estimateur $\hat{\theta}$ et $\sigma^2(\theta,\hat{\theta})$ est la **variance** de $\hat{\theta}$. On a $\sigma^2(\theta,\hat{\theta}) = \mathbf{Var}_\theta(\hat{\theta})$. On dit qu'un estimateur $\hat{\theta}$ est **sans biais** si $\mathbb{E}_\theta[\hat{\theta}] = \theta$ (i.e. $b(\theta,\hat{\theta}) = 0$) pour tout $\theta \in \Theta$. Dans le cas contraire, on dit que $\hat{\theta}$ est **biaisé**. Si $\mathbb{E}_\theta[\hat{\theta}_n] \to \theta$ lorsque $n \to \infty$, on dit que $\hat{\theta}_n$ est **asymptotiquement sans biais**.

Un bon estimateur est caractérisé par un petit biais et une petite variance. Mais en général, ces deux termes sont en compétition et on ne peut pas les minimiser simultanément, et il n'existe pas d'estimateur dont le risque quadratique est 0. En minimisant le risque quadratique, on cherche un compromis entre le biais et la variance.

Plus la valeur du risque est petite, plus l'estimateur $\hat{\theta}$ est performant. Afin de comparer deux estimateurs, on peut comparer leurs risques quadratiques. Soient $\hat{\theta}^{(1)}$ et $\hat{\theta}^{(2)}$ deux estimateurs de θ dans le modèle statistique $\{\mathbb{P}_\theta, \theta \in \Theta\}$. Si

$$R(\theta,\hat{\theta}^{(1)}) \leq R(\theta,\hat{\theta}^{(2)}) \quad \text{pour tout } \theta \in \Theta,$$

et si, de plus, il existe $\theta' \in \Theta$ tel que l'inégalité est stricte, alors on dit que $\hat{\theta}^{(1)}$ est **plus efficace** que $\hat{\theta}^{(2)}$ (ou meilleur que $\hat{\theta}^{(2)}$) et que $\hat{\theta}^{(2)}$ est **inadmissible**.

1.2.3 À propos de l'EMV

La définition de l'EMV étant très générale, on peut considérer l'EMV dans des modèles très variés. Il n'est p.ex. pas limité au cas d'observations i.i.d. (comme la méthode de substitution) ou aux lois intégrables (comme la méthode des moments).

Sous des conditions assez faibles, l'EMV est consistant.

Le calcul explicite du risque quadratique dépend du modèle. Très souvent, on n'a pas d'expression explicite, mais en général on peut approcher la valeur du risque quadratique par des simulations de Monte-Carlo (voir Chapitre 2, Section 2.1).

Sous des hypothèses de régularité du modèle – on parle de *modèle régulier* – on peut montrer que l'EMV est (asymptotiquement) optimal (dans un certain sens). Par conséquent, dans de très nombreux contextes les statisticiens souhaitent utiliser l'EMV. En revanche, le calcul de l'EMV n'est explicite que dans certains modèles statistiques jouet. En pratique, il est rare de pouvoir exhiber la formule explicite de l'estimateur du maximum de vraisemblance. Le plus souvent, et notamment dans des modèles pertinents pour la pratique, le problème de maximisation (1.1) n'admet pas de solution explicite. Par conséquent, il est nécessaire de recourir à des méthodes numériques. Une des méthodes numériques les plus répandues en statistique est la méthode de Newton-Raphson, qui sera présentée en Section 1.3.2

Exemple. EMV de la loi Gamma.

Revenons aux données sur les lynx. Au vu de l'histogramme de la Figure 1.1 a) on peut se demander si une loi Gamma ne s'ajusterait pas mieux aux données qu'une exponentielle, car la densité de la loi Gamma est à valeurs positives, unimodale (un pic) et asymétrique.

Notons $f_{\alpha,\beta}$ la densité de la loi Gamma $\Gamma(\alpha,\beta)$ avec $\alpha>0,\beta>0$ donnée par

$$f_{\alpha,\beta}(x)=\frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}\mathrm{e}^{-\beta x},\quad x>0,$$

où $\Gamma(\alpha)=\int_0^\infty t^{\alpha-1}\mathrm{e}^{-t}\mathrm{d}t$ désigne la fonction Gamma.

Pour des réalisations $x_1,\dots,x_n$ i.i.d. de la loi Gamma $\Gamma(\alpha,\beta)$ la fonction de vraisemblance vaut

$$\begin{aligned}\mathcal{L}(x_1,\dots,x_n;\alpha,\beta)&=\prod_{i=1}^n f_{\alpha,\beta}(x_i)=\prod_{i=1}^n\frac{\beta^\alpha}{\Gamma(\alpha)}x_i^{\alpha-1}\mathrm{e}^{-\beta x_i}\\&=\frac{\beta^{n\alpha}}{(\Gamma(\alpha))^n}\left(\prod_{i=1}^n x_i\right)^{\alpha-1}\exp\left\{-\beta\sum_{i=1}^n x_i\right\}.\end{aligned}$$

La fonction de log-vraisemblance et ses dérivées partielles sont données par

$$\begin{aligned}\ell(\alpha,\beta)&=\log(\mathcal{L}(x_1,\dots,x_n;\alpha,\beta))\\&=n\alpha\log\beta-n\log(\Gamma(\alpha))+(\alpha-1)\sum_{i=1}^n\log x_i-\beta\sum_{i=1}^n x_i\\\frac{\partial}{\partial\alpha}\ell(\alpha,\beta)&=n\log\beta-\frac{n\Gamma'(\alpha)}{\Gamma(\alpha)}+\sum_{i=1}^n\log x_i\\\frac{\partial}{\partial\beta}\ell(\alpha,\beta)&=\frac{n\alpha}{\beta}-\sum_{i=1}^n x_i,\end{aligned}$$

où $\Gamma'(\cdot)$ désigne la dérivée de la fonction $\Gamma(\cdot)$. Il est clair que $\nabla\ell(\alpha,\beta)=0$ n'a pas de solution explicite, notamment parce que la fonction $\Gamma(\cdot)$ est définie par une intégrale et $\Gamma'(\cdot)$ n'a pas d'expression explicite. Néanmoins, cela ne met pas en question l'existence de l'EMV ! Mais pour le calculer il faut utiliser des méthodes numériques comme par exemple la méthode de Newton-Raphson.

1.3 Optimisation d'une fonction

En vue du calcul de l'EMV, et plus précisément, de la solution du problème de maximisation en (1.1), nous rappelons dans cette partie d'abord les techniques d'optimisation classiques d'analyse. Ensuite, nous présenterons la méthode de Newton-Raphson qui est une méthode numérique de maximisation répandue en statistique.

1.3.1 Rappel : Techniques d'optimisation classiques

Rappelons qu'il existe des fonctions non bornées, qui n'ont donc pas de maximum global. Il existe alors des problèmes de maximisation qui n'admettent pas de solution.

Fonction concave

Soit $f : I \subset \mathbb{R} \to \mathbb{R}$ une fonction deux fois dérivable définie sur un intervalle I. Si

$$f''(x) < 0, \quad \text{pour tout } x \in I,$$

alors f est strictement concave. Par conséquence, si f admet un maximum global, il est unique. Pour le trouver, il suffit de chercher la solution de $f'(x) = 0$. Si $f'(x) = 0$ n'a pas de solutions dans I, le maximum se trouve aux bords de l'intervalle I.

Fonction deux fois dérivable

Soit $f : I \subset \mathbb{R} \to \mathbb{R}$ une fonction deux fois dérivable, mais pas nécessairement concave. Alors on calcule tous les points critiques de la fonction f, i.e. toutes les solutions de $f'(x) = 0$, et on étudie le comportement de f aux bords de l'intervalle I.

Si un point critique x^* est tel que $f''(x^*) < 0$, alors x^* est un maximum local.

Si un point critique x^* vérifie $f''(x^*) > 0$, il s'agit d'un minimum local.

Si un point critique x^* est tel que $f''(x^*) = 0$, il peut s'agir d'un point d'inflexion ou d'un maximum local ou d'un minimum local. Dans ce cas, on peut étudier le comportement de la dérivée f' dans un voisinage V de x^*. Si

$$f'(x) > 0, \quad \forall x < x^* \text{ et } x \in V \qquad \text{et} \qquad f'(x) < 0, \quad \forall x > x^* \text{ et } x \in V, \tag{1.6}$$

alors x^* est bien un maximum (local).

Une fois qu'on a déterminé tous les maxima locaux, on détermine le maximum global en comparant les valeurs de f en ses maxima locaux et en prenant en compte le comportement de f aux bords de l'intervalle I. Rappelons que si l'intervalle I est ouvert, il est possible que f n'admette pas de maximum global.

Fonction dérivable

Si f n'est dérivable qu'une fois, on détermine tous les points critiques de f, i.e. toutes les solutions de $f'(x) = 0$. Puis on détermine les maxima locaux en vérifiant (1.6). Enfin, afin de trouver le maximum global (et afin de voir s'il existe), on compare les valeurs de f en ces maxima et on prend en compte le comportement de f aux bords de l'intervalle I.

Fonction non dérivable

Si f n'est pas dérivable, on peut établir le tableau de variation de la fonction et/ou tracer l'allure de la fonction pour trouver le maximum global de f (s'il existe).

Fonction de plusieurs variables

Soit $f : D \subset \mathbb{R}^d \to \mathbb{R}$ une fonction deux fois dérivable définie sur D. Si la matrice hessienne

$$H(x) = \nabla^2 f(x) < 0, \quad \text{pour tout } x \in D,$$

alors f est strictement concave et la solution de $\nabla f(x) = 0$ (si elle existe) est le maximum global de f.

Si f est deux fois dérivable mais pas concave, alors un point x_0 tel que

$$\nabla f(x_0) = 0 \quad \text{et} \quad H(x_0) < 0,$$

est un maximum local de f.

Rappelons une propriété de l'algèbre sur les matrice symétrique $A = (a_{i,j})_{1 \leq i,j \leq r}$ de taille $r \times r$. Notons $A_s = (a_{i,j})_{1 \leq i,j \leq s}$. On appelle *s-ième mineur principal dominant* de A le déterminant de A_s, $\det(A_s)$. La matrice A est définie négative, notée $A < 0$, si et seulement si tous les mineurs principaux dominants avec s pair sont strictement positifs, et tous les mineurs principaux dominants avec s impair sont négatifs.

Maximisation sous contrainte

La *méthode des multiplicateurs de Lagrange* est une façon de ramener un problème d'optimisation sous contrainte à un problème d'optimisation sans contrainte. Elle repose sur le résultat suivant.

Théorème 3 (Théorème des extrema liés)**.** *Soient $f, \psi_1, \ldots, \psi_p \in \mathcal{C}^1(\mathbb{R}^d, \mathbb{R})$ et $\mathcal{D} = \{x \in \mathbb{R}^d,\ \psi_1(x) = \cdots = \psi_p(x) = 0\}$. Si x_0 est un extremum local de f sur $\mathcal{D}$ et si les vecteurs $\nabla\psi_1(x_0), \ldots, \nabla\psi_p(x_0)$ forment une famille libre de $\mathbb{R}^d$, alors il existe $\lambda_1, \ldots, \lambda_p \in \mathbb{R}$ tels que $\nabla f(x_0) = \lambda_1 \nabla\psi_1(x_0) + \cdots + \lambda_p \nabla\psi_p(x_0)$.*

Afin de prendre en compte des contraintes dans un problème de maximisation, on introduit des multiplicateurs de Lagrange $\lambda = (\lambda_1, \ldots, \lambda_p)^T$ et la fonction de Lagrange

$$L(x, \lambda) = f(x) + \sum_{k=1}^{p} \lambda_k \psi_k(x).$$

Soit un point critique (x_0, λ) de L tel que

$$0 \ = \ \nabla L(x_0, \lambda) \ = \ \begin{pmatrix} \nabla f(x_0) + \sum_{k=1}^{p} \lambda_k \nabla\psi_k(x_0) \\ \psi_1(x_0) \\ \vdots \\ \psi_p(x_0) \end{pmatrix}.$$

Autrement dit, tout point critique (x_0, λ) de L vérifie les contraintes $\psi_1(x_0) = \cdots = \psi_p(x_0) = 0$. Donc, $x_0 \in \mathcal{D}$. Par ailleurs, $\nabla f(x_0)$ est une combinaison linéaire des $\nabla\psi_i(x_0)$, ce qui d'après le théorème des extrema liés est une condition nécessaire pour que x_0 soit un extremum de f sur $\mathcal{D}$. On a donc ramené le problème de maximisation sous contraintes à un problème de recherche de points critiques. Une fois les points critiques déterminé, il faut les étudier pour identifier le maximum global.

1.3.2 Méthode de Newton-Raphson

La méthode de Newton-Raphson est une procédure numérique pour déterminer un point critique d'une fonction f. En appliquant cette méthode à la fonction de log-vraisemblance $\ell(\theta)$, on peut espérer de trouver son maximum et donc l'estimateur de maximum de vraisemblance. Bien évidemment, trouver un point critique n'est pas équivalent à déterminer le point du maximum global d'une fonction, car ce point peut être un maximum local seulement, ou même un minimum ou un point d'inflexion. Néanmoins, dans des nombreux cas, cette méthode donne des résultats satisfaisants pour détecter le point maximum d'une fonction.

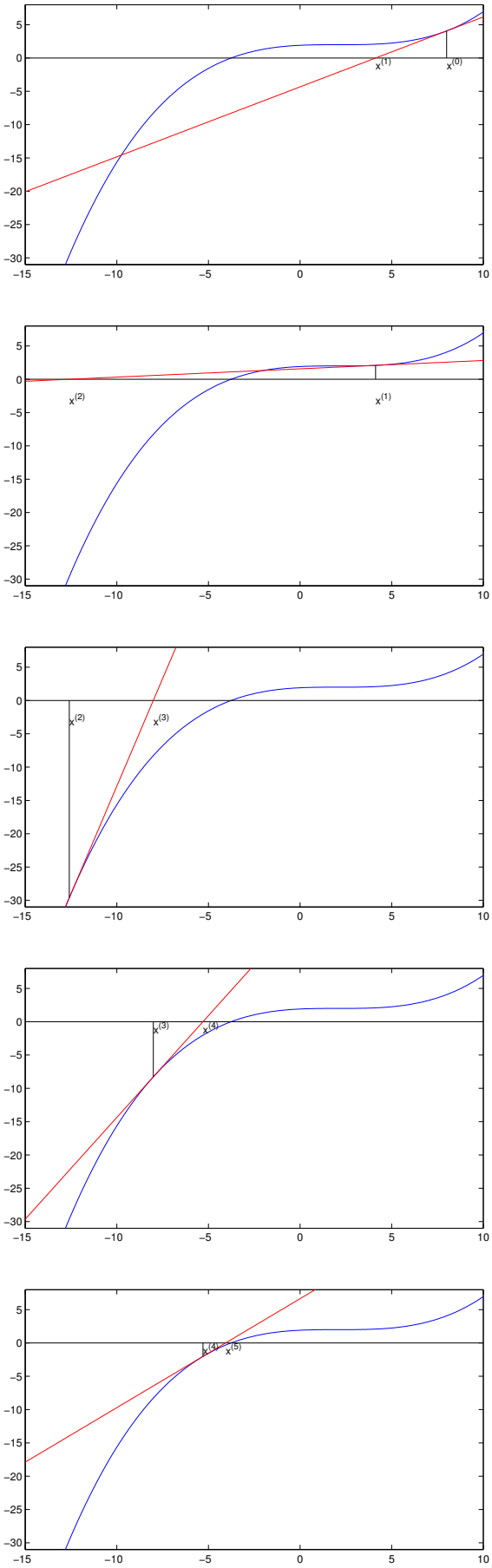


FIGURE 1.3 – Illustration des 5 premières itérations de la méthode de Newton-Raphson.

La méthode de Newton-Raphson

La méthode de Newton-Raphson est une procédure itérative pour trouver des points critiques d'une fonction réelle $f : \mathcal{X} \to \mathbb{R}$ où $\mathcal{X} \subset \mathbb{R}^d$. On suppose que f est deux fois dérivable et on cherche la (ou les) solutions de $\nabla f(x) = 0$. La méthode de Newton-Raphson repose sur le développement de Taylor, plus précisément sur l'approximation linéaire du gradient $\nabla f(x)$ par

$$\nabla f(x) = \nabla f(\xi) + H(\xi)(x - \xi) + r(x, \xi), \tag{1.7}$$

où $\xi \in \mathcal{X}$, $H(\xi) = \nabla^2 f(\xi)$ est la matrice hessienne de f en ξ et $r(x, \xi)$ est un terme de reste.

Si ξ est près de x, le reste r est négligeable comparé au terme linéaire. Au lieu de résoudre $\nabla f(x) = 0$ directement, la méthode de Newton-Raphson consiste à négliger le terme r en (1.7) et résoudre

$$\nabla f(\xi) + H(\xi)(x - \xi) = 0 \tag{1.8}$$

par rapport à x. En fait, on procède itérativement : on se donne un point initial $\xi = x^{(0)}$, puis on résout l'équation (1.8) par rapport à x et on appelle la solution $x^{(1)}$. Ensuite, on pose $\xi = x^{(1)}$ et on recommence à résoudre l'équation (1.8) par rapport à x et ainsi de suite. Plus généralement, l'itération t consiste à calculer

$$x^{(t)} = x^{(t-1)} - \left[H(x^{(t-1)})\right]^{-1} \nabla f(x^{(t-1)}), \tag{1.9}$$

où $x^{(t-1)}$ est le résultat de l'itération précédente.

Pour que l'algorithme soit bien défini, il est nécessaire que l'inverse $\left[H(x^{(t)})\right]^{-1}$ existe pour tout t.

Interprétation géométrique

Pour une interprétation géométrique de la méthode de Newton-Raphson remarquons que le terme à gauche de l'équation (1.8) est la tangente à $\nabla f(x)$ au point $x = \xi$. Au lieu de chercher le point où le gradient $\nabla f(x)$ s'annule, on cherche donc le zéro de la tangente. Étant donné que la tangente est une fonction linéaire, il est beaucoup plus facile de déterminer ce point que de trouver un zéro de $\nabla f(x)$.

Figure 1.3 illustre les cinq premières étapes de la méthode de Newton-Raphson dans un exemple. La suite des $x^{(t)}$ avec point initial $x^{(0)} = 8$ est la suivante :

$x^{(0)} = 8$	$\nabla f(x^{(0)}) = 4{,}07$
$x^{(1)} = 4{,}12$	$\nabla f(x^{(1)}) = 2{,}08$
$x^{(2)} = -12{,}49$	$\nabla f(x^{(2)}) = -29{,}61$
$x^{(3)} = -8{,}01$	$\nabla f(x^{(3)}) = -8{,}28$
$x^{(4)} = -5{,}31$	$\nabla f(x^{(4)}) = -2{,}02$
$x^{(5)} = -4{,}06$	$\nabla f(x^{(5)}) = -0{,}32$
$x^{(6)} = -3{,}7800$	$\nabla f(x^{(6)}) = -0{,}0145$
$x^{(7)} = -3{,}7659$	$\nabla f(x^{(7)}) = -3{,}48\ 10^{-5}$
$x^{(8)} = -3{,}7659$	$\nabla f(x^{(8)}) = -2{,}02\ 10^{-10}$
$x^{(9)} = -3{,}7659$	$\nabla f(x^{(9)}) = 4{,}44\ 10^{-16}$

On observe que l'algorithme converge après quelques itérations seulement. En effet, à toute itération (sauf la deuxième), on s'approche du zéro de la fonction $\nabla f(x)$.

Critères d'arrêt

Plusieurs critères d'arrêt sont envisageable pour cet algorithme. Les deux plus courants sont les suivants. Soit $\varepsilon > 0$ un seuil fixé.

— On arrête dès que $\|x^{(t)} - x^{(t-1)}\| < \varepsilon$.

— On arrête dès que $|\nabla f(x^{(t)})| < \varepsilon$.

Quelque soit le critère d'arrêt, il est possible que la condition soit vérifiée en des points qui ne correspondent pas à des zéros de $\nabla f(x)$.

En général, il est difficile de garantir que la suite $(x^{(t)})_t$ converge. En revanche, et c'est ce qui rend cette méthode attractive, *si* elle converge, elle converge assez vite comme dans l'exemple ci-dessus.

Vitesse de convergence

Supposons que x^* est une solution de $\nabla f(x) = 0$ et que $\|x^{(t)} - x^*\|$ est petit. Alors par (1.9)

$$\begin{aligned} x^{(t+1)} - x^* &= x^{(t)} - x^* - \left[H(x^{(t)})\right]^{-1} \nabla f(x^{(t)}) \\ &= x^{(t)} - x^* - \left[H(x^{(t)})\right]^{-1} (\nabla f(x^{(t)}) - \nabla f(x^*)), \end{aligned}$$

car $\nabla f(x^*) = 0$. Or, par le développement limité (1.7) on a

$$\nabla f(x^*) = \nabla f(x^{(t)}) + H(x^{(t)})(x^* - x^{(t)}) + r(x^*, x^{(t)}).$$

D'où

$$\begin{aligned} x^{(t+1)} - x^* &= x^{(t)} - x^* + \left[H(x^{(t)})\right]^{-1} \left\{H(x^{(t)})(x^* - x^{(t)}) + r(x^*, x^{(t)})\right\} \\ &= \left[H(x^{(t)})\right]^{-1} r(x^*, x^{(t)}). \end{aligned}$$

D'après le théorème de Taylor, $r(x^*, x^{(t)}) \leq c\|x^{(t)} - x^*\|^2$, où c dénote le maximum de $\|\nabla^3 f(x)\|$ sur le rectangle dont les extrémités sont déterminées par $x^{(t)}$ et x^*. Plus précisément, dans le cas unidimensionnel où $x \in \mathbb{R}$, ce rectangle est l'intervalle $[x^{(t)}, x^*]$ si $x^{(t)} < x^*$, ou bien $[x^*, x^{(t)}]$ si $x^* < x^{(t)}$.

On obtient alors

$$\|x^{(t+1)} - x^*\| \leq c \left\| \left[H(x^{(t)})\right]^{-1} \right\| \|x^* - x^{(t)}\|^2.$$

Ceci montre qu'en une itération l'erreur diminue de façon quadratique : on passe de $\|x^* - x^{(t)}\|$ à un terme d'ordre $\|x^* - x^{(t)}\|^2$. On dit que la vitesse de convergence de la méthode de Newton-Raphson est quadratique.

Exemple : Loi de Cauchy

Considérons la loi de Cauchy centrée ($\mu = 0$) de paramètre $\sigma > 0$ inconnu dont la densité est donnée par

$$f_\sigma(x) = \frac{\sigma}{\pi(\sigma^2 + x^2)}, \quad x \in \mathbb{R}.$$

Soient $\mathbf{x} = (x_1, \ldots, x_n)$ des réalisations i.i.d. d'une variable aléatoire X de loi de Cauchy$(0, \sigma)$. Essayons de calculer l'EMV de σ.

La fonction de vraisemblance s'écrit

$$\mathcal{L}(\mathbf{x};\sigma) = \prod_{i=1}^{n} f_\sigma(x_i) = \frac{\sigma^n}{\pi^n \prod_{i=1}^n (\sigma^2 + x_i^2)}.$$

On peut considérer la log-vraisemblance

$$\ell(\sigma) = \log \mathcal{L}(\mathbf{x};\sigma) = n\log\sigma - n\log\pi - \sum_{i=1}^{n} \log(\sigma^2 + x_i^2).$$

On dérive

$$\ell'(\sigma) = \frac{n}{\sigma} - 2\sum_{i=1}^{n} \frac{\sigma}{\sigma^2 + x_i^2}.$$

Résoudre l'équation $\ell'(\sigma) = 0$ est équivalent à trouver les zéros d'un polynôme d'ordre $2n$. Ce n'est pas faisable par un calcul explicite, et donc l'EMV n'est pas explicite dans ce modèle.

En revanche, on peut appliquer la méthode de Newton-Raphson pour trouver des points critiques de la fonction de log-vraisemblance $\ell(\sigma)$ afin d'approcher l'EMV numériquement. Pour cela, on calcule la dérivée seconde de $\ell(\sigma)$

$$\ell''(\sigma) = -\frac{n}{\sigma^2} - 2\sum_{i=1}^{n} \frac{\sigma^2 + x_i^2 - 2\sigma^2}{(\sigma^2 + x_i^2)^2} = -\frac{n}{\sigma^2} - 2\sum_{i=1}^{n} \frac{x_i^2 - \sigma^2}{(\sigma^2 + x_i^2)^2}$$

Or, la méthode de Newton-Raphson consiste à calculer iterativement (pour un point initial $\sigma^{(0)}$ choisi par l'utilisateur) pour $t = 0, 1, \ldots$

$$\begin{aligned}
\sigma^{(t+1)} &= \sigma^{(t)} - \frac{\ell'(\sigma^{(t)})}{\ell''(\sigma^{(t)})} \\
&= \sigma^{(t)} + \frac{\frac{n}{\sigma^{(t)}} - 2\sum_{i=1}^n \frac{\sigma^{(t)}}{(\sigma^{(t)})^2 + x_i^2}}{\frac{n}{(\sigma^{(t)})^2} + 2\sum_{i=1}^n \frac{x_i^2 - (\sigma^{(t)})^2}{[(\sigma^{(t)})^2 + x_i^2]^2}} \\
&= \frac{n\sigma^{(t)} - 2(\sigma^{(t)})^5 \sum_{i=1}^n \frac{1}{[(\sigma^{(t)})^2 + x_i^2]^2}}{\frac{n}{2} + (\sigma^{(t)})^2 \sum_{i=1}^n \frac{x_i^2 - (\sigma^{(t)})^2}{[(\sigma^{(t)})^2 + x_i^2]^2}}.
\end{aligned}$$

1.4 Intervalle de confiance

Jusuq'ici nous avons mis en évidence le rôle d'un estimateur en tant que pourvoyeur d'une "approximation" de la valeur inconnue du paramètre θ. Cela étant, une estimation sans degré de précision est douteuse, dans la mesure où elle est variable (il suffit d'ajouter ou de retrancher une observation pour changer sa valeur). Ainsi, lorsqu'un statisticien propose, au vu des observations $x_1, \ldots, x_n$, une estimation $\hat{\theta}_n$ de θ, quelle confiance peut-il avoir en son résultat ? Lorsque l'estimateur est consistant, tout ce qu'on sait, c'est que plus n est grand, plus $\hat{\theta}_n$ a des chances d'être voisin de θ.

Prenons comme exemple la moyenne empirique comme estimateur de l'espérance μ d'une loi (intégrable) F. Supposons que la moyenne empirique vaut 5,2. Intuitivement, nous accordons à cette estimation beaucoup plus de confiance lorsque la taille d'échantillon n est grande (p. ex. $n = 100\,000$) que quand elle est petite (p. ex. $n = 3$). À part de la

taille d'échantillon, d'autres facteurs peuvent influencer la précision d'une estimation $\hat{\theta}_n$, comme par exemple la variance ou la forme de la loi F. Nous voyons l'importance de savoir quantifier la précision ou volatilité d'un estimateur.

Dans ce paragraphe, nous introduisons la notion d'*intervalle de confiance*. L'idée consiste à calculer tout un intervalle (par opposition à un estimateur ponctuel) qui est susceptible de contenir la vraie valeur du paramètre θ avec une certaine probabilité prescrite. L'intervalle de confiance est un moyen pour quantifier l'incertitude d'un estimateur de θ.

1.4.1 Définition

Notons $\{\mathbb{P}_\theta, \theta \in \Theta\}$ un modèle statistique avec $\Theta \subset \mathbb{R}$. Supposons que $\mathbf{x} = (x_1, \ldots, x_n)$ est la réalisation de $\mathbf{X} = (X_1, \ldots, X_n)$ de loi $\mathbb{P}_\theta$. Soient $a(\cdot)$ et $b(\cdot)$ des fonctions boréliennes à valeurs dans $\mathbb{R}$, telles que $a(\mathbf{x}) < b(\mathbf{x})$ pour tout $\mathbf{x}$. Soit $0 < \alpha < 1$ un niveau de confiance donné. L'intervalle $[a(\mathbf{X}), b(\mathbf{X})]$ est dit **intervalle de confiance de niveau** $1-\alpha$ pour θ si

$$\mathbb{P}_\theta \left(a(\mathbf{X}) \leq \theta \leq b(\mathbf{X})\right) \geq 1 - \alpha \ , \tag{1.10}$$

pour tout $\theta \in \Theta$. On le note $\mathrm{IC}_{1-\alpha}(\theta)$.

On dit que $\mathrm{IC}_{1-\alpha}(\theta)$ est un intervalle de confiance de *taille* $1-\alpha$ pour θ si, pour tout $\theta \in \Theta$, on a égalité en (1.10).

En pratique, on choisit une valeur faible de α, typiquement de l'ordre de 0,1 ou 0,05, et on parle alors d'intervalle de confiance de niveau 90 % ou 95 %.

On doit comprendre un intervalle de confiance de niveau $1-\alpha$ comme un intervalle *aléatoire* qui a une probabilité $1-\alpha$ de contenir le vrai paramètre θ et non comme une région fixée auquel θ aléatoire appartient avec une probabilité $1-\alpha$. Dans la pratique, le statisticien calcule les réalisations numériques $a(\mathbf{x})$ et $b(\mathbf{x})$ de $a(\mathbf{X})$ et $b(\mathbf{X})$ à partir de l'observation $\mathbf{x}$, et cela lui fournit une *réalisation* de l'intervalle de confiance. Supposons par exemple que $\alpha = 0{,}05$ et que l'on ait trouvé $a = 2$ et $b = 7$. Même si la tentation est forte, on ne peut pas dire à proprement parler que l'intervalle $[2, 7]$ contient θ avec probabilité 0,95. Soit il contient θ, soit il ne le contient pas. Tout ce que l'on peut dire, c'est que la probabilité que l'intervalle qu'on vient de calculé contient θ est de 95 %. Ou encore : si l'on construit l'intervalle de confiance de niveau 0,95 pour 100 échantillons $\mathbf{x}$ différents, il est probable que 95 d'entre eux contiennent la vraie valeur de θ (mais on ne sait évidemment pas lesquels !).

Bien entendu, l'intervalle $\mathrm{IC}_{1-\alpha}(\theta) = (-\infty, \infty)$ convient toujours, mais n'est guère intéressant. En effet, on est intéressé de rendre l'intervalle $[a(\mathbf{X}), b(\mathbf{X})]$ le plus petit possible. On notera

$$\ell_{\mathrm{IC}} = b(\mathbf{X}) - a(\mathbf{X})$$

la *longueur de l'intervalle de confiance* $\mathrm{IC}_{1-\alpha}(\theta) = [a(\mathbf{X}), b(\mathbf{X})]$.

Il arrive parfois qu'on ne soit intéressé que par une borne inférieure ou une borne supérieure pour θ, $a(\mathbf{X})$ ou $b(\mathbf{X})$ étant rejeté à l'infini. On parle alors d'intervalle de confiance *unilatéral* (par opposition à *bilatéral*).

De façon analogue, on définit l'intervalle de confiance asymptotique.

Un **intervalle de confiance asymptotique de niveau** $1-\alpha$ pour θ est un intervalle aléatoire $[a_n(\mathbf{X}_n), b_n(\mathbf{X}_n)]$ tel que, pour tout $\theta \in \Theta$,

$$\liminf_{n\to\infty} \mathbb{P}_\theta \left(a_n(\mathbf{X}_n) \leq \theta \leq b_n(\mathbf{X}_n)\right) \geq 1 - \alpha \ . \tag{1.11}$$

Étant valables pour tout n fini, es intervalles de confiance sont préférables aux intervalles de confiance asymptotiques. En effet, pour les derniers, on ne contrôle pas exactement l'erreur, on ne fait que dire qu'elle est asymptotiquement de l'ordre fixé, sans préciser à partir de quelle taille n de l'échantillon l'approximation devient raisonnable. Cependant, dans de nombreux modèles, il est plus facile de construire des intervalles de confiance asymptotiques.

Remarquons que si le paramètre θ est un vecteur de dimension d avec $d > 1$, on peut généraliser la notion d'intervalle de confiance pour θ. Dans ce cas on cherchera plutôt une *région de confiance* $\mathcal{C}(\mathbf{X}) \subset \mathbb{R}^d$ qui contient θ avec probabilité au moins $1 - \alpha$.

1.4.2 Construction d'intervalle de confiance

On imagine assez facilement qu'un estimateur ponctuel $\hat{\theta}$ de θ sera un bon point de départ pour construire un intervalle de confiance pour θ. En effet, puisque $\hat{\theta}$ est censé de prendre des valeurs près de θ, il est naturel d'utiliser un voisinage du type $[\hat{\theta} - \hat{\delta}_1, \hat{\theta} + \hat{\delta}_2]$ de $\hat{\theta}$ comme intervalle de confiance. Afin de déterminer la taille exacte de ce voisinage (pour que l'intervalle de confiance soit de niveau $1 - \alpha$), la connaissance de la loi de l'estimateur est indispensable.

Un procédé assez général pour la construction d'intervalle de confiance repose sur l'utilisation de fonctions pivotales. Une fonction $\theta \mapsto \mathcal{T}(\hat{\theta}, \theta)$ dont la loi ne dépend pas du paramètre θ (ou d'autres paramètres inconnus du modèle) est dite **fonction pivotale** (ou **pivot**) pour le modèle statistique $\{\mathbb{P}_\theta, \theta \in \Theta\}$.

On procède de la manière suivante :

1. On détermine un estimateur ponctuel $\hat{\theta}$ de θ.
2. On détermine la loi de l'estimateur $\hat{\theta}$.
3. On cherche une transformation $\mathcal{T}(\hat{\theta}, \theta)$ de $\hat{\theta}$ dont la loi ne dépend plus de paramètres inconnus. Autrement dit, on cherche une fonction pivotale $\mathcal{T}(\hat{\theta}, \theta)$ dont on détermine la loi.
4. On choisit $\gamma_1 \in [0, 1]$ et $\gamma_2 \in [0, 1]$ tels que $\gamma_2 - \gamma_1 = 1 - \alpha$. On détermine les quantiles q_{γ_1} et q_{γ_2} d'ordre γ_1 et γ_2 de la loi de $\mathcal{T}(\hat{\theta}, \theta)$ tels que
$$\mathbb{P}_\theta \left(q_{\gamma_1} \leq \mathcal{T}(\hat{\theta}, \theta) \leq q_{\gamma_2} \right) = \gamma_2 - \gamma_1 = 1 - \alpha.$$
5. En "inversant" $\mathcal{T}$ (lorsque c'est possible...), on encadre alors θ par deux quantités aléatoires A et B, fonctions *uniquement* de $\hat{\theta}, q_{\gamma_1}, q_{\gamma_1}$ et de paramètres connus, telles que
$$\mathbb{P}_\theta \left(A \leq \theta \leq B \right) = 1 - \alpha.$$

Exemple. Loi normale

Appliquons cette démarche à un exemple : la construction d'un intervalle de confiance pour le paramètre μ de la loi normale $\mathcal{N}(\mu, 1)$ à partir d'un échantillon i.i.d. $\mathbf{X} = (X_1, \ldots, X_n)$ de loi $\mathcal{N}(\mu, 1)$. Dans ce modèle, l'estimateur du maximum de vraisemblance de μ est la moyenne empirique $\bar{X}$ (étape 1). On sait que $\bar{X}$ suit la loi normale $\mathcal{N}(\mu, 1/n)$ (étape 2). On en déduit que $\mathcal{T} = \sqrt{n}(\bar{X} - \mu)$ suit la loi normale standard $\mathcal{N}(0, 1)$, ce qui convient pour l'étape 3. Choisissons $\gamma_1 = \alpha/2$ et $\gamma_2 = 1 - \alpha/2$ et notons z_γ le quantile d'ordre γ de

la loi normale standard. On obtient alors

$$\begin{aligned} 1-\alpha &= \mathbb{P}(z_{\alpha/2} \leq \mathcal{T} \leq z_{1-\alpha/2}) \qquad \text{(étape 4)} \\ &= \mathbb{P}\left(z_{\alpha/2} \leq \sqrt{n}(\bar{X}_n - \mu) \leq z_{1-\alpha/2})\right) \\ &= \mathbb{P}\left(\bar{X}_n - \frac{z_{1-\alpha/2}}{\sqrt{n}} \leq \mu \leq \bar{X}_n - \frac{z_{\alpha/2}}{\sqrt{n}}\right). \end{aligned}$$

On en déduit que $\mathrm{IC}_{1-\alpha}(\mu) = \left[\bar{X}_n - z_{1-\alpha/2}/\sqrt{n}, \bar{X}_n - z_{\alpha/2}/\sqrt{n}\right]$ est un intervalle de confiance de taille $1-\alpha$ pour μ (étape 5). La longueur de cet intervalle est

$$\ell_{\mathrm{IC}} = 2z_{1-\alpha/2}/\sqrt{n},$$

car $-z_{\alpha/2} = z_{1-\alpha/2}$, par symétrie de la loi de la loi normale standard.

Il est courant de choisir $\gamma_1 = \alpha/2$ et $\gamma_2 = 1-\alpha/2$. En fait, quand la loi de la fonction pivotale $\mathcal{T}(\hat{\theta}, \theta)$ est symétrique et unimodale, ce choix minimise la longueur d'intervalle parmi tous les γ_1, γ_2 tels que $\gamma_1 + \gamma_2 = 1-\alpha$. Il permet donc de localiser la vraie valeur du paramètre θ avec plus de précision que tout intervalle asymétrique.

Le procédé décrit ci-dessus pour la construction d'intervalle de confiance n'est pas incontournable. D'autres approches sont possibles. Dans certaines situations on peut par exemple utiliser des inégalités comme l'inégalité de Markov ou de Hoeffding pour obtenir un intervalle de confiance.

Dans des nombreux cas, l'étape 2 et/ou l'étape 3 du procédé ci-dessus ne sont pas évidentes. Il peut s'avérer plus facile de considérer la *loi limite* de l'estimateur $\hat{\theta}_n$ au lieu de la loi pour n fini. Il faut alors trouver une transformation $\mathcal{T}_n(\hat{\theta}_n, \theta)$ telle que sa *loi limite* soit indépendante de tout paramètre inconnu, et puis, on pourra en déduire un intervalle de confiance *asymptotique.*

Chapitre 2

Bootstrap

Pour la construction d'intervalles de confiance par la méthode pivotale il est indispensable de connaître la loi de l'estimateur $\hat{\theta}$ (ou sa loi limite). Mais que faire lorsqu'elle est inconnue ? Une solution numérique est l'approche du bootstrap. Le bootstrap est une méthode de rééchantillonnage très similaire aux simulations de Monte-Carlo. La différence est que l'on simule selon la loi empirique des données.

Dans ce chapitre nous donnons d'abord une introduction général au bootstrap. Ensuite, plusieurs méthodes de bootstrap pour la construction d'intervalles de confiance sont présentées. Pour une introduction très accessible voir Efron and Tibshirani (1993).

2.1 Exemple introductif

On observe une réalisation $\mathbf{x} = (x_1, \ldots, x_n)$ du vecteur aléatoire $\mathbf{X} = (X_1, \ldots, X_n)$ qui est composé de n copies i.i.d. de la variable aléatoire

$$X = T + \theta, \tag{2.1}$$

où T suit la loi de Student t_q à q degrés de liberté et le paramètre de position $\theta \in \mathbb{R}$ est inconnu. Notons que la loi de X est symétrique par rapport à θ, puisque la loi de Student est symétrique.

Pour estimer un paramètre de position, on peut considérer (au moins) deux estimateur différents : la moyenne empirique $\bar{X}_n$ et la médiane empirique

$$M_n := \begin{cases} X_{(\frac{n+1}{2})}, & \text{si } n \text{ impair} \\ X_{(\frac{n}{2})}, & \text{si } n \text{ pair,} \end{cases}$$

où $X_{(1)}, \ldots, X_{(n)}$ désignent les statistiques d'ordre associées à l'échantillon $X_1, \ldots, X_n$ obtenues en classant les observations par ordre croissant, c'est-à-dire

$$X_{(1)} \leq \cdots \leq X_{(n)} \qquad \text{et} \qquad X_{(j)} \in \{X_1, \ldots, X_n\} \quad \forall j = 1, \ldots, n.$$

Lequel des estimateurs $\bar{X}_n$ et M_n est meilleur pour estimer θ ? On pourrait comparer leurs risques quadratiques. Pour cela, il faut connaître la loi des estimateurs, ce qui n'est pas évident pour la médiane empirique M_n. Un calcul explicit de son risque quadratique n'est pas possible. On peut alors utiliser des simulations de Monte-Carlo afin d'approcher la valeur de son risque quadratique. Comment faire précisément ?

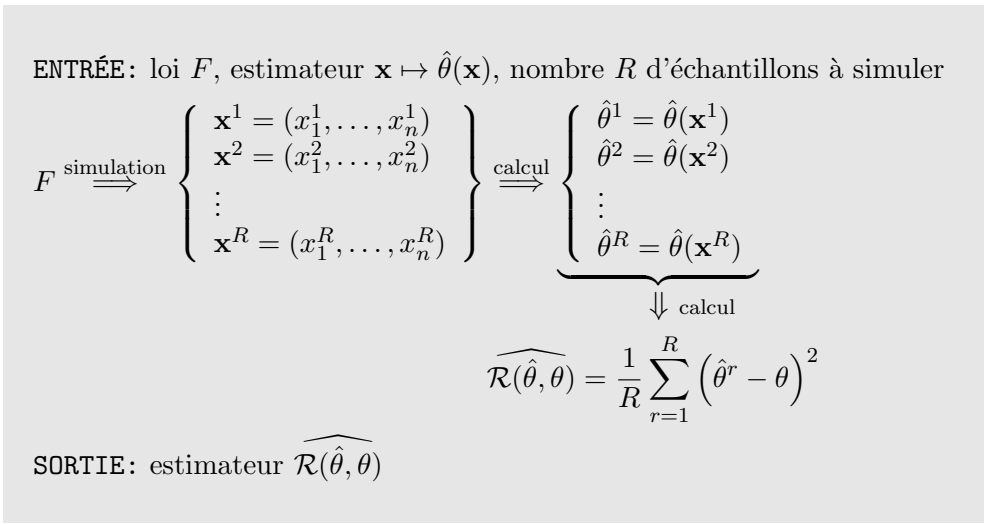


FIGURE 2.1 – Schéma des simulations de Monte Carlo pour l'estimation du risque quadratique $\mathcal{R}(\hat{\theta}, \theta)$ d'un estimateur $\hat{\theta}$.

2.1.1 Risque quadratique par Monte-Carlo

Rappelons que le risque quadratique d'un estimateur $\hat{\theta}$ de θ est défini par l'espérance

$$\mathcal{R}(\hat{\theta}, \theta) = \mathbb{E}_\theta[(\hat{\theta} - \theta)^2].$$

Notons $U := (\hat{\theta} - \theta)^2$. Si on arrive à simuler des copies i.i.d. U_j de U, on peut approcher le risque quadratique par la moyenne empirique des U_j. Or, la loi de U est inconnue, car la loi de $\hat{\theta}$ l'est. En revanche, on peut voir U comme une fonctionnelle des données $\mathbf{X}$, c'est-à-dire $U = H(\mathbf{X})$, car $\hat{\theta} = \hat{\theta}(\mathbf{X})$ et il est facile de générer des réalisations de $\mathbf{X} = (X_1, \ldots, X_n)$ avec des X_i définis en (2.1). Plus précisément, nous procédons de la façon suivante :

1. Tout d'abord on choisit les valeurs des paramètres θ et q et la taille d'échantillon n. Posons $r = 1$.
2. Ensuite on génère un échantillon, noté $\mathbf{x}^r = (x_1^r, \ldots, x_n^r)$, de taille n de la loi de $X = T + \theta$ où T suit la loi de Student t_q.
3. On évalue l'estimateur $\hat{\theta}$ sur l'échantillon $\mathbf{x}^r$: $\hat{\theta}^r = \hat{\theta}(\mathbf{x}^r)$.
   On incrémente r : $r = r + 1$.
4. On répète les étapes 2. et 3. R fois pour une valeur de R assez grande, donnant lieu à un échantillon de taille R de l'estimateur $\hat{\theta}$:
$$(\hat{\theta}^1, \ldots, \hat{\theta}^R).$$
5. On calcule le risque quadratique empirique associé à cet échantillon d'estimateurs :
$$\widehat{\mathcal{R}(\hat{\theta}, \theta)} = \frac{1}{R} \sum_{r=1}^{R} \left(\hat{\theta}^r - \theta \right)^2.$$

La mise en œuvre des simulations de Monte Carlo sous `R` est assez simple :

```
riskMC ← function(theta, ddl, nb.obs, REP = 1000){
  # simulation Monte-Carloa pour approcher le risque quadratique
  # de la moyenne et de la mediane empirique du parametre de
  # position d'une Student translatee
  # theta: parametre de position a estimer
  # ddl: nombre de degres de liberte de la loi de Student
  # nb.obs: taille d'echantillon
  # REP: nombre d'echantillons a simuler
  estim.moy ← rep(0,REP)
  estim.med ← rep(0,REP)
  for (i in 1:REP){
    # simuler un jeu de donnees:
    data ← rt(nb.obs, ddl) + theta
    # evaluer les estimateurs sur cet echantillon
    estim.moy[i] ← mean(data)
    estim.med[i] ← median(data)
  }
  # Risque quadratique
  risk ← c(mean((estim.moy - theta)^2),
           mean((estim.med - theta)^2))
  names(risk) ← c('risque.moyenne', 'risque.mediane')
  return(risk)
}
```

Dans les résultats ci-dessous, le paramètre θ est fixé à 5, la taille d'échantillon n vaut 100 et on génère `REP`=1000 échantillons à chaque appel. Quant au degré de liberté q de la loi de Student, il varie entre 1 et 50.

```
> riskMC(5, 1, 100)
risque.moyenne risque.mediane
  1.062213e+03   2.450985e-02
> riskMC(5, 1, 100)
risque.moyenne risque.mediane
  6.679630e+04   2.570921e-02
> riskMC(5, 3, 100)
risque.moyenne risque.mediane
    0.02935593     0.01875502
> riskMC(5, 3, 100)
risque.moyenne risque.mediane
    0.03035038     0.01843931
> riskMC(5, 5, 100)
risque.moyenne risque.mediane
    0.01582819     0.01608757
> riskMC(5, 10, 100)
risque.moyenne risque.mediane
    0.01205372     0.01477995
> riskMC(5, 50, 100)
risque.moyenne risque.mediane
   0.009011147    0.014845728
```

En fait, pour $q = 1$ la loi de Student conincïde avec la loi de Cauchy, qui est une loi à queues lourdes et qui n'est pas intégrable. Par conséquent, la moyenne empirique ne converge pas. Ainsi, on observe que les deux appels de `riskMC` avec $q = 1$ degré de liberté donnent deux valeurs assez différentes pour le risque quadratique de la moyenne empirique, alors que les valeurs du risque de la médiane empiriques sont stables. À partir de $q > 2$ la loi de Student est de variance finie, et donc le risque quadratique de la moyenne empirique est stable pour les deux appels de `riskMC` avec $q = 3$ degrés de liberté. Globalement, on observe que le

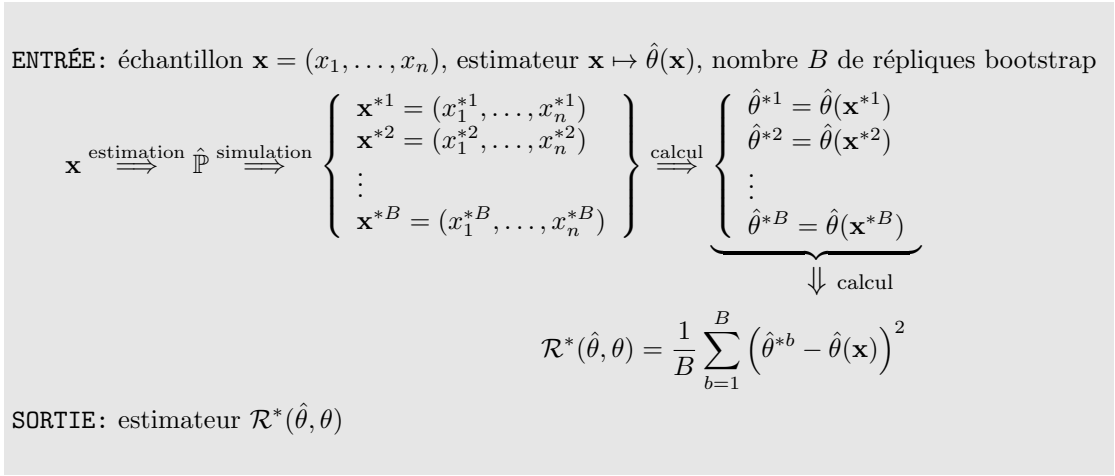


FIGURE 2.2 – Schéma bootstrap pour estimer le risque quadratique $\mathcal{R}(\hat{\theta}, \theta)$ d'un estimateur $\hat{\theta}$.

risque quadratique diminue pour les deux estimateurs lorsque le nombre q de degrés de liberté augmente. En effet, quand q tend vers l'infini la loi de Student converge vers la loi normale standard qui est une loi à queues légères.

Les résultats ci-dessus montrent que les deux estimateurs ont des comportements différents en fonction des queues de la loi. En effet, la médiane est nettement meilleur pour la loi de Cauchy en terme de risque quadratique, alors que la moyenne empirique a une petite avance pour des degrés de liberté élevées, c'est-à-dire pour des lois à queues légères.

Une v.a. T de loi de Student t_q avec $q > 2$ a moyenne 0 et variance $\mathbf{Var}(T) = q/(q-2)$. Donc, le risque quadratique de la moyenne empirique dans cet exemple est donnée par

$$\mathcal{R}(\tilde{\theta}, \theta) = \frac{q}{(q-2)n}.$$

Comparons les valeurs obtenues par des simulations de Monte Carlo ci-dessus avec les valeurs théoriques du risque quadratique. On observe que les estimations sont très proche des vraies valeurs :

```
> risk.moy.theo ← function(ddl, nb.obs) return(ddl/(ddl-2)/nb.obs)
> risk.moy.theo(3, 100)
[1] 0.03
> risk.moy.theo(5, 100)
[1] 0.01666667
> risk.moy.theo(10, 100)
[1] 0.0125
> risk.moy.theo(50, 100)
[1] 0.01041667
```

2.1.2 Risque quadratique par le Bootstrap

Nous venons de voir que lorsque la loi $\mathbb{P}$ des données est *connue*, on peut procéder par des simulations de Monte Carlo pour approcher le risque quadratique d'un estimateur. Mais comment faire quand la loi $\mathbb{P}$ est inconnue, ce qui est généralement le cas dans des applications ? Pour revenir à l'exemple précédent : que faire quand le degré de liberté q de la loi de Student est inconnu ? Comment savoir lequel des deux estimateurs, la moyenne empirique ou la médiane empirique, est préférable sur un échantillon observé ? En fait,

on peut adapter la méthode de Monte Carlo à ce scénario. Cette procédure est appelée le **bootstrap** et elle est décrite et mise en œuvre pour notre exemple dans ce paragraphe.

En toute généralité, le bootstrap est une *méthode de rééchantillonnage* pour approcher des caractéristiques de la loi d'un estimateur en utilisant rien d'autre que les données. Ici nous verrons qu'elle permet d'approcher le risque quadratique d'un estimateur.

Comme précédemment, notons $\mathbf{x} = (x_1, \dots, x_n)$ un échantillon i.i.d. de loi $\mathbb{P}$ *inconnue* et $\hat{\theta}$ un estimateur d'un paramètre θ. Le but est de déterminer le risque quadratique de $\hat{\theta}$ à partir d'un échantillon $\mathbf{x}$. Étant inconnue, on peut approcher la loi $\mathbb{P}$ par une loi estimée $\hat{\mathbb{P}}$, par exemple par la loi empirique donnée par la fonction de répartition empirique $\hat{F}$ associée aux observations $\mathbf{x}$. L'idée du bootstrap consiste à effectuer des simulations de Monte Carlo *en simulant de la loi estimée* $\hat{\mathbb{P}}$ (au lieu de la loi $\mathbb{P}$).

La démarche du bootstrap est la suivante :

1. On détermine une approximation $\hat{\mathbb{P}}$ de la loi $\mathbb{P}$ à partir de l'échantillon observé $\mathbf{x}$. Posons $b = 1$.
2. Ensuite on génère une réalisation, notée $\mathbf{x}^{*b} = (x_1^{*b}, \dots, x_n^{*b})$, de $\mathbf{X}$ de la loi $\hat{\mathbb{P}}$.
3. On évalue l'estimateur $\hat{\theta}$ sur l'échantillon $\mathbf{x}^{*b}$: $\hat{\theta}^{*b} = \hat{\theta}(\mathbf{x}^{*b})$. On incrémente $b = b+1$.
4. On répète les étapes 2. et 3. B fois pour une valeur de B assez grande, donnant lieu à un échantillon de taille B de l'estimateur $\hat{\theta}^* = \hat{\theta}(\mathbf{X})$ où $\mathbf{X}$ est de la loi $\hat{\mathbb{P}}$:

$$(\hat{\theta}^{*1}, \dots, \hat{\theta}^{*B}).$$

5. On calcule le risque quadratique empirique associé à cet échantillon d'estimateurs :

$$\mathcal{R}^*(\hat{\theta}, \theta) = \frac{1}{B} \sum_{b=1}^{B} \left(\hat{\theta}^{*b} - \hat{\theta}(\mathbf{x}) \right)^2 .$$

   Remarquons que la vraie valeur de θ étant inconnue, on la remplace par la valeur estimée sur les observations, à savoir $\hat{\theta}(\mathbf{x})$.

Cette démarche est également illustrée par la Figure 2.2. On remarque la grande similitude avec le schéma des simulations de Monte Carlo de la Figure 2.1.

Il est courant d'utiliser des étoiles * pour désigner des objets créés lors de la procédure bootstrap. En particulier, $\mathbf{x}^{*b}$ désigne un échantillon de la loi empirique $\hat{F}$, dit *échantillon bootstrap*, et $\hat{\theta}^{*b} = \hat{\theta}(\mathbf{x}^{*b})$ est appelé une *réplique bootstrap* de l'estimateur $\hat{\theta}$.

Pour faire du bootstrap il nous faut une approximation $\hat{\mathbb{P}}$ de la loi $\mathbb{P}$ des observations $\mathbf{x}$. On distingue deux approches :

— *Bootstrap paramétrique :* Si on a un modèle paramétrique $\{\mathbb{P}_\varphi, \varphi \in \Phi\}$ pour $\mathbb{P}$, on calcule un estimateur $\hat{\varphi}$ du paramètre φ et on utilise $\hat{\mathbb{P}} = \mathbb{P}_{\hat{\varphi}}$. Remarquons que le paramètre φ peut être identique au paramètre θ ou pas.

— *Bootstrap nonparamétrique :* On utilise la loi empirique $\hat{F}$ définie par la fonction de répartition empirique associée à $\mathbf{x}$ comme estimation de $\mathbb{P}$.

Simuler selon la loi empirique

Comment obtenir un échantillon bootstrap dans le cas du bootstrap nonparamétrique ? Autrement dit, comment générer des réalisations de la loi $\hat{F}$ associée aux observations $\mathbf{x} = (x_1, \dots, x_n)$? Nous avons vu que $\hat{F}$ est la loi discrète à valeurs dans $\{x_1, \dots, x_n\}$

qui associe le poids $1/n$ à chaque observation x_i. Plus précisément, soit X^* une variable aléatoire de la loi $\hat{F}$. Si toutes les observations x_i sont deux à deux distinctes, on a

$$\mathbb{P}_{\hat{F}}(X^* = x_i) = \frac{1}{n}, \quad i = 1, \ldots, n.$$

Donc, la loi empirique $\hat{F}$ est la loi uniforme discrète sur $\{x_1, \ldots, x_n\}$. En conséquence, simuler une réalisation de la fonction de répartition empirique $\hat{F}$ est équivalent à tirer un point x_i au hasard dans l'échantillon observé $\mathbf{x}$. On génère alors un échantillon $x_1^*, \ldots, x_n^*$ de la loi $\hat{F}$ par *tirage avec remise* de n valeurs dans l'échantillon observé $\mathbf{x}$.

Dans le cas général, si on admet la possibilité que l'échantillon $\mathbf{x}$ contient certaines valeurs plusieurs fois, ce qui est le cas lorsque F est une loi discrète, on a

$$\mathbb{P}_{\hat{F}}(X^* = x_i) = \frac{\#\{j : x_i = x_j\}}{n}, \quad i = 1, \ldots, n.$$

Un échantillon bootstrap est alors de la même taille que l'échantillon original à valeurs issues de l'échantillon original $\mathbf{x}$, mais avec des fréquences potentiellement différentes. On parle souvent de *rééchantillonnage* car on reconstruit un ensemble d'échantillons en partant de l'échantillon observé.

Sous `R` on peut tout simplement utiliser la fonction `sample()` avec l'option `replace = TRUE` pour générer des échantillons bootstrap.

Exemple. Risque quadratique par le bootstrap nonparamétrique

Reprenons l'exemple du début de ce chapitre : de l'estimation du risque quadratique de la moyenne empirique et de la médiane empirique pour estimer un paramètre θ de position. La fonction suivante met en œuvre le bootstrap nonparamétrique pour ce problème.

```
risk.boot ← function(data, B = 1000){
  # bootstrap nonparametrique pour estimer le risque quadratique
  # de la moyenne et de la mediane du parametre de position
  # d'une Student translatee
  # data: echantillon
  # B: nombre de repliques bootstrap
  nb.obs ← length(data)
  estim.moy ← rep(0, B)
  estim.med ← rep(0, B)
  for (i in 1:B){
    # tirer un echantillon bootstrap
    data.boot ← sample(data, nb.obs, replace = TRUE)
    # evaluer les repliques bootstrap
    estim.moy[i] ← mean(data.boot)
    estim.med[i] ← median(data.boot)
  }
  # Risque quadratique
  risk ← c(mean((estim.moy-mean(data))^2),
           mean((estim.med-median(theta))^2))
  names(risk) ← c('risque.moyenne', 'risque.mediane')
  return(risk)
}
```

Les deux fonctions `riskMC()` et `risk.boot()` sont très similaires. La différence principale est située à la première ligne de la boucle `for` dans la génération des échantillons `data` resp. `data.boot`.

L'exemple suivant montre qu'avec un échantillon de taille 100 on arrive à estimer les risques quadratiques des deux estimateurs avec une bonne précision sans connaître la loi $\mathbb{P}$ des observations ! On peut comparer les résultats aux valeurs obtenues par des simulations de Monte Carlo auparavant.

```
> data ← rt(100, 1) + 5
> risk.boot(data)
risque.moyenne risque.mediane
    1.49543404     0.03029092
> data ← rt(100, 1) + 5
> risk.boot(data)
risque.moyenne risque.mediane
    0.51335247     0.01697121
> data ← rt(100, 3) + 5
> risk.boot(data)
risque.moyenne risque.mediane
    0.04299717     0.01562267
> data ← rt(100, 3) + 5
> risk.boot(data)
risque.moyenne risque.mediane
    0.01994292     0.01157183
> data ← rt(100, 5) + 5
> risk.boot(data)
risque.moyenne risque.mediane
    0.01050988     0.01663342
> data ← rt(100, 10) + 5
> risk.boot(data)
risque.moyenne risque.mediane
    0.01402456     0.01070962
> data ← rt(100, 50) + 5
> risk.boot(data)
risque.moyenne risque.mediane
   0.008391878    0.017091082
```

2.2 Le principe du bootstrap

Dans ce paragraphe nous présentons le principe du bootstrap dans un cadre plus général et nous donnons d'autres exemples pour le bootstrap nonparamétrique.

Notons toujours $\mathbf{x} = (x_1, \ldots, x_n)$ une réalisation de $\mathbf{X} = (X_1, \ldots, X_n)$ où les X_i sont des variables aléatoires i.i.d. de loi F. Soit $T = T(\mathbf{X})$ un estimateur d'un paramètre ou d'une quantité θ inconnu. Notons $t = T(\mathbf{x})$ la valeur de l'estimateur T observé sur les données $\mathbf{x}$. L'objectif du bootstrap consiste à caractériser *la loi* de l'estimateur T. Par exemple, on souhaite déterminer la moyenne, la variance ou des quantiles de la loi de T.

En général, toute caractéristique$c$ de la loi de l'estimateur T s'écrit comme une fonctionnelle de la loi F, à savoir

$$c = c(F).$$

Le bootstrap propose d'approcher $c(F)$ par l'estimateur

$$c(\hat{F}),$$

(ou par $c_n(\hat{F}_n)$ avec $c_n \to c$ lorsque $n \to \infty$) où $\hat{F}$ est la fonction de répartition empirique dans le cas du bootstrap nonparamétrique.

Quelques exemples courants de caractéristiques de la loi de T :

— *Biais de l'estimateur T*

$$\beta := b(F) := \mathbb{E}_F[T] - \theta = \mathbb{E}[T|F] - \theta$$

— *Variance de T*

$$v := \mathbf{Var}_F(T) = \mathbf{Var}(T|F) = \mathbb{E}\left[(T - \mathbb{E}[T|F])^2|F\right]$$

— *Fonction de répartition de T en y*

$$G(y) := \mathbb{P}(T \leq y|F) = \mathbb{E}[\mathbb{1}\{T \leq y\}|F]$$

— *Quantile d'ordre α de la loi de T*

$$q_\alpha := G^{-1}(\alpha) = \inf\{y \in \mathbb{R} : G(y) \geq \alpha\} = \inf\{y \in \mathbb{R} : \mathbb{P}(T \leq y|F) \geq \alpha\}$$

Dans des applications, nous avons souvent des estimateurs dont nous ne connaissons pas la loi comme illustre l'exemple suivant.

Exemple. Données sur la population américaine.

Nous disposons d'un jeu de données sur la population de 49 villes américaines. Plus précisément, les observations sont bivariées $(x_i, y_i), i = 1, \ldots, n = 49$ où x_i et y_i désignent la population de la i-ème ville en 1920 et 1930, respectivement (Source : Davison and Hinkley (1997), p. 7). Nous supposons que les observations $(x_i, y_i), i = 1, \ldots, n$ sont des réalisations i.i.d. d'un vecteur aléatoire (X, Y) où X et Y représentent la population d'une ville en 1920 et 1930, respectivement.

La Figure 2.3 montre l'histogramme des lois marginales ainsi que le nuage des points (x_i, y_i). On observe une forte corrélation positive entre les deux variables. Les deux lois marginales pourraient être des lois Gamma, mais de paramètres différents. En revanche, il n'est pas évident de choisir un modèle paramétrique pour le couple (X, Y).

L'objectif de l'analyse de ces données est de comprendre la dynamique de croissance des villes. Pour cela on introduit l'indicateur de croissance θ défini par

$$\theta = \frac{\mathbb{E}[Y]}{\mathbb{E}[X]} = \frac{\int y \mathrm{d}F(x, y)}{\int x \mathrm{d}F(x, y)}.$$

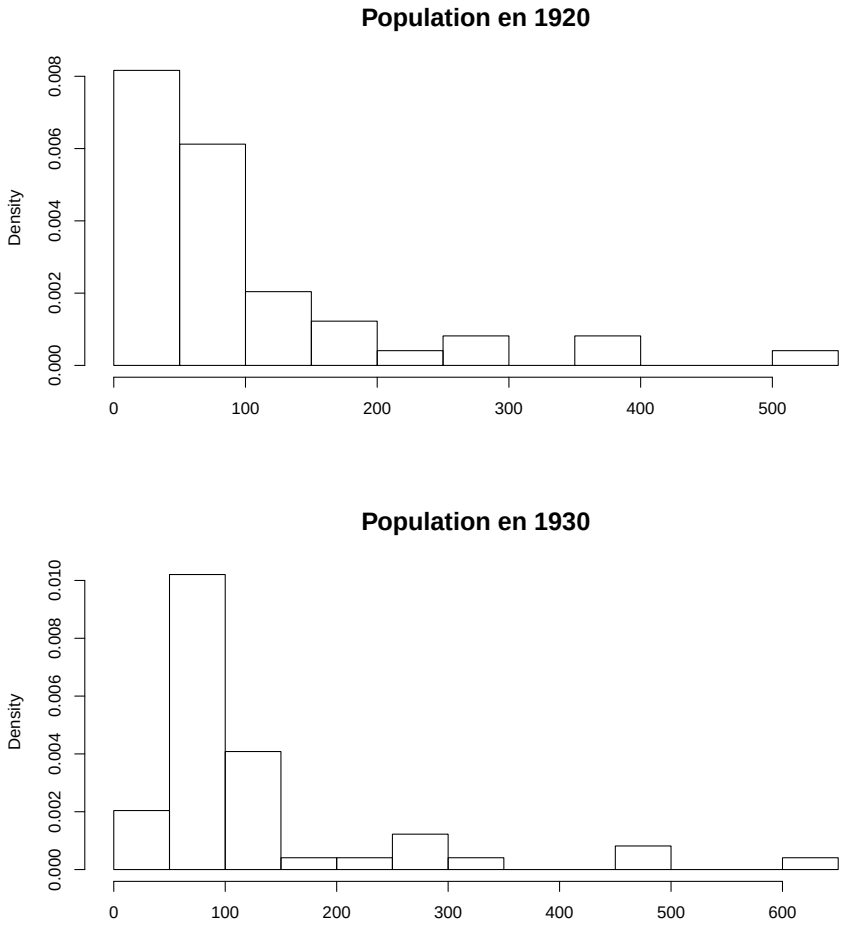
Un estimateur naturel de θ, obtenu par la méthode de substitution, est donnée par

$$T = \frac{\bar{Y}_n}{\bar{X}_n}.$$

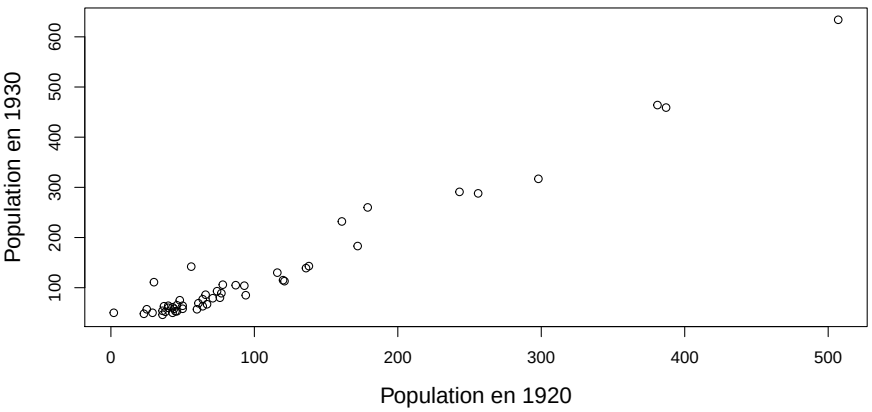
Vu que nous n'avons pas de modèle paramétrique pour la loi F des observations, nous pouvons effectuer du bootstrap nonparamétrique afin de déterminer le biais, la variance ou le risque quadratique de cet estimateur.

Estimateurs bootstrap

Notons $\mathbb{E}^*$ l'espérance conditionnelle sachant que $(X_i, \ldots, X_n) = (x_1, \ldots, x_n)$. Une réplique bootstrap de T est donnée par $T^* = T(X_1^*, \ldots, X_n^*)$ avec $(X_1^*, \ldots, X_n^*) \sim \hat{F}$ i.i.d. où $\hat{F}$ est



a)



b)

FIGURE 2.3 – Données sur la population de 49 villes américaines en 1920 et 1930 (Source : Davison and Hinkley (1997)).

une approximation de la loi F des données calculée à partir des observations $(x_1, \ldots, x_n)$. D'après la méthode de substitution, des estimateurs de β, v, $G(y)$ ou q_α sont donnés par

$$\begin{aligned}
\beta^* &= b(\hat{F}) = \mathbb{E}[T|\hat{F}] - T = \mathbb{E}^*[T^*] - t \\
v^* &= \mathbf{Var}(T|\hat{F}) = \mathbf{Var}^*(T^*) = \mathbb{E}^*\left[(T^* - \mathbb{E}^*[T^*])^2\right] \\
G^*(y) &= \mathbb{P}(T \leq y|\hat{F}) = \mathbb{P}^*(T^* \leq y) = \mathbb{E}^*[\mathbb{1}\{T^* \leq y\}] \\
q^*_\alpha &= \inf\{y \in \mathbb{R} : \mathbb{P}(T \leq y|\hat{F}) \geq \alpha\} = G^{*-1}(\alpha).
\end{aligned}$$

Les estimateurs β^*, v^*, $G^*(y)$ ou q^*_α sont appelés des **estimateurs bootstrap ideaux**. Seulement dans des cas particuliers on peut les calculer de façon explicite. Dans tous les autres cas, on les approche via des simulations. On procède comme suit :

1. Générer B échantillons bootstrap $(X^*_{b,1}, \ldots, X^*_{b,n})$, pour $b = 1, \ldots, B$ de loi $\hat{F}$.
2. Évaluer les répliques bootstrap $T^*_b = T(X^*_{b,1}, \ldots, X^*_{b,n})$ pour $b = 1, \ldots, B$.
3. Calculer les approximations suivantes :

$$\beta^*_B = \bar{T}^* - t, \qquad \text{où } \bar{T}^* = \frac{1}{B}\sum_{b=1}^{R} T^*_b \tag{2.2}$$

$$v^*_B = \frac{1}{B}\sum_{b=1}^{B}\left(T^*_b - \bar{T}^*\right)^2 \tag{2.3}$$

$$\begin{aligned}
G^*_B(y) &= \frac{1}{B}\sum_{b=1}^{B}\mathbb{1}\{T^*_b \leq y\} \\
q^*_{\alpha,B} &= G^{*-1}_B(\alpha).
\end{aligned}$$

Notons que G^*_B est la fonction de répartition empirique associée à $(T^*_1, \ldots, T^*_B)$.

Théorème 4. *Soit $\mathbf{x} = (x_1, \ldots, x_n)$ et $0 < \alpha < 1$. Le quantile empirique $\hat{q}_\alpha$ d'ordre α, défini par $\hat{q}_\alpha := \hat{F}^{-1}(\alpha) := \inf\{t \in \mathbb{R} : \hat{F}(t) \geq \alpha\}$, est donné par*

$$\hat{q}_\alpha = x_{(\lceil \alpha n \rceil)},$$

où $\lceil a \rceil$ désigne le plus petit entier supérieur ou égal à a.

Démonstration. cf. TD. □

On peut également montrer que les quantiles empiriques $\hat{q}^n_\alpha$ converge en probabilité vers les quantiles théoriques q^F_α de la loi F lorsque n tend vers l'infini.

Théorème 5. *Soient $X_1, X_2, \ldots$ des v.a. i.i.d. de loi F. Soit F strictement croissante en le quantile $q^F_\alpha := \inf\{t \in \mathbb{R} : F(t) \geq \alpha\}$ d'ordre α de F. Notons $\hat{q}^n_\alpha$ le quantile empirique d'ordre α associé à $(X_1, \ldots, X_n)$. Alors,*

$$\hat{q}^n_\alpha \xrightarrow{P} q^F_\alpha, \quad \text{lorsque } n \to \infty.$$

Démonstration. cf. TD. □

Exemple. Données sur la population américaine (suite).

Rappelons que les observations $(x_i, y_i), i = 1, \ldots, n$ sont considérées comme des réalisations i.i.d. de loi F inconnue. La quantité à estimer est l'indicateur de croissance θ défini par $\theta = \mathbb{E}[Y]/\mathbb{E}[X]$, et un estimateur naturel de θ est donné par $T = \bar{Y}_n/\bar{X}_n$. Nous effectuons du bootstrap nonparamétrique afin d'estimer le biais, la variance ou le risque quadratique de l'estimateur T. Ainsi, on génère des échantillons bootstrap nonparamétrique $\{(x^*_{b,i}, y^*_{b,i}), i = 1, \ldots, n\}$ pour $b = 1, \ldots, B$ en tirant des *couples* (x_i, y_i) avec remise dans $\{(x_i, y_i), i = 1, \ldots, n\}$. Ensuite, on évalue les répliques bootstrap T^*_b de T donnés par

$$T^*_b = \frac{\bar{y}^*_{b,n}}{\bar{x}^*_{b,n}} = \frac{\frac{1}{n}\sum_{i=1}^n y^*_{b,i}}{\frac{1}{n}\sum_{i=1}^n x^*_{b,i}}, \qquad b = 1, \ldots, B.$$

Enfin, on peut évaluer l'estimateur bootstrap du biais β^* et de la variance v^* de T par (2.2) et (2.3), respectivement. Le risque quadratique de T est estimé par $\mathcal{R}^*(T, \theta) = (\beta^*)^2 + v^*$.

Voyons comment ça marche précisément sous R. Pour mieux observer certains aspects du bootstrap, nous n'utiliserons pas la totalité des données, mais seulement les 10 premières observations.

Nous commençons par importer les données :

```
city ← read.csv('UScitypopulation.csv', header = TRUE, sep = '\t')
city ← city[1:10, ]
head(city)
    X   Y
1 138 143
2  93 104
3  61  69
4 179 260
5  48  75
6  37  63
attach(city)
# estimateur de theta = E[Y]/E[X]
theta.hat ← mean(Y)/mean(X)
theta.hat
[1] 1.520312
```

Créons d'abord un seul échantillon bootstrap :

```
# tirer n indices dans 1,...,n pour l'echantillon bootstrap
n ← nrow(city)
ind ← sample(1:n, n, replace = TRUE)
ind
[1] 10  6  6  9  7  1  2  5  9  8
```

On voit bien que l'indice 6, par exemple, est tiré plusieurs fois alors que les indices 3 et 4 n'y figurent pas. Enfin, on évalue la réplique bootstrap T^* sur cet échantillon bootstrap :

```
> theta.boot ← mean(Y[ind])/mean(X[ind])
> theta.boot
[1] 1.751606
```

La valeur de la réplique bootstrap (1.751606) est relativement près de la valeur de l'estimateur (1.520312) sur les données. Maintenant, répétons la même démarche 1000 fois afin de créer un ensemble de répliques bootstrap $T^*_b, b = 1, \ldots, 1000$ pour estimer le biais, la variance et le bootstrap :

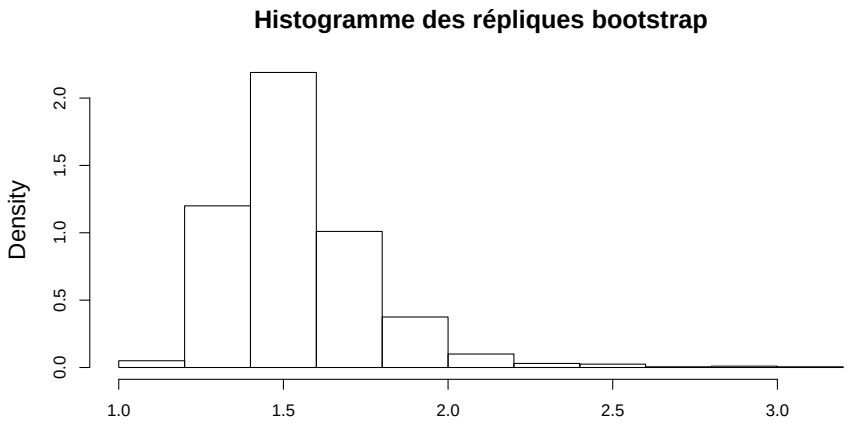


FIGURE 2.4 – Histogramme de 1000 répliques bootstrap de $T = \bar{Y}_n/\bar{X}_n$ pour un sous-ensemble de taille 10 des données sur la population américaine.

```
B ← 1000
theta.boot ← rep(NA, B)
for (i in 1:B){
   # generer un echantillon bootstrap
   ind ← sample(1:n, size = n, replace = TRUE)
   # replique bootstrap associee
   theta.boot[i] ← mean(Y[ind])/mean(X[ind])
 }
bias.boot ← mean(theta.boot) - theta.hat
bias.boot
[1] 0.02960482
var.boot ← var(theta.boot)
var.boot
[1] 0.04952014
risk.boot ← bias.boot^2 + var.boot
risk.boot
[1] 0.05039659
```

On peut également tracer l'histogramme des répliques bootstrap $T_b^*, b = 1, \ldots, 1000$, voir Figure 2.4. On observe que celui-ci est unimodal et asymétrique ; l'approcher par une loi normale serait grossier.

La boucle `for` dans le code ci-dessus pourrait être remplacé par un appel à la fonction `apply()`.

2.2.1 ERREURS

En utilisant le bootstrap, on effectue deux approximations qui donnent lieu à deux erreurs différentes. Rappelons que la cible est une caractéristique $c = c(F)$ de la loi de T. Selon la méthode de substitution du Chapitre 1, on l'approche par $c(\hat{F}_n)$ (ou $c_n(\hat{F}_n)$), l'estimateur bootstrap idéal, où $\hat{F}_n$ est une estimation (paramétrique ou non) de la loi F des données $\mathbf{x} = (x_1, \ldots, x_n)$. La différence entre $c(F)$ et $c(\hat{F}_n)$ est une *erreur statistique*, essentiellement due à l'estimation de F à partir d'un jeu de données. Pour que cette erreur soit petite, il faut que d'une part $\hat{F}_n$ soit un bon estimateur de F (p. ex. convergence uniforme

dans un sens approprié), ce qui est généralement le cas quand la taille d'échantillon n est grande, et d'autre part, que la fonctionnelle c soit suffisamment régulière (pour que $c(\hat{F}_n)$ ou $c_n(\hat{F}_n)$ se comporte bien). Pour établir des résultats comme la vitesse de convergence des estimateurs bootstrap ideaux, on utilise des développements d'Edgeworth (c'est technique et hors programme de ce cours).

L'estimateur bootstrap idéal $c(\hat{F}_n)$ est rarement connu explicitement. Et en général, il est également impossible de l'évaluer de façon exacte par un calcul numérique. Par conséquent, on utilise l'approximation par des *simulations* bootstrap pour obtenir l'estimateur bootstrap c_B^*, où B désigne le nombre d'échantillons bootstrap. Par la loi des grands nombres, on a, conditionnellement aux observations $\mathbf{x} = (x_1, \ldots, x_n)$,

$$c_B^* \xrightarrow{P} c(\hat{F}_n), \quad B \to \infty.$$

Remarquons qu'ici c'est le nombre B d'échantillons bootstrap qui tend vers l'infini, alors que la taille n d'échantillon reste fixée ! Ainsi, il suffit de choisir le nombre B d'échantillons bootstrap suffisamment grand, pour rendre l'erreur due aux simulations négligeable. Seul inconvénient, plus B est grand, plus le temps de calcul est long. En pratique, pour l'approximation du biais, de la variance ou du risque quadratique, il est suffisant de choisir le nombre B de répliques bootstrap entre 25 et 200.

Pour résumer, on a

$$c = c(\hat{F}) \underbrace{\approx}_{\substack{\text{erreur statistique} \\ \text{petite si } n \text{ grand}}} c(\hat{F}_n) \underbrace{\approx}_{\substack{\text{erreur de simulation} \\ \text{petite si } B \text{ est grand}}} c_B^*.$$

L'utilisateur a la main sur l'erreur de simulation, mais pas sur l'erreur statistique, car en général on ne choisit pas ses données.

2.2.2 Analyse de la moyenne empirique

Dans ce paragraphe, nous analysons l'impact du bootstrap nonparamétrique sur l'estimation du biais et de la variance de la moyenne empirique comme estimateur de l'espérance de la loi des données. La moyenne empirique est l'estimateur pour lequel les calculs de l'erreur de simulation due au bootstrap sont encore relativement simples. Les résultats permettent de nous guider dans le choix du nombre B d'échantillons bootstrap.

Fixons d'abord le cadre. Soit $\mathbf{x} = (x_1, \ldots, x_n)$ une réalisation de $\mathbf{X} = (X_1, \ldots, X_n)$ où les X_i sont des v.a. i.i.d. de loi F intégrable et inconnue. Considérons la moyenne empirique $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ comme estimateur de $\mathbb{E}[X_1]$. Notons $\bar{X}_b^* = \frac{1}{n}\sum_{i=1}^n X_{b,i}^*$ pour $b = 1, \ldots, B$ les répliques bootstrap de $\bar{X}_n$, où $X_{b,i}^*$ sont des v.a. i.i.d. de la loi empirique $\hat{F}$, où $\hat{F}$ désigne la fonction de répartition empirique associée à $\mathbf{x} = (x_1, \ldots, x_n)$.

Lemme 1. *Conditionnellement aux observations* $\mathbf{x} = (x_1, \ldots, x_n)$*, on a*

$$\mathbb{E}^*\left[\bar{X}_b^*\right] := \mathbb{E}\left[\bar{X}_b^* \mid \hat{F}, \mathbf{X} = \mathbf{x}\right] = \bar{x}_n \qquad \mathbf{Var}^*(\bar{X}_b^*) = \frac{s_{\mathbf{x}}^2}{n},$$

où $s_{\mathbf{x}}^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x}_n)^2$.

Démonstration. D'une part, on a

$$\mathbb{E}^*\left[\bar{X}_b^*\right] = \mathbb{E}^*\left[\frac{1}{n}\sum_{i=1}^n X_{b,i}^*\right] = \mathbb{E}^*\left[X_{1,i}^*\right] \qquad \text{(car les } X_{b,i}^* \text{ sont i.i.d.)}$$
$$= \mathbb{E}\left[X_{1,i}^* \mid \hat{F}, \mathbf{X} = \mathbf{x}\right] = \frac{1}{n}\sum_{i=1}^n x_i = \bar{x}_n,$$

d'après (1.5). D'autre part, on a

$$\begin{aligned}\mathbf{Var}^*(\bar{X}_b^*) &= \mathbf{Var}^*\left(\frac{1}{n}\sum_{i=1}^n X_{b,i}^*\right) = \frac{1}{n}\mathbf{Var}^*\left(X_{1,i}^*\right) \quad \text{(car les } X_{b,i}^* \text{ sont i.i.d.)} \\ &= \frac{1}{n}\left\{\mathbb{E}^*\left[(X_{1,i}^*)^2\right] - \left(\mathbb{E}^*\left[X_{1,i}^*\right]\right)^2\right\} \\ &= \frac{1}{n}\left\{\frac{1}{n}\sum_{i=1}^n x_i^2 - (\bar{x}_n)^2\right\} \\ &= \frac{s_{\mathbf{x}}^2}{n}.\end{aligned}$$

□

Proposition 1. *L'estimateur bootstrap* $\beta_B^* = \frac{1}{B}\sum_{b=1}^B \bar{X}_b^* - \bar{x}_n$ *du biais de la moyenne empirique* $\bar{X}_n$ *vérifie :*

(i) Conditionnellement aux observations $\mathbf{x} = (x_1, \ldots, x_n)$*, on a*

$$\mathbb{E}^*\left[\beta_B^*\right] = 0$$
$$\mathbf{Var}^*(\beta_B^*) = \frac{s_{\mathbf{x}}^2}{nB}.$$

(ii) Sans conditionnement, en prenant l'espérance sur $\mathbf{X}$*, on a*

$$\mathbb{E}\left[\beta_B^*\right] := \mathbb{E}_{\mathbf{X}}\left[\mathbb{E}^*\left[\beta_B^*\right]\right] = 0$$
$$\mathbf{Var}(\beta_B^*) = \frac{n-1}{n^2 B}\mathbf{Var}(X_1) \approx \frac{\mathbf{Var}(X_1)}{nB}.$$

On observe que le biais estimé par le bootstrap β_B^* de la moyenne empirique pour approcher $\mathbb{E}[X_1]$ est 0 en moyenne. Par ailleurs, la variance de β_B^* s'approche de 0 si le nombre B de répliques bootstrap est grand. Plus précisément, pour diviser la variance par deux, il faut doubler le nombre B d'échantillons bootstrap.

Démonstration. D'abord, on a

$$\mathbb{E}^*\left[\beta_B^*\right] = \mathbb{E}^*\left[\frac{1}{B}\sum_{b=1}^B \bar{X}_b^*\right] - \bar{x}_n = \mathbb{E}^*\left[\bar{X}_1^*\right] - \bar{x}_n = \bar{x}_n - \bar{x}_n = 0,$$

par Lemme 1. De même, puisque les $\bar{X}_b^*$ sont i.i.d., on a

$$\mathbf{Var}^*(\beta_B^*) = \mathbf{Var}^*\left(\frac{1}{B}\sum_{b=1}^B \bar{X}_b^*\right) = \frac{1}{B}\mathbf{Var}^*(\bar{X}_1^*) = \frac{s_{\mathbf{x}}^2}{nB}.$$

Maintenant, en prenant l'espérance en $\mathbf{X}$ (sans conditionnement), on obtient

$$\mathbb{E}\left[\beta_B^*\right] = \mathbb{E}_{\mathbf{X}}\left[\mathbb{E}^*\left[\beta_B^*\right]\right] = 0.$$

Enfin, comme β_B^* est centré, on trouve que

$$\begin{aligned}
\mathbf{Var}(\beta_B^*) &= \mathbb{E}\left[(\beta_B^*)^2\right] = \mathbb{E}_{\mathbf{X}}\left[\mathbb{E}^*\left[(\beta_B^*)^2\right]\right] \\
&= \mathbb{E}_{\mathbf{X}}\left[\mathbf{Var}^*(\beta_B^*) + \left(\mathbb{E}^*\left[\beta_B^*\right]\right)^2\right] \\
&= \mathbb{E}_{\mathbf{X}}\left[\frac{s_{\mathbf{X}}^2}{nB}\right] \qquad \text{(par Lemme 1)} \\
&= \frac{1}{nB}\mathbb{E}_{\mathbf{X}}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2\right] \\
&= \frac{1}{nB}\frac{n-1}{n}\mathbf{Var}(X_1),
\end{aligned}$$

où on a utilisé un résultat de la Feuille de TD n°6, Exercice 1. □

Proposition 2. *Sous l'hypothèse supplémentaire que les observations ont une loi normale, i.e. $X_i \sim \mathcal{N}(\mu, \sigma^2)$, l'estimateur bootstrap de la variance $v_B^* = \frac{1}{B}\sum_{b=1}^{B}\left(\bar{X}_b^* - \frac{1}{B}\sum_{b=1}^{B}\bar{X}_b^*\right)^2$ de la moyenne empirique $\bar{X}_n$ vérifie :*

(i) Conditionnellement aux observations $\mathbf{x} = (x_1, \ldots, x_n)$, on a

$$\begin{aligned}
\mathbb{E}^*\left[v_B^*\right] &= \frac{s_{\mathbf{x}}^2}{n}\frac{B-1}{B} \approx \frac{s_{\mathbf{x}}^2}{n} \\
\mathbf{Var}^*(v_B^*) &\approx \left(\frac{s_{\mathbf{x}}^2}{n}\right)^2 \frac{2}{B}\left(1 + \frac{1}{2}\kappa_4\right),
\end{aligned}$$

où κ_4 est le coefficient d'aplatissement ou kurtosis de la loi $\hat{F}$ défini par

$$\kappa_4 = \frac{\mathbb{E}^*\left[(X_1^* - \mathbb{E}^*[X_1^*])^4\right]}{\left(\mathbf{Var}^*(X_1^*)\right)^2} - 3.$$

(ii) Sans conditionnement, en prenant l'espérance en $\mathbf{X}$, on a

$$\begin{aligned}
\mathbb{E}\left[v_B^*\right] &= \frac{n-1}{n^2}\frac{B-1}{B}\sigma^2 \approx \frac{\sigma^2}{n} - \frac{\sigma^2}{nB} \\
\mathbf{Var}(v_B^*) &\approx \frac{2\sigma^4}{n^3} + \frac{2\sigma^4}{Bn^2}\left(1 + \frac{2}{n}\right).
\end{aligned}$$

Sans preuve.

On observe que l'espérance de la variance bootstrap $\mathbb{E}\left[v_B^*\right]$ est composé de deux termes : le premier terme, σ^2/n, est purement dû à l'approximation statistique de F par la loi empirique $\hat{F}$, car ce terme ne disparaît pas quand B tend vers l'infini. À l'opposé, le deuxième terme, $-\sigma^2/(nB)$, reflète l'erreur due aux simulations, et il s'annule quand B tend vers l'infini. D'ailleurs, on a bien

$$\mathbb{E}[v_B^*] \longrightarrow \mathbf{Var}(\bar{X}_n) = \frac{\sigma^2}{n}, \qquad B \to \infty.$$

De même, la variance de la variance bootstrap $\mathbf{Var}(v_B^*)$ se décompose en un terme dû à l'erreur statistique et en un terme dû à l'erreur de simulation. Comment interpréter le résultat ci-dessus ? Si on souhaite p. ex. que la variance de l'estimateur v_B^* due à l'erreur de simulation soit de l'ordre de 10% de la variance due à l'erreur statistique, il faut choisir $B = 10n$.

On peut montrer des résultats similaires pour d'autres types de loi et pour de nombreux estimateurs T.

2.3 Intervalles de confiance par le bootstrap

Pour la construction d'intervalles de confiance pour un paramètre θ par la méthode pivotale il est indispensable de connaître la loi de l'estimateur T (ou sa loi limite pour un intervalle asymptotique). Que fait-on si elle est inconnue ? Comme dans le cas du biais ou de la variance de d'un estimateur T, on peut utiliser le bootstrap pour s'en sortir.

De façon générale, la connaissance des quantiles q_α de la loi de $T-\theta$, où les quantiles sont tels que

$$\mathbb{P}_\theta(T-\theta \leq q_{\alpha/2}) = \frac{\alpha}{2} \quad \text{et} \quad \mathbb{P}_\theta(T-\theta \geq q_{1-\alpha/2}) = \frac{\alpha}{2},$$

permettent de déduire un intervalle de confiance de niveau $1-\alpha$. En effet,

$$\begin{aligned} 1-\alpha &= 1 - \left\{\mathbb{P}_\theta(T-\theta \leq q_{\alpha/2}) + \mathbb{P}_\theta(T-\theta \geq q_{1-\alpha/2})\right\} \\ &= 1 - \mathbb{P}_\theta\left(\left\{T-\theta \leq q_{\alpha/2}\right\} \cup \left\{T-\theta \geq q_{1-\alpha/2}\right\}\right) \\ &= \mathbb{P}_\theta\left(T-\theta \in [q_{\alpha/2}, q_{1-\alpha/2}]\right) \\ &= \mathbb{P}_\theta\left(\theta \in [T-q_{1-\alpha/2}, T-q_{\alpha/2}]\right). \end{aligned}$$

On en déduit qu'un intervalle de confiance pour θ de niveau $1-\alpha$ est donné par

$$[t-q_{1-\alpha/2}, t-q_{\alpha/2}],$$

où $t = T(\mathbf{x})$ est l'estimateur évalué sur les observations $\mathbf{x}$.

On voit bien l'importance de la connaissance des quantiles q_α de la loi de $T-\theta$. Quand on ne les connaît pas, le bootstrap peut servir pour les approcher. Nous présentons plusieurs approches bootstrap pour le faire.

2.3.1 Approximation normale

La première méthode bootstrap est inspirée par l'intervalle de confiance asymptotique pour la moyenne empirique. Rappelons que pour un échantillon $\mathbf{X} = (X_1, \ldots, X_n)$ i.i.d. de loi F avec moyenne μ et variance finie, on a par le théorème central limite,

$$\frac{\bar{X}_n - \mu}{\sqrt{s^2_{\mathbf{X}}/n}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1), \quad n \to \infty.$$

où $s^2_{\mathbf{X}} = \sqrt{\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2}$ désigne la variance empirique associée aux données $\mathbf{X}$. Par le lemme de Slutsky, l'intervalle

$$\mathrm{IC}_{1-\alpha}(\mu) = \left[\bar{X} - z_{1-\alpha/2}\sqrt{\frac{s^2_{\mathbf{X}}}{n}}, \bar{X} - z_{\alpha/2}\sqrt{\frac{s^2_{\mathbf{X}}}{n}}\right],$$

où z_α désigne le quantile d'ordre α de la loi normale standard, est un intervalle de confiance asymptotique de niveau $1-\alpha$ pour la moyenne μ. Remarquons que le terme $s^2_{\mathbf{X}}/n$ est un estimateur de la variance de la moyenne empirique $\bar{X}_n$.

Or, considérons un contexte plus général d'un estimateur T de θ. Dans de nombreux cas, comme p. ex. la médiane empirique et la moyenne tronquée, la loi limite de T est gaussienne, à savoir

$$T - \theta \dot{\sim} \mathcal{N}(0, v),$$

où $v > 0$ représente la variance de l'estimateur T et la notation $A_n \dot{\sim} B_n$ signifie que A_n a approximativement la même loi que B_n lorsque la taille n d'échantillon est grand. Ainsi, il découle un intervalle de confiance asymptotique du calcul suivant :

$$\begin{aligned} 1-\alpha &\approx \mathbb{P}_\theta\left(\frac{T-\theta}{\sqrt{v}} \in [z_{\alpha/2}, z_{1-\alpha/2}]\right) \\ &= \mathbb{P}_\theta\left(\theta \in \left[T-\sqrt{v}z_{1-\alpha/2}, T-\sqrt{v}z_{\alpha/2}\right]\right). \end{aligned}$$

En général, la variance v de l'estimateur T est inconnue et doit être estimée. Nous avons vu précédemment que le bootstrap peut fournir un tel estimateur, noté v_B^*, de v.

En pratique, on obtient des intervalles encore plus exacts, si on inclut une correction du biais. Ainsi, on écrit

$$T - \theta \dot{\sim} \mathcal{N}(\beta, v), \tag{2.4}$$

avec $\beta \in \mathbb{R}$ inconnu. Un intervalle de confiance asymptotique de niveau $1-\alpha$ est alors donné par

$$\left[T-\beta-z_{1-\alpha/2}\sqrt{v}, T-\beta-z_{\alpha/2}\sqrt{v}\right].$$

Comme la variance v de l'estimateur T, on peut estimer le biais β de T par le bootstrap, noté β_B^*. On en déduit un premier l'**intervalle bootstrap par approximation normale** $\mathcal{I}_{\text{norm}}^*$ défini par

$$\mathcal{I}_{\text{norm}}^* = \left[t-\beta_B^*-z_{1-\alpha/2}\sqrt{v_B^*}, t-\beta_B^*-z_{\alpha/2}\sqrt{v_B^*}\right],$$

où $t = T(\mathbf{x})$ est la valeur de l'estimateur T sur les observations $\mathbf{x}$.

L'intervalle $\mathcal{I}_{\text{norm}}^*$ est un bon intervalle de confiance pour θ *si* l'approximation normale de la loi $T-\theta$ supposée en (2.4) est justifiée. En pratique, on vérifie cette hypothèse sur l'échantillon des répliques bootstrap $T_b^*, b = 1, \ldots, B$ en traçant l'histogramme ou un QQ-plot.

Exemple. Données sur la population américaine (suite).

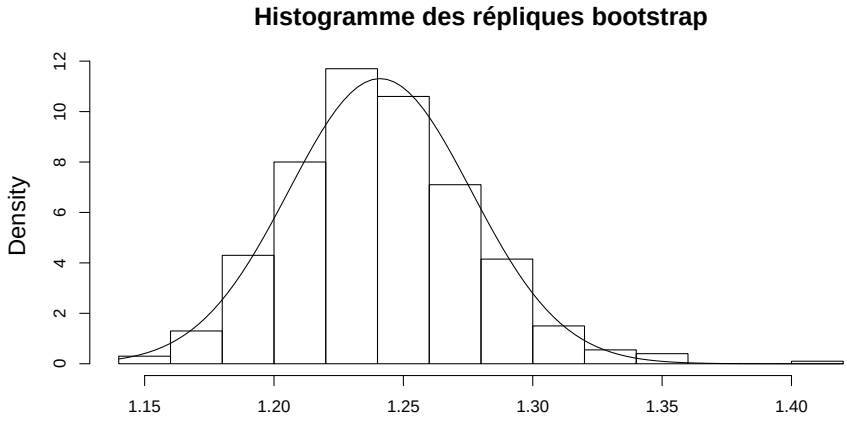
Pour les données sur la population de 49 villes américaines vues en Section 2.2, on peut calculer l'intervalle bootstrap $\mathcal{I}_{\text{norm}}^*$ pour $\theta = \mathbb{E}[Y]/\mathbb{E}[X]$. L'estimateur considéré est $T = \bar{Y}_n/\bar{X}_n$. En utilisant un sous-ensemble de taille 10 des données, Figure 2.4 montre l'histogramme des répliques bootstrap $T_b^*, b = 1, \ldots, B = 1000$ qui n'a pas l'air très gaussien à cause de son asymétrie. Ce manque de normalité indique que l'intervalle bootstrap $\mathcal{I}_{\text{norm}}^*$ ne sera pas un bon intervalle de confiance pour θ. En revanche, en utilisant toutes les 49 observations, l'histogramme des répliques bootstrap T_b^* devient bien plus symétrique et le QQ-plot qui compare la loi empirique des répliques bootstrap T_b^* à la loi normale indique que l'approximation est relativement bonne, voir Figure 2.5. L'intervalle bootstrap $\mathcal{I}_{\text{norm}}^*$ devrait être de bonne qualité, et il vaut

$$\mathcal{I}_{\text{norm}}^* = [1.17, 1.31].$$

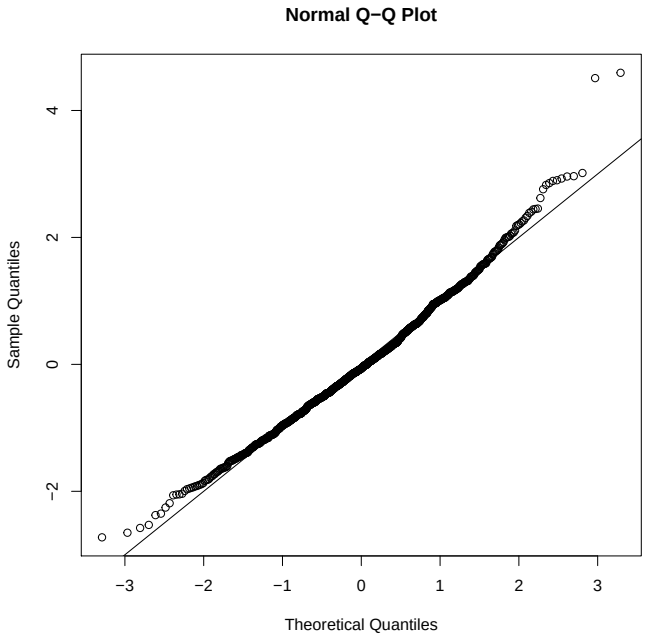
2.3.2 Intervalle bootstrap de base

Une approche différente consiste à employer le bootstrap pour estimer directement les quantiles de la loi de $T-\theta$, que l'on notera q_α. En effet, l'intervalle

$$\left[T-q_{1-\alpha/2}, T-q_{\alpha/2}\right]$$



a)



b)

FIGURE 2.5 – Pour l'ensemble des 49 observations sur la population américaine a) Histogramme de 1000 répliques bootstrap T_b^* de $T = \bar{Y}_n/\bar{X}_n$ et densité d'une la loi normale et b) QQ-plot pour comparaison de la distribution des $T_b^*, b = 1, \ldots, B = 1000$ à la loi normale.

ENTRÉE :

$\mathbf{x} = (x_1, \ldots, x_n)$	échantillon
$\mathbf{x} \mapsto T(\mathbf{x})$	estimateur de θ
α	intervalle de confiance de niveau nominal $1 - \alpha$
B	nombre d'échantillons bootstrap

1. On évalue l'estimateur T sur les observations $\mathbf{x}$: $t = T(\mathbf{x})$.
2. `for (b = 1:B)`
   (a) On génère un échantillon bootstrap $\mathbf{x}_b^* = (x_{b,1}^*, \ldots, x_{b,n}^*)$ par tirage avec remise dans l'échantillon $\mathbf{x}$.
   (b) On évalue l'estimateur T sur l'échantillon bootstrap : $T_b^* = T(\mathbf{x}_b^*)$.
3. On détermine les statistiques d'ordre $T^*_{(\lceil \alpha B/2 \rceil)}$ et $T^*_{(\lceil (1-\alpha/2)B \rceil)}$ associées à $T_b^*, b = 1, \ldots, B$.

SORTIE : Intervalle bootstrap de base :

$$\mathcal{I}^*_{\text{basic}} = \left[2t - T^*_{(\lceil (1-\alpha/2)B \rceil)}, 2t - T^*_{(\lceil \alpha B/2 \rceil)}\right].$$

FIGURE 2.6 – Algorithme pour le calcul de l'intervalle bootstrap de base.

est un intervalle de confiance de niveau $1 - \alpha$ pour θ, car

$$\mathbb{P}_\theta\left(\theta \in \left[T - q_{1-\alpha/2}, T - q_{\alpha/2}\right]\right) = \mathbb{P}_\theta\left(T - \theta \in \left[q_{\alpha/2}, q_{1-\alpha/2}\right]\right) = 1 - \alpha.$$

Notons $q^*_{\alpha,B}$ le quantile empirique d'ordre α associé aux répliques bootstrap $T_b^* - t, b = 1, \ldots, B$. On en déduit l'**intervalle bootstrap de base** $\mathcal{I}^*_{\text{basic}}$ défini par

$$\mathcal{I}^*_{\text{basic}} = \left[t - q^*_{1-\alpha/2,B}, t - q^*_{\alpha/2,B}\right].$$

D'après le Théorème 4, le quantile empirique $q^*_{\alpha,B}$ d'ordre α associé aux répliques bootstrap $T_b^* - t, b = 1, \ldots, B$ est donné par

$$q^*_{\alpha,B} = T^*_{(\lceil \alpha B \rceil)} - t,$$

où $T^*_{(m)}$ désigne la m-ième statistique d'ordre associé à $T_b^*, b = 1, \ldots, B$. Par conséquent, l'intervalle boostrap de base se réécrit comme

$$\mathcal{I}^*_{\text{basic}} = \left[2t - T^*_{(\lceil (1-\alpha/2)B \rceil)}, 2t - T^*_{(\lceil \alpha B/2 \rceil)}\right].$$

En pratique, on observe que cette approche échoue lorsque la loi de $T - \theta$ n'est pas approximativement pivotale, ce qui veut dire qu'il faut que la loi de $T - \theta = T(\mathbf{X}) - \theta$ avec $\mathbf{X} \sim F$ soit approximativement la même que celle de $T^* - t = T(\mathbf{X}^*) - t$ avec $\mathbf{X}^* \sim \hat{F}$. Autrement dit, les quantiles bootstrap $q^*_{\alpha,B}$ n'ont rien avoir avec les quantiles q_α recherchés.

Le schéma de l'algorithme pour le calcul de l'intervalle bootstrap de base est donné dans la Figure 2.6.

2.3.3 Intervalle bootstrap studentisé

En combinant les deux premières approches, on peut construire un troisième intervalle de confiance bootstrap avec des meilleures performances. Nous avons vu que le problème

ENTRÉE :

$\mathbf{x} = (x_1, \ldots, x_n)$	échantillon
$\mathbf{x} \mapsto T(\mathbf{x})$	estimateur de θ
α	intervalle de confiance de niveau nominal $1 - \alpha$
B, C	nombre d'échantillons bootstrap

1. On évalue l'estimateur T sur les observations $\mathbf{x}$: $t = T(x_1, \ldots, x_n)$.
2. `for (b = 1:B)`
   (a) On génère un échantillon bootstrap $\mathbf{x}_B^* = (x_{b,1}^*, \ldots, x_{b,n}^*)$ par tirage avec remise dans l'échantillon $\mathbf{x}$.
   (b) On évalue la réplique bootstrap de l'estimateur $T_b^* = T(\mathbf{x}_b^*)$.
   (c) `for (c = 1:C)`
      (i) On génère un échantillon bootstrap $(x_{c,1}^{**}, \ldots, x_{c,n}^{**})$ en tirant avec remise dans $(x_{b,1}^*, \ldots, x_{b,n}^*)$.
      (ii) On évalue la réplique bootstrap correspondant de l'estimateur : $T_c^{**} = T(x_{c,1}^{**}, \ldots, x_{c,n}^{**})$.
   (d) On estime la variance v_b^* de T_b^* par

$$v_{b,C}^* = \frac{1}{C} \sum_{c=1}^{C} \left(T_c^{**} - \frac{1}{C} \sum_{c=1}^{C} T_c^{**} \right)^2 .$$

   (e) On évalue la réplique bootstrap de $\mathcal{Z}$ définie par

$$\mathcal{Z}_b^* = \frac{T_b^* - t}{\sqrt{v_{b,C}^*}}.$$

3. On détermine les quantiles empiriques $q_{\alpha/2,B}^{\mathcal{Z}*}$ et $q_{1-\alpha/2,B}^{\mathcal{Z}*}$ d'ordre $\alpha/2$ et $1 - \alpha/2$ associés à l'échantillon $(\mathcal{Z}_1^*, \ldots, \mathcal{Z}_B^*)$:

$$q_{\alpha/2,B}^{\mathcal{Z}*} = \mathcal{Z}_{(\lceil B\alpha/2 \rceil)}^* \quad \text{et} \quad q_{1-\alpha/2,B}^{\mathcal{Z}*} = \mathcal{Z}_{(\lceil B(1-\alpha/2) \rceil)}^* .$$

4. On calcule l'estimateur bootstrap de la variance v de T :

$$v_B^* = \frac{1}{B} \sum_{b=1}^{B} \left(T_b^* - \frac{1}{B} \sum_{b=1}^{B} T_b^* \right)^2 .$$

SORTIE : Intervalle bootstrap studentisé :

$$\mathcal{I}_{\text{stud}}^* = \left[t - \sqrt{v_B^*} q_{1-\alpha/2,B}^{\mathcal{Z}*}, t - \sqrt{v_B^*} q_{\alpha/2,B}^{\mathcal{Z}*} \right].$$

FIGURE 2.7 – Algorithme pour le calcul de l'intervalle bootstrap studentisé

de l'intervalle bootstrap de base réside dans le fait que $T - \theta$ n'est généralement pas une statistique pivotale. Or, en divisant par l'écart-type $\sqrt{v}$ de l'estimateur T, on arrive à stabiliser la variance, et la statistique

$$\mathcal{Z} = \frac{T - \theta}{\sqrt{v}}$$

est souvent pivotale, c'est-à-dire la loi de $\mathcal{Z}$ sous F est près de la loi de $\mathcal{Z}$ sous $\hat{F}$. En fait, $\mathcal{Z}$ est la même statistique que celle considérée pour l'intervalle bootstrap par approximation normale. Cependant, maintenant, nous ne supposons plus que sa loi est une loi normale et nous utilisons directement le bootstrap pour approcher cette loi, en particulier, pour estimer ses quantiles afin de construire un intervalle de confiance.

Plus précisément, en notant $q_\alpha^{\mathcal{Z}}$ les quantiles de la loi de $\mathcal{Z}$, un intervalle de confiance de niveau $1-\alpha$ pour θ est donné par

$$\left[T - q_{1-\alpha/2}^{\mathcal{Z}}\sqrt{v}, T - q_{\alpha/2}^{\mathcal{Z}}\sqrt{v}\right].$$

Les quantités inconnues dans cet intervalle sont la variance v de T et les quantiles $q_\alpha^{\mathcal{Z}}$ de $\mathcal{Z}$. La variance peut être approchée par l'estimateur bootstrap habituel v_B^*. Pour l'estimation des quantiles $q_\alpha^{\mathcal{Z}}$, il faut générer un échantillon bootstrap de la statistique $\mathcal{Z}$. Autrement dit, on crée des répliques bootstrap $\mathcal{Z}_b^*, b = 1, \ldots, B$ définies par

$$\mathcal{Z}_b^* = \frac{T_b^* - t}{\sqrt{v_{b,C}^*}},$$

où $v_{b,C}^*$ désigne un estimateur de la variance de l'estimateur T_b^* (et non de T !). Pour obtenir un tel estimateur il faut effectuer un *autre* bootstrap pour chaque $b = 1, \ldots, B$. Ainsi, soit $(x_{b,1}^*, \ldots, x_{b,n}^*)$ l'échantillon bootstrap à l'iteration b sur lequel on calcule l'estimateur bootstrap $T_b^* = T(x_{b,1}^*, \ldots, x_{b,n}^*)$. Pour estimer la variance v_b^* de T_b^*, on procède ainsi :

On crée C échantillons bootstrap $(x_{c,1}^{**}, \ldots, x_{c,n}^{**})$ en tirant avec remise dans l'échantillon bootstrap $(x_{b,1}^*, \ldots, x_{b,n}^*)$ actuel (et non dans $(x_1, \ldots, x_n)$) ! On évalue les répliques bootstrap $T_c^{**} = T(x_{c,1}^{**}, \ldots, x_{c,n}^{**})$ et on estime la variance de T_b^* par

$$v_{b,C}^* = \frac{1}{C}\sum_{c=1}^{C}\left(T_c^{**} - \frac{1}{C}\sum_{c=1}^{C}T_c^{**}\right)^2.$$

Or, des estimateurs bootstrap des quantiles $q_{\alpha/2}^{\mathcal{Z}}$ et $q_{1-\alpha/2}^{\mathcal{Z}}$ sont donnés par

$$q_{\alpha/2,B}^{\mathcal{Z}*} = \mathcal{Z}_{(\lceil B\alpha/2\rceil)}^* \quad \text{et} \quad q_{1-\alpha/2,B}^{\mathcal{Z}*} = \mathcal{Z}_{(\lceil B(1-\alpha/2)\rceil)}^*.$$

Au final on obtient l'**intervalle bootstrap studentisé** défini par

$$\mathcal{I}_{\text{stud}}^* = \left[t - \sqrt{v_B^*}q_{1-\alpha/2,B}^{\mathcal{Z}*}, t - \sqrt{v_B^*}q_{\alpha/2,B}^{\mathcal{Z}*}\right].$$

Parfois cet intervalle est aussi appelé *intervalle bootstrap-t*.

Le schéma de l'algorithme du calcul de cet intervalle est donné en Figure 2.7.

La technique qui consiste à effectuer un autre bootstrap à l'intérieur du bootstrap principal est appelée *double bootstrap*. Le nombre total d'échantillons bootstrap de taille n à générer afin de calculer un intervalle bootstrap studentisé s'elève à BC, ce qui est vite très coûteux en terme de calcul. Si par exemple $B = 1000$ et $C = 100$, cela fait 100 000 échantillons à simuler. En utilisant la fonction d'influence, d'autres estimateurs de la variance de T_b^* peuvent être construits qui sont moins chers en termes de calcul.

Il est possible d'intégrer une correction du biais dans l'intervalle bootstrap studentisé, mais elle apporte rarement une amélioration significative en pratique.

En général, l'intervalle bootstrap studentisé produit des meilleurs résultats que l'intervalle bootstrap de base à condition que l'estimateur de la variance soit de bonne qualité. Le double bootstrap ne donne pas toujours des résultats très stables surtout quand la taille n d'échantillon est faible. En fait la statistique $\mathcal{Z}$ est plus souvent pivotale que la statistique $T - \theta$ utilisé pour l'intervalle bootstrap de base, c'est-à-dire la loi de $\mathcal{Z}$ sous F est égale à (ou près de) la loi de $\mathcal{Z}$ sous $\hat{F}$. Par conséquent, les quantiles bootstrap $q_{\alpha,B}^{\mathcal{Z}*}$ sont des estimateurs adéquats des quantiles théoriques $q_\alpha^{\mathcal{Z}}$.

On sait que les intervalles bootstrap studentisés $\mathcal{I}_{\text{stud}}^*$ sont particulièrement bien lorsque T est un estimateur d'un paramètre de position ou de la tendance centrale de la loi des observations X_i, comme par exemple la moyenne empirique, la moyenne tronquée ou la médiane empirique.

2.3.4 Intervalle bootstrap par transformation du paramètre

Lorsque $T - \theta$ ne suit pas de loi normale, il se peut qu'il existe une transformation h telle que $h(T) - h(\theta)$ a une loi normale. Plus précisément, notons $U = h(T)$ et $\eta = h(\theta)$ et supposons que h est une fonction strictement croissante telle que

$$U - \eta = h(T) - h(\theta) \dot{\sim} \mathcal{N}(0, w),$$

avec une constante $w > 0$. Alors, on pourrait calculer l'intervalle bootstrap par approximation normale pour η donné par

$$\left[u - \sqrt{w_B^*} z_{1-\alpha/2}, u - \sqrt{w_B^*} z_{\alpha/2}\right],$$

où $u = h(t)$ et w_B^* désigne l'estimateur bootstrap de la variance de $U = h(T)$ donné par

$$w_B^* = \frac{1}{B} \sum_{b=1}^{B} \left(U_b^* - \frac{1}{B} \sum_{b=1}^{B} U_b^* \right),$$

où $U_b^* = h(T_b^*) = h(T(x_{b,1}^*, \ldots, x_{b,n}^*))$ pour $b = 1, \ldots, B$ sont des répliques bootstrap de U. Autrement dit, on a

$$\begin{aligned} 1 - \alpha &\approx \mathbb{P}\left(\eta \in \left[U - \sqrt{w_B^*} z_{1-\alpha/2}, U - \sqrt{w_B^*} z_{\alpha/2}\right]\right) \\ &= \mathbb{P}\left(h(\theta) \in \left[h(T) - \sqrt{w_B^*} z_{1-\alpha/2}, h(T) - \sqrt{w_B^*} z_{\alpha/2}\right]\right) \\ &= \mathbb{P}\left(\theta \in \left[h^{-1}\left(h(T) - \sqrt{w_B^*} z_{1-\alpha/2}\right), h^{-1}\left(h(T) - \sqrt{w_B^*} z_{\alpha/2}\right)\right]\right), \end{aligned}$$

car h est strictement croissante. Il en découle un intervalle bootstrap pour θ donné par

$$\left[h^{-1}\left(h(t) - \sqrt{w_B^*} z_{1-\alpha/2}\right), h^{-1}\left(h(t) - \sqrt{w_B^*} z_{\alpha/2}\right)\right]. \tag{2.5}$$

En pratique, afin de vérifier que $h(T)$ suit une loi normale, on trace le QQ-plot des répliques bootstrap $U_1^*, \ldots, U_B^*$.

De même, on construit un intervalle bootstrap par transformation de paramètre en utilisant l'intervalle bootstrap de base. Ainsi, notons

$$\left[2u - U^*_{(\lceil B(1-\alpha/2)\rceil)}, 2u - U^*_{(\lceil B\alpha/2\rceil)}\right]$$

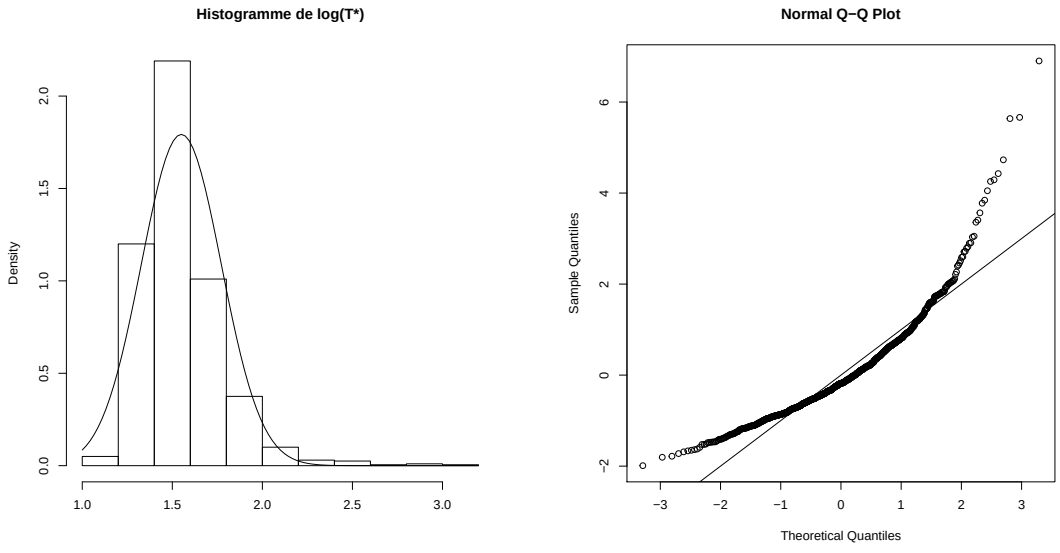
l'intervalle boostrap de base pour η. On obtient un intervalle bootstrap pour θ par

$$\left[h^{-1}\left(2h(t) - h(T^*_{(\lceil B(1-\alpha/2)\rceil)})\right), h^{-1}\left(2h(t) - h(T^*_{(\lceil B\alpha/2\rceil)})\right)\right], \tag{2.6}$$

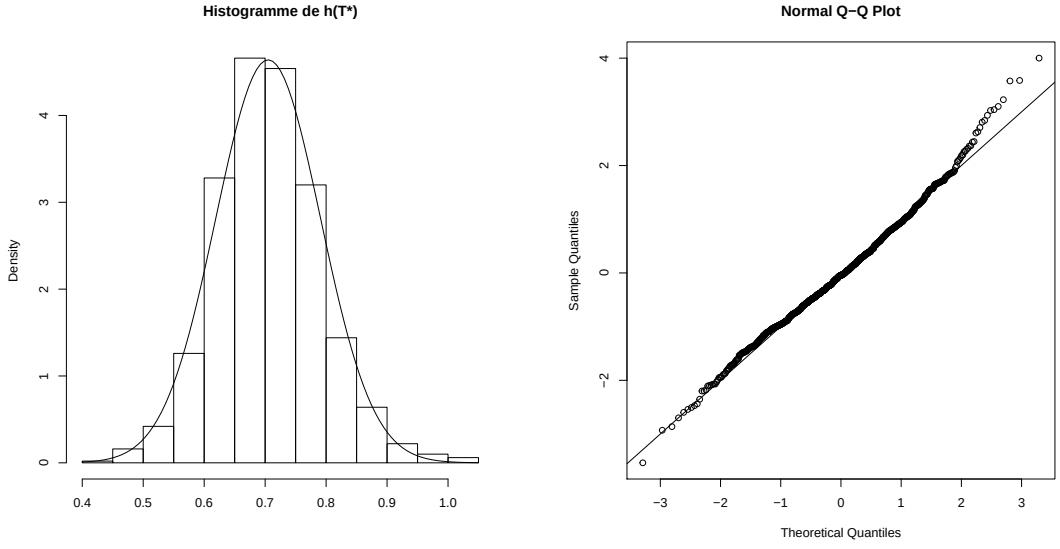
car h est strictement croissante.

Exemple. Données sur la population américaine (suite).

Revenons aux données sur la population de villes américaines. En utilisant un sous-ensemble de taille 10 des données, l'histogramme des répliques bootstrap $T_b^*, b = 1, \ldots, B = 1000$ (cf. Figure 2.4) affiche une certaine asymétrie. On peut chercher une transformation h qui rend l'histogramme plus symétrique, d'allure gaussienne. Nous analysons deux transformations possibles : la première est le simple logarithme, *i.e.* $h(t) = \log(t)$, la



a) b)



c) d)

FIGURE 2.8 – Pour un sous-échantillon de taille 10 des données sur la population de 49 villes américaines en 1920 et 1930 : a) Histogramme et b) QQ-plot de $\{\log(T_b^*), b = 1, \ldots, B\}$; c) Histogramme et d) QQ-plot de $\{(\log(T_b^*))^{2/5}, b = 1, \ldots, B\}$.

deuxième est une transformation plus sophistiquée donnée par $h(t) = (\log(t))^{2/5}$. Figure 2.8 montre l'histogramme et le QQ-plot des répliques bootstrap transformées, à savoir de $\{h(T_b^*), b = 1, \ldots, B\}$. On observe que le logarithme ne fait pas l'affaire, alors que la deuxième transformation réussi à rendre la distribution de $h(T_b^*)$ très proche d'une loi normale.

En utilisant la fonction $h(t) = (\log(t))^{2/5}$, l'intervalle boostrap pour θ donné en (2.5) qui repose sur une approximation normale vaut $[1.236, 2.045]$, alors que l'intervalle bootstrap utilisant l'intervalle bootstrap de base défini en (2.6) vaut $[1.224, 2.024]$.

2.3.5 Méthodes des percentiles

On a vu qu'une transformation du paramètre peut améliorer l'intervalle bootstrap. En revanche, en pratique il n'est pas toujours évident de trouver une telle transformation. Il existe une approche de construction d'intervalles bootstrap qui n'utiliser qu'implicitement l'existence d'une telle transformation **sans que celle-ci soit connue !** On distingue deux méthodes. Commençons par la méthode de base.

Méthode de base

Supposons qu'il existe une transformation h strictement croissante telle que la loi de $U - \eta$ où $U = h(T)$ et $\eta = h(\theta)$ est **symétrique** autour de 0. Un intervalle de confiance pour η est $[U - q_{1-\alpha/2}, U - q_{\alpha/2}]$ avec q_γ les quantiles de $U - \eta$. Par symétrie, $q_\gamma = -q_{1-\gamma}$ pour tout γ, donc l'intervalle de confiance théorique se réécrit $[U + q_{\alpha/2}, U + q_{1-\alpha/2}]$. On estime ensuite les quantiles de $U - \eta$ par bootstrap (comme pour l'intervalle de confiance bootstrap de base), ce qui nous amène à l'intervalle

$$\mathcal{I}_\eta^* = \left[u + q_{\alpha/2,B}^*, u + q_{1-\alpha/2,B}^*\right]$$

où $u = h(t)$ et $q_{\gamma,B}^*$ est le quantile bootstrap associé à $U_1^* - u, \ldots, U_B^* - u$. Remarquons que ces quantiles bootstrap sont donnés par

$$q_{\gamma,B}^* = U_{(\lceil B\gamma \rceil)}^* - u.$$

On en déduit que

$$\begin{aligned}\mathcal{I}_\eta^* &= \left[U_{(\lceil B\alpha/2 \rceil)}^*, U_{(\lceil B(1-\alpha/2) \rceil)}^*\right] \\ &= \left[h(T_{(\lceil B\alpha/2 \rceil)}^*), h(T_{(\lceil B(1-\alpha/2) \rceil)}^*)\right],\end{aligned}$$

par définition des $U_b^* = h(T_b^*)$. Ce dernier intervalle est un intervalle de confiance pour $\eta = h(\theta)$. En inversant h, un intervalle bootstrap pour le paramètre θ est donné par

$$\mathcal{I}_{\text{perc}}^* = \left[T_{(\lceil B\alpha/2 \rceil)}^*, T_{(\lceil B(1-\alpha/2) \rceil)}^*\right]. \tag{2.7}$$

Remarquons que cet intervalle est calculable **sans connaissance** de la transformation h.

Une autre interprétation de l'intervalle donné en (2.7) est de le voir comme un encadrement du paramètre θ par des quantiles de la loi de T. Mais ce raisonnement n'est valable que si la loi de l'estimateur T est symétrique.

En pratique, cet intervalle n'est pas très précis. En effet, on peut l'améliorer en incluant une correction du biais (de $U = h(T)$ comme estimateur de $\eta = h(\theta)$). De plus, l'intervalle

hérite des inconvénients de l'intervalle bootstrap de base. En particulier, assez souvent la statistique $U - \eta = h(T) - h(\theta)$ n'est pas pivotale, autrement dit, sous F elle n'a pas la même loi que sous $\hat{F}$. Par conséquent, l'estimation des quantiles par le bootstrap est erronée.

Une nette amélioration de cette approche est présentée dans le paragraphes suivant.

Méthodes des percentiles ajustées ou intervalle bootstrap BC_a

On peut apporter une amélioration à l'intervalle $\mathcal{I}^*_{\text{perc}}$ défini en (2.7) par un meilleur choix de l'ordre des quantiles bootstrap utilisés. On va construire l'**intervalle bootstrap** BC_a, où BC_a signifie *bias corrected and accelerated*, sous la forme

$$\mathcal{I}^*_{BC_a} = \left[T^*_{(\lceil B\hat{\alpha}_1 \rceil)}, T^*_{(\lceil B\hat{\alpha}_2)\rceil)}\right]. \tag{2.8}$$

avec des ordres $\hat{\alpha}_1$ et $\hat{\alpha}_2$ à déterminer.

Supposons qu'il existe une transformation $h(\cdot)$ strictement croissante, et notons $\eta = h(\theta)$ et $U = h(T)$, telle que

$$\frac{U - \eta}{\sigma_\eta} \sim \mathcal{N}(-\beta, 1),$$

avec $\beta \in \mathbb{R}$ et $\sigma_\eta = \sqrt{\mathbf{Var}(U)}$. En particulier, la quantité $-\sigma_\eta\beta$ représente l'éventuel biais de l'estimateur U comme estimateur de η. Par ailleurs, nous supposons qu'il existe une constante $a \in \mathbb{R}$ telle que σ_η vérifie la relation suivante

$$\sigma_\eta = 1 + a\eta.$$

En utilisant des quantiles de la loi normale standard, on construit un intervalle de confiance de niveau $1 - \alpha$ pour η de la façon suivante :

$$\begin{aligned}
1 - \alpha &= \mathbb{P}\left(z_{\alpha/2} \leq \frac{U - \eta}{\sigma_\eta} + \beta \leq z_{1-\alpha/2}\right) \\
&= \mathbb{P}\left((z_{\alpha/2} - \beta)(1 + a\eta) \leq U - \eta \leq (z_{1-\alpha/2} - \beta)(1 + a\eta)\right) \\
&= \mathbb{P}\left(\frac{U - (z_{1-\alpha/2} - \beta)}{1 + a(z_{1-\alpha/2} - \beta)} \leq \eta \leq \frac{U - (z_{\alpha/2} - \beta)}{1 + a(z_{\alpha/2} - \beta)}\right) \\
&= \mathbb{P}\left(U - \frac{(Ua + 1)(z_{1-\alpha/2} - \beta)}{1 + a(z_{1-\alpha/2} - \beta)} \leq \eta \leq U - \frac{(Ua + 1)(z_{\alpha/2} - \beta)}{1 + a(z_{\alpha/2} - \beta)}\right).
\end{aligned}$$

Notons

$$\tau_1 = -\frac{(Ua + 1)(z_{1-\alpha/2} - \beta)}{1 + a(z_{1-\alpha/2} - \beta)} \qquad \text{et} \qquad \tau_2 = -\frac{(Ua + 1)(z_{\alpha/2} - \beta)}{1 + a(z_{\alpha/2} - \beta)}.$$

En utilisant que pour toute variable aléatoire V de loi F_V, la variable aléatoire $F_V(V)$ suit

la loi uniforme $U[0,1]$, on a

$$\begin{aligned}
1-\alpha &= \mathbb{P}(\eta \in [U+\tau_1, U+\tau_2]) \\
&= \mathbb{P}(-\tau_2 \leq U-\eta \leq -\tau_1) \\
&= \mathbb{P}\left(-\frac{\tau_2}{\sigma_\eta}+\beta \leq \frac{U-\eta}{\sigma_\eta}+\beta \leq -\frac{\tau_1}{\sigma_\eta}+\beta\right) \\
&= \mathbb{P}\left(\Phi\left(-\frac{\tau_2}{\sigma_\eta}+\beta\right) \leq \underbrace{\Phi\left(\frac{U-\eta}{\sigma_\eta}+\beta\right)}_{\sim U[0,1]} \leq \Phi\left(-\frac{\tau_1}{\sigma_\eta}+\beta\right)\right) \quad (\text{car } (U-\eta)/\sigma_\eta+\beta \sim \mathcal{N}(0,1)) \\
&= \mathbb{P}\left(\Phi\left(-\frac{\tau_2}{\sigma_\eta}+\beta\right) \leq G(T) \leq \Phi\left(-\frac{\tau_1}{\sigma_\eta}+\beta\right)\right),
\end{aligned}$$

où G désigne la fonction de répartition de la loi de T et où on a utilisé que $G(T) \sim U[0,1]$. En inversant G on obtient

$$1-\alpha = \mathbb{P}\left(G^{-1}\left(\Phi\left(-\frac{\tau_2}{\sigma_\eta}+\beta\right)\right) \leq T \leq G^{-1}\left(\Phi\left(-\frac{\tau_1}{\sigma_\eta}+\beta\right)\right)\right).$$

Autrement dit, en notant

$$\alpha_1 = \Phi\left(-\frac{\tau_2}{\sigma_\eta}+\beta\right) \quad \text{et} \quad \alpha_2 = \Phi\left(-\frac{\tau_1}{\sigma_\eta}+\beta\right).$$

et q_γ le quantile de niveau γ de la loi de T, on a établi que

$$\mathbb{P}(q_{\alpha_1} \leqslant T \leqslant q_{\alpha_2}) \;=\; 1-\alpha\,.$$

On peut maintenant appliquer la méthode des percentiles (puisque l'hypothèse d'existence de h est bien satisfaite) mais avec les quantiles q_{α_1} et q_{α_2} plutôt que $q_{\alpha/2}$ et $q_{1-\alpha/2}$, ce qui nous amène à un intervalle de confiance de la forme

$$\left[T^*_{(\lceil \alpha_1 B \rceil)}, T^*_{(\lceil \alpha_2 B \rceil)}\right].$$

Cependant, maintenant, α_1 et α_2 sont inconnus, il faut donc les estimer. Notons que

$$\begin{aligned}
-\frac{\tau_2}{\sigma_\eta}+\beta &= \frac{(Ua+1)(z_{\alpha/2}-\beta)}{\sigma_\eta(1+a(z_{\alpha/2}-\beta))}+\beta \\
&\approx \frac{z_{\alpha/2}-\beta}{1+a(z_{\alpha/2}-\beta)}+\beta,
\end{aligned}$$

car $Ua+1 \approx \eta a+1 = \sigma_\eta$ puisque U est un estimateur de η. De même,

$$-\frac{\tau_1}{\sigma_\eta}+\beta \approx \frac{z_{1-\alpha/2}-\beta}{1+a(z_{1-\alpha/2}-\beta)}+\beta.$$

Il nous reste à trouver comment estimer les constantes β et a. En fait, la constante β vérifie

$$\Phi(\beta) = \mathbb{P}\left(\frac{U-\eta}{\sigma_\eta}+\beta \leq \beta\right) = \mathbb{P}\left(U \leq \eta\right) = \mathbb{P}\left(h(T) \leq h(\theta)\right) = \mathbb{P}(T \leq \theta),$$

car h^{-1} est monotone croissant. Cela implique qu'un estimateur bootstrap β^*_B de β est donné par

$$\beta^*_B = \Phi^{-1}\left(\frac{1}{B}\sum_{b=1}^{B} \mathbb{1}\{T^*_b < t\}\right).$$

On voit que si θ est la médiane de la loi de T, alors β_B^* sera près de 0. Dans ce sens, β_B^* effectue une correction d'une sorte de biais.

Quant à la constante d'accélération a, elle mesure le taux de changement de l'écart type de U. On peut montrer qu'elle est liée au coefficient d'asymétrie de la loi de T. Un estimateur de a est donné par

$$\hat{a} = \frac{\sum_{i=1}^n (T_{(\cdot)} - T_{(-i)})^3}{6\{\sum_{i=1}^n (T_{(\cdot)} - T_{(-i)})^2\}^{3/2}},$$

où $T_{(-i)}$ est l'estimateur T évalué sur l'échantillon $\mathbf{x}$ après suppression de la i-ème observation, à savoir $T_{(-i)} = T(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$, et $T_{(\cdot)} = \frac{1}{n}\sum_{i=1}^n T_{(-i)}$. En fait, cette technique d'évaluation d'un estimateur sur un échantillon diminué d'une observation s'appelle du *Jackknife* ou du *leave-one-out*.

Au final, l'intervalle bootstrap BC_a est donné par (2.8) avec

$$\hat{\alpha}_1 = \Phi\left(\frac{z_{\alpha/2} - \beta_B^*}{1 + \hat{a}(z_{\alpha/2} - \beta_B^*)} + \beta_B^*\right) \quad \text{et} \quad \hat{\alpha}_2 = \Phi\left(\frac{z_{1-\alpha/2} - \beta_B^*}{1 + \hat{a}(z_{1-\alpha/2} - \beta_B^*)} + \beta_B^*\right).$$

Remarquons que si $\beta_B^* = 0$ et $\hat{a} = 0$, on a $\hat{\alpha}_1 = \alpha/2$ et $\hat{\alpha}_2 = 1 - \alpha/2$, et donc l'intervalle $\mathcal{I}_{BC_a}^*$ coincïde avec l'intervalle $\mathcal{I}_{\text{perc}}^*$ en (2.7).

Il est possible de montrer de façon rigoureuse que l'intervalle BC_a est un intervalle de confiance pour θ de niveau asymptotique $1 - \alpha$ sous des conditions assez faibles.

2.3.6 Comparaison de différents intervalles bootstrap

L'approche traditionnelle de la statistique inférentielle repose sur des modèles idéalisés avec des hypothèses fortes sur le type de la distribution des observations. Le bootstrap n'a pas besoin de modèle statistique bien précis. Bien qu'il existe des méthodes bootstrap dites paramétriques qui s'appliquent aux modèles paramétriques (*i.e.* la loi des observations est fixée à un paramètre $\theta \in \mathbb{R}^d$ près), le bootstrap est, avant tout, utilisé pour des modèles nonparamétriques. De ce fait, le bootstrap s'avère très utile pour des problèmes pratiques où la définition d'un modèle statistique proche de la réalité est difficile à trouver (comme dans l'exemple sur les populations urbaines aux États-Unis) et/ou un modèle paramétrique est trop contraignant.

L'idée fondamentale des méthodes bootstrap est qu'en absence d'informations précises sur la distribution des données, l'échantillon observé contient toute information disponible sur la distribution sous-jacente, ce qui justifie la méthode de substitution et le rééchantillonnage.

Quelques conditions d'application

La condition pour que l'intervalle bootstrap $\mathcal{I}_{\text{norm}}^*$ soit un intervalle de confiance asymptotique est que l'approximation normale de la loi de T soit vérifiée asymptotiquement. Si la distribution de T n'est pas gaussienne et par exemple asymétrique, l'intervalle $\mathcal{I}_{\text{norm}}^*$ est très erratique. En fait, cet intervalle est peu utilisé dans la pratique, car l'hypothèse gaussienne est très contraignante. Par ailleurs, l'intervalle $\mathcal{I}_{\text{norm}}^*$ nécessite que la taille n de l'échantillon soit plutôt élevée.

Pour que l'intervalle bootstrap $\mathcal{I}_{\text{basic}}^*$ de base soit performant, il faut que la statistique $T - \theta$ soit un *pivot approximatif*. Autrement dit, la loi de $T - \theta$ sous F doit être identique

ou près de la loi de $T - \theta$ sous $\hat{F}$, ce qui est rarement le cas en pratique et implique en général des mauvaises performances de l'intervalle bootstrap $\mathcal{I}^*_{\text{basic}}$.

Quant à l'intervalle bootstrap studentisé $\mathcal{I}^*_{\text{student}}$, l'approche conduit à de très bons résultats si la loi de la statistique $\mathcal{Z}$ est plus ou moins la même quelque soit la valeur de θ, autrement dit si $\mathcal{Z}$ est un pivot approximatif. En revanche, si la loi de $\mathcal{Z}$ varie beaucoup avec le paramètre θ, les résultats peuvent s'avérer catastrophiques. Dans ce cas, un intervalle bootstrap basé sur une transformation h du paramètre θ telle que $h(T) - h(\theta)$ soit de loi normale ou pivotale peut améliorer les résultats significativement. La difficulté de cette démarche est de trouver la bonne transformation h. Un autre inconvénient de l'intervalle studentisé est l'utilisation d'un double bootstrap, très gourmand en termes de temps de calcul.

Concernant la méthode des percentiles de base, elle nécessite la symétrie de la loi de T. En cas d'asymétrie, il est préférable d'utiliser l'intervalle bootstrap BC_a. En effet, ce dernier intervalle bootstrap $\mathcal{I}^*_{BC_a}$ affiche de très bonnes performances en pratique.

Justesse des intervalles bootstrap

Pour tous les intervalles bootstrap présentés ici on peut déduire des conditions sous lesquelles ils sont des intervalles de confiance asymptotique. Plus précisément, on peut montrer que la couverture, c'est-à-dire la probabilité $\mathbb{P}_\theta(\theta \in \mathcal{I}_n)$, tend vers la valeur nominale $1 - \alpha$ lorsque la taille n de l'échantillon tend vers l'infini. On dit qu'un intervalle est **juste au premier ordre** si la vitesse de convergence de la couverture vers $1 - \alpha$ est de l'ordre $n^{-1/2}$. Autrement dit, si

$$\lim_{n\to\infty} \mathbb{P}_\theta(\theta \in \mathcal{I}_n) = 1 - \alpha + O(n^{-1/2}).$$

En général, les intervalles de confiance asymptotique usuels qui repose sur le théorème central limite sont juste au premier ordre. On peut également montrer que les intervalles bootstrap $\mathcal{I}^*_{\text{basic}}$ et $\mathcal{I}^*_{\text{perc}}$ sont aussi justes au premier ordre.

En revanche, sous des conditions appropriées (mais assez souples), l'intervalle bootstrap studentisé ainsi que l'intervalle BC_a sont **juste au second d'ordre**, autrement dit

$$\lim_{n\to\infty} \mathbb{P}_\theta(\theta \in \mathcal{I}_n) = 1 - \alpha + O(n^{-1}).$$

La convergence de la couverture a lieu plus rapidement que dans les autres cas. Ceci n'est pas seulement un avantage théorique, mais entraîne de meilleures couvertures sur des échantillons à taille finie. Concernant l'intervalle bootstrap studentisé, cette amélioration s'explique en quelque sorte par le fait de considérer une normalisation de la loi de l'estimateur T, c'est-à-dire de passer à une échelle normalisée par opposition à la méthode de base de l'intervalle $\mathcal{I}^*_{\text{basic}}$. Quant à l'intervalle BC_a, l'amélioration est due au choix intelligent des quantiles de la distribution bootstrap.

En fait, même si l'intervalle bootstrap studentisé est juste au second ordre, on peut souvent observer qu'il n'est pas très performant sur des petits échantillons. La raison est que l'approche demande de bootstrapper le rapport de deux variables aléatoires, et c'est seulement quand la taille d'échantillon est suffisamment grande que la variabilité du dénominateur est suffisamment petite pour ne pas jouer négativement sur la justesse de l'intervalle.

En conclusion, l'intervalle BC_a est l'approche recommandée. Il est préférable aux autres méthodes dans la grande majorité de cas.

Chapitre 3

Modèles de mélange

Dans ce chapitre, nous présenterons une classe de modèles très utilisée en statistique comme en machine learning : les modèles de mélange. Le modèle de mélange fait partie de la famille de modèles à variables latentes. Nous montrerons son utilité et son importance pour l'analyse des données, avant de présenter deux algorithmes pour ajuster les paramètres d'un modèle de mélange, l'algorithme EM et l'échantillonneur de Gibbs. Mais tout d'abord, faisons un rappel sur la notion de la loi conditionnelle, qui est un outil nécessaire pour la définition et manipulation des modèles de mélange.

Quelques références : sur les modèles de mélange en général (Droesbeke et al., 2013; McLachlan and Peel, 2000), sur l'algorithme EM (McLachlan and Krishnan, 2008) et sur l'échantillonneur de Gibbs pour les mélanges (Marin and Robert, 2007; Robert and Casella, 2004).

3.1 Rappel : Loi conditionnelle

Soient A et B deux événements aléatoires $A, B \in \mathcal{A}$ tels que $\mathbb{P}(B) > 0$. La **probabilité conditionnelle** $\mathbb{P}(A|B)$ **de** A **sachant** B est définie par

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

Si A et B sont des événements indépendants, on a

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A)\mathbb{P}(B)}{\mathbb{P}(B)} = \mathbb{P}(B).$$

De la définition des probabilités conditionnelles découle immédiatement la **formule de Bayes**

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)},$$

pourvu que $\mathbb{P}(A) > 0$ et $\mathbb{P}(B) > 0$, et la **formule des probabilités totales**

$$\mathbb{P}(A) = \sum_{i=1}^{\infty} \mathbb{P}(A|B_i)\mathbb{P}(B_i),$$

si $B_1, B_2, \ldots$ forment une partition de Ω telle que $\mathbb{P}(B_i) > 0$ pour tout i.

Soient $\mathbf{X} \in \mathbb{R}^p$ et $\mathbf{Y} \in \mathbb{R}^q$ des vecteurs aléatoires. Notons $p_{(\mathbf{X},\mathbf{Y})}$ la densité jointe du vecteur aléatoire $(\mathbf{X}, \mathbf{Y}) \in \mathbb{R}^{p+q}$ par rapport à une mesure de référence donnée $\mu_{(X,Y)} = \mu_X \otimes \mu_Y$

sur $\mathbb{R}^{p+q}$. Notons $p_{\mathbf{X}}$ et $p_{\mathbf{Y}}$ les densités marginales de $\mathbf{X}$ et de $\mathbf{Y}$ données par

$$p_{\mathbf{X}}(x) = \int_{\mathbb{R}^q} p_{(\mathbf{X},\mathbf{Y})}(x,y)\mu_Y(\mathrm{d}y) \quad \text{et} \quad p_{\mathbf{Y}}(y) = \int_{\mathbb{R}^p} p_{(\mathbf{X},\mathbf{Y})}(x,y)\mu_X(\mathrm{d}x).$$

Pour $x \in \mathbb{R}^p$ fixée, on définit la fonction $y \mapsto p_{\mathbf{Y}|\mathbf{X}}(y|x)$ comme

$$p_{\mathbf{Y}|\mathbf{X}}(y|x) = \begin{cases} \frac{p_{(\mathbf{X},\mathbf{Y})}(x,y)}{p_{\mathbf{X}}(x)} & \text{si } p_{\mathbf{X}}(x) > 0, \\ p_{\mathbf{Y}}(y) & \text{sinon.} \end{cases}$$

On remarque que $p_{\mathbf{Y}|\mathbf{X}}(\cdot|x)$ est une densité de probabilité par rapport à la mesure μ_Y pour tout $x \in \mathbb{R}^p$, car

$$p_{\mathbf{Y}|\mathbf{X}}(y|x) \geq 0, \forall y \in \mathbb{R}^q \quad \text{et} \quad \int_{\mathbb{R}^q} p_{\mathbf{Y}|\mathbf{X}}(y|x)\mu_Y(\mathrm{d}y) = 1.$$

On appelle $p_{\mathbf{Y}|\mathbf{X}}(\cdot|x)$ la **densité conditionnelle de $\mathbf{Y}$ sachant que $\mathbf{X} = x$.**

La **loi conditionnelle de Y sachant que $X = x$** est alors donnée par

$$\mathbb{P}(\mathbf{Y} \in A|\mathbf{X} = x) = \int_A p_{\mathbf{Y}|\mathbf{X}}(y|x)\mu_Y(\mathrm{d}y), \quad A \in \mathcal{B}, x \in \mathbb{R}.$$

et l'**espérance conditionnelle de $\mathbf{Y}$ sachant que $\mathbf{X} = x$** par

$$\mathbb{E}[\mathbf{Y}|\mathbf{X} = x] = \int_{\mathbb{R}^q}^{\infty} y p_{\mathbf{Y}|\mathbf{X}}(y|x)\mu_Y(\mathrm{d}y).$$

La condition $\mathbb{E}[\|\mathbf{Y}\|] < \infty$ est suffisante pour assurer l'existence de l'espérance conditionnelle $\mathbb{E}[\mathbf{Y}|\mathbf{X} = x]$ pour tout x.

On peut également définir la **fonction de répartition conditionnelle de $\mathbf{Y}$ sachant que $\mathbf{X} = x$**, notée $F_{\mathbf{Y}|\mathbf{X}}(\cdot|x)$: c'est la f.d.r. qui correspond à la mesure de probabilité $\mathbb{P}(\mathbf{Y} \in \cdot|\mathbf{X} = x)$. Elle est donnée par

$$F_{\mathbf{Y}|\mathbf{X}}(y|x) = \int_{-\infty}^{y} p_{\mathbf{Y}|\mathbf{X}}(t|x)\mu_Y(\mathrm{d}t).$$

Dans le cas discret, quand $\mathbf{X}$ et $\mathbf{Y}$ sont deux vecteurs aléatoires discrets à valeurs dans $\mathcal{V} = \{v_1, v_2, \ldots\}$ et $\mathcal{W} = \{w_1, w_2, \ldots\}$, resp., la loi conditionnelle de $\mathbf{Y}$ sachant que $\mathbf{X} = v$, pour $v \in \mathcal{V}$ fixé, est donnée par les probabilités

$$\mathbb{P}(\mathbf{Y} = w_k|\mathbf{X} = v) = \frac{\mathbb{P}(\mathbf{Y} = w_k, \mathbf{X} = v)}{\mathbb{P}(\mathbf{X} = v)}, \quad \forall k \geq 1.$$

Dans le cas continu, où $\mathbf{X}$ et $\mathbf{Y}$ sont deux vecteurs aléatoires de densité jointe $f_{(\mathbf{X},\mathbf{Y})}(x,y)$ par rapport à la mesure de Lebesgue sur $\mathbb{R}^{p+q}$, la densité conditionnelle $f_{\mathbf{Y}|\mathbf{X}}$ de $\mathbf{Y}$ sachant $\mathbf{X}$ est donnée par

$$f_{\mathbf{Y}|\mathbf{X}}(y|x) = \begin{cases} \frac{f_{(\mathbf{X},\mathbf{Y})}(x,y)}{f_{\mathbf{X}}(x)} & \text{si } f_{\mathbf{X}}(x) > 0 \\ f_{\mathbf{Y}}(y) & \text{si } f_{\mathbf{X}}(x) = 0, \end{cases}$$

où $f_{\mathbf{X}}$ et $f_{\mathbf{Y}}$ dénotent les densités marginales de $\mathbf{X}$ et $\mathbf{Y}$.

Remarquons que ces définitions sont cohérentes avec les définitions plus générales dans la littérature. En particulier, si l'on note $g(x) = \mathbb{E}[\mathbf{Y}|\mathbf{X} = x]$ (tel que défini ci-dessus), alors $g(\mathbf{X})$ est bien l'espérance conditionnelle de $\mathbf{Y}$ sachant $\mathbf{X}$ au sens où c'est l'unique variable aléatoire mesurable par rapport à $\mathbf{X}$ telle que

$$\mathbb{E}\left[f(\mathbf{X})g(\mathbf{X})\right] = \mathbb{E}\left[f(\mathbf{X})\mathbf{Y}\right]$$

TABLE 3.1 – Fréquences des longueurs des ailes en mm de 381 passereaux.

Longueur	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	98
Fréquence	5	3	12	36	55	45	21	13	15	34	59	48	16	12	6	1

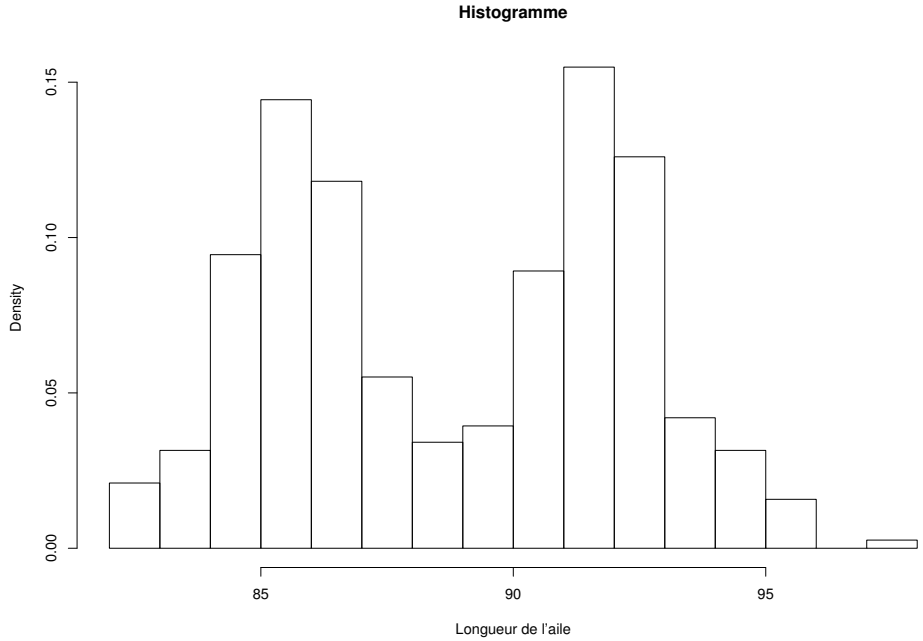


FIGURE 3.1 – Histogramme des longueurs des ailes.

pour toute fonction f mesurable bornée (l'unicité étant à comprendre au sens où deux variables aléatoires $\mathbf{X}$-mesurables qui satisfont cette propriété sont égales presque sûrement).

3.2 Modèles de mélange

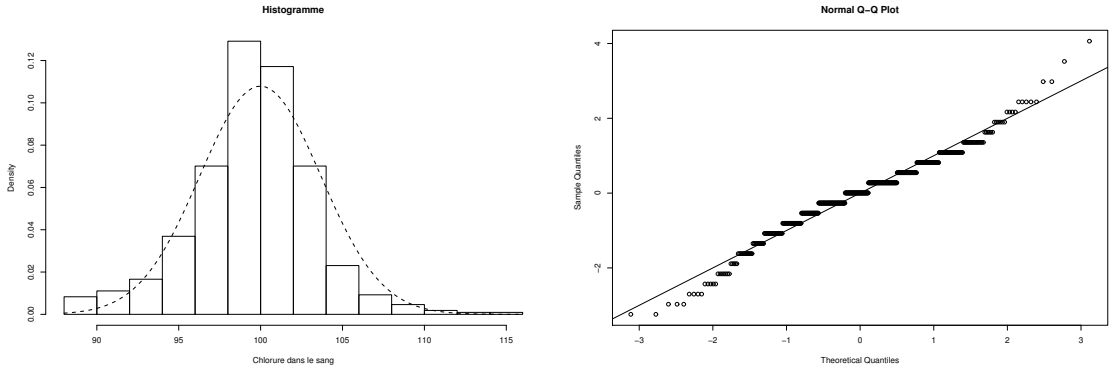
Les modèles de mélange sont très fréquemment utilisés dans la modélisation des données issues de tous les domaines d'application ainsi qu'en apprentissage statistique pour des tâches comme le clustering d'individus. Les modèles de mélange définissent des familles de lois de probabilité très riche, avec relativement peu de paramètres en combinant des lois usuelles. Ils permettent de modéliser et estimer une structure ou organisation sous-jacente des observations sous forme de plusieurs groupes ou populations avec des comportements variables.

3.2.1 Exemple : Longueurs des ailes de passereaux

Les données suivantes proviennent d'une étude sur la migration des passereaux. Pour ne pas trop perturber les oiseaux capturés brièvement, seulement quelques mesures rapides sont effectuées. La Table 3.1 reporte les mesures de la longueur d'une aile (en mm) de 381 passereaux et la Figure 3.1 représente les données graphiquement. La forme bimodale de l'histogramme laisse penser à la présence de deux populations différentes dans l'échantillon. En effet, il y a une petite différence en la taille des mâles et femelles. En revanche, lors

TABLE 3.2 – Taux de chlorure dans le sang (mmol/L) pour 542 individus.

Taux	88	89	90	91	92	93	94	95	96	97	98	99	100
Fréquence	2	3	4	5	7	5	13	13	27	36	40	72	68
Taux	101	102	103	104	105	106	107	108	109	111	113	115	
Fréquence	80	47	43	33	19	6	6	4	5	2	1	1	



(a) (b)

FIGURE 3.2 – (a) Histogramme des données de chlorure dans le sang et la densité de la loi normale $\mathcal{N}(\hat{\mu}, \hat{\sigma}^2)$ où $\hat{\mu}$ et $\hat{\sigma}^2$ sont l'EMV. (b) QQ-plot des données standardisées en comparaison à la loi normale standard.

de cette étude, le sexe des oiseaux n'a pas été observé. Par conséquent, nous disposons alors d'un seul jeu de données qui mélange les observations pour les mâles avec celles des femelles.

Pour modéliser une telle situation où on observe deux populations de comportement différent, il convient d'associer une loi de probabilité à chacune des populations. Autrement dit, on introduit une loi $\mathbb{P}_F$ pour les longueurs d'aile des femelles et une loi $\mathbb{P}_M$ pour les mâles. Vu la forme bimodale de l'histogramme, on pourrait choisir des lois normales pour $\mathbb{P}_F$ et $\mathbb{P}_M$ (de paramètres différents).

Si on avait noté le sexe de chaque oiseau, on pourrait faire deux sous-échantillons pour estimer les paramètres des lois $\mathbb{P}_F$ et $\mathbb{P}_M$ séparément. Malheureusement, ce n'est pas le cas. Par conséquent, il faut intégrer dans le modèle le fait que l'on ignore si la i-ème observation x_i est la longueur d'aile d'une femelle ou d'un mâle.

3.2.2 EXEMPLE : TAUX DE CHLORURE DANS LE SANG

Une autre étude porte sur le taux de chlorure dans le sang chez des adultes. Nous disposons d'un échantillon de taille 542 (cf. Table 3.2). La Figure 3.2 (a) montre l'histogramme qui est de forme unimodale et relativement symétrique. La densité d'une loi normale superposée à l'histogramme est en bonne adéquation, sans être parfaite. En revanche, le QQ-plot (Figure 3.2 (b)) qui compare les données (standardisées) à la loi normale standard est nettement moins favorable à l'hypothèse d'une simple loi normale.

En fait, ici, il est plus réaliste de supposer que la population observée contient une grande proportion π d'individus en bonne santé et une petite proportion $(1 - \pi)$ d'individus malades avec des taux de chlorure extrêmes (soit très bas, soit trop élevé). On peut faire l'hypothèse que les individus sains suivent une loi normale $\mathcal{N}(\mu_1, \sigma_1^2)$ alors que les individus malades suivent une loi normale $\mathcal{N}(\mu_2, \sigma_2^2)$ avec à peu près la même moyenne ($\mu_1 \approx$

μ_2) mais avec une plus grande variance que les personnes en bonne santé ($\sigma_1^2 < \sigma_2^2$). Nous soulignons que nous ignorons qui sont les personnes malades dans cette étude. Il est alors nécessaire d'utiliser un modèle qui modélise à la fois le comportement des personnes malades comme des personnes en bonne santé.

3.2.3 Définition d'un modèle de mélange

Les deux exemples précédents appartiennent à la famille des modèles de mélange, qui est la modélisation simultanée du comportement de plusieurs populations différentes. La définition d'une population, classe ou groupe dépend de l'application. Elle peut reposer sur par exemple les tranches d'âge, les milieux sociaux, les nationalités ou des antécédents médicaux pour ne parler que de critères sur des populations humaines. Mais à vrai dire, il n'est pas nécessaire de définir ces groupes (ou de savoir les interpréter), il suffit de supposer qu'ils existent.

Soit $m \geq 2$ le nombre de sous-populations différentes dont nous cherchons à modéliser le comportement. Notons $\mathbb{P}_j$ la loi associée à la j-ème classe. Pour simplifier les notations, nous supposons que toutes les lois $\mathbb{P}_j$ appartiennent à une même famille $\mathcal{H} = \{h_\phi, \phi \in \Phi\}$ de lois connues où $\Phi \subset \mathbb{R}^d$ et h_ϕ désignent des densités par rapport à une mesure de références μ. On notera $\phi_j \in \Phi$ le paramètre de la loi $\mathbb{P}_j$, autrement dit, $\mathbb{P}_j$ est la loi de densité h_{ϕ_j}. En plus, notons π_j la proportion d'individus de la j-ème classe dans la population totale, de sorte que $\pi_j \in [0,1]$ pour tout $j = 1, \ldots, m$ et $\sum_{j=1}^m \pi_j = 1$.

Modélisation

Pour définir une variable aléatoire X qui représente m populations différentes, il convient d'introduire d'abord une variable aléatoire U pour modéliser l'appartenance d'un individu à une des m populations. Plus précisément, la loi de U est discrète à valeurs dans $\{1, \ldots, m\}$ avec

$$\mathbb{P}(U = j) = \pi_j, \quad j = 1, \ldots, m.$$

Ce qui est équivalent à la notation

$$U \sim \sum_{j=1}^m \pi_j \delta_{\{j\}},$$

où $\delta_{\{j\}}$ désigne la mesure de Dirac avec masse 1 en $\{j\}$. Par ailleurs, on introduit des variables aléatoires $V_j, j = 1, \ldots, m$ où V_j est de densité h_{ϕ_j} avec $\phi_j \in \Phi$. On suppose que les variables aléatoires $U, V_1, \ldots, V_m$ sont mutuellement indépendantes. Enfin, on définit la variable aléatoire X par

$$X = \sum_{j=1}^m \mathbb{1}_{\{U=j\}} V_j. \tag{3.1}$$

Dans les deux exemples précédents, on peut considérer les données $\mathbf{x} = (x_1, \ldots, x_n)$ comme des réalisations i.i.d. d'une telle variable aléatoire X avec $m = 2$ classes.

Loi de mélange

Calculons la loi de X. Par la formule des probabilités totales et l'indépendance des variables V_j et U, on a

$$\begin{aligned} F_X(x) = \mathbb{P}(X \leq x) &= \sum_{k=1}^{m} \mathbb{P}(X \leq x | U = k)\mathbb{P}(U = k) \\ &= \sum_{k=1}^{m} \pi_k \mathbb{P}\left(\left(\sum_{j=1}^{m} \mathbb{1}\{U = j\} V_j \right) \leq x \middle| U = k \right) \\ &= \sum_{k=1}^{m} \pi_k \mathbb{P}(V_k \leq x | U = k) = \sum_{k=1}^{m} \pi_k \mathbb{P}(V_k \leq x) = \sum_{k=1}^{m} \pi_k F_{V_k}(x). \end{aligned}$$

Comme les lois des V_j admettent des densités par rapport à une mesure μ, on en déduit que la loi de X admet également une densité p_X par rapport à μ. En dérivant F_X on obtient pour la densité

$$p_X(x) = \sum_{j=1}^{m} \pi_j h_{\phi_j}(x). \tag{3.2}$$

La densité p_X est dite **densité de mélange**. On appelle h_{ϕ_j} la **j-ième composante du mélange** et π_j son **poids**. Notons que, si $\mathcal{H} = \{h_\phi, \phi \in \Phi\}$ est une famille de lois continues (ou discrètes), p_θ est également continue (ou discrète). La densité de mélange p_X est une combinaison convexe des densités $h_{\phi_j}, j = 1, \ldots, m$.

Les paramètres du modèle de mélange sont, d'une part, les paramètres $\phi_1, \ldots, \phi_m \in \Phi$ des différentes composantes du mélange, et d'autre part, les probabilités discrètes $\pi_1, \ldots, \pi_m$ de la loi de U. Puisque $\sum_{j=1}^{m} \pi_j = 1$, la valeur de π_m est déterminée par les valeurs de $\pi_1, \ldots, \pi_{m-1}$. Il en résulte que le vecteur de paramètres θ d'un modèle de mélange comme décrit ci-dessus est donné par

$$\theta = (\phi_1, \ldots, \phi_m, \pi_1, \ldots, \pi_{m-1}).$$

Si les ϕ_j sont des nombres réels, le modèle de mélange contient alors $2m - 1$ paramètres inconnus.

Le nombre m de populations ou classes est appelé **ordre du mélange**. Dans ce cours, on supposera que l'ordre m du modèle de mélange est connu. Cependant, ce n'est que rarement le cas en pratique. Il existe différentes méthodes pour sélectionner l'ordre m d'un mélange au vu des données, mais elles dépassent le cadre de ce cours. Remarquons simplement qu'en augmentant l'ordre m on ne peut qu'enrichir le modèle et donc améliorer l'adéquation aux données puisqu'un modèle d'ordre m peut toujours être vu comme un modèle d'ordre $m+1$ où le poids d'un des groupes est nul, ou alors où un groupe est divisé en deux groupes avec des paramètres ϕ_j identiques. Néanmoins, utiliser un modèle trop riche conduit en général au problème du surapprentissages. Cela signifie que le modèle explique très bien les données sur lesquelles on ajuste les paramètres, mais il est défaillant sur des nouvelles observations. Il convient d'opérer un compromis entre un modèle suffisamment expressif et un modèle avec un nombre de paramètres limité.

On appelle les U_i des **variables latentes** ou **manquantes** ou **cachées**, car elles ne sont pas observées. L'appartenance de groupe d'une observation X_i, qui correspond à la valeur de la variable U_i, est l**étiquette** ou le **label** de X_i. Un modèle de mélange est adéquat lorsqu'on ne dispose pas de ces labels, comme dans l'exemple des passereaux où le sexe des oiseaux n'est pas observé. Ce manque d'information se produit pour des raisons variées. Cela est

parfois dû à un oubli pendant l'acquisition des données. Parfois il est impossible (ou très cher) d'obtenir cette information. Mais le plus souvent, on ignore quelle variable ou quel critère définit bien les groupes qui explique la loi des observations. En général, on ne fait que l'hypothèse qu'un tel partitionnement existe et on cherche à l'identifier et ensuite à interpréter les groupes obtenus.

En effet, si les variables latentes étaient observées, on ne s'embêterait pas à faire un modèle de mélange, mais on utiliserait simplement un modèle par population. Pour les passeraux, on aurait un modèle gaussien pour les femmelles et un autre pour les mâles.

Loi mélangeante ou loi latente

Définissions la variable aléatoire W à valeurs dans $\{\phi_1, \dots, \phi_m\}$ avec

$$\mathbb{P}(W = \phi_j) = \pi_j, \qquad j = 1, \dots, m.$$

Notons $\mathbb{Q}$ la loi de W donnée par

$$\mathbb{Q} = \sum_{j=1}^{m} \pi_j \delta_{\{\phi_j\}}.$$

La loi $\mathbb{Q}$ détermine la loi de la variable X définie par le modèle de mélange associé. On appelle $\mathbb{Q}$ la **loi mélangeante** ou la **loi latente** du mélange.

On peut réécrire la densité de mélange p_X donné en (3.2) comme

$$p_X(x) = \int_{\phi \in \Phi} h_\phi(x) \mathrm{d}\mathbb{Q}(\phi). \tag{3.3}$$

Quand $\mathbb{Q}$ est une loi discrète comme ci-dessus, p_X donné en (3.3) est la densité d'un **mélange fini** avec m composantes. En revanche, on constate que l'intégrale en (3.3) fait aussi sens lorsque $\mathbb{Q}$ est une loi continue sur Φ : on parle de **mélange continu**, ce qui correspond à un mélange avec une infinité de composantes.

3.2.4 Simulation d'un mélange

La simulation de réalisations d'un mélange se fait assez naturellement en deux étapes en utilisant la variable latente.

Pour simuler une réalisation X d'un mélange fini, on utilise la construction de X par des variables aléatoires $U, V_1, \dots, V_m$ donnée par (3.1) :

1. Tirer l'étiquette $U \sim \sum_{j=1}^{m} \pi_j \delta_{\{j\}}$, c'est-à-dire générer une réalisation de la loi discrète à valeurs dans $\{1, \dots, m\}$ avec probabilités $\mathbb{P}(U = j) = \pi_j, j = 1, \dots, m$.
2. Tirer des réalisations $V_j \sim h_{\phi_j}, j = 1, \dots, m$ indépendantes.
3. Évaluer $X = \sum_{j=1}^{m} \mathbb{1}_{\{U=j\}} V_j$.

En effet, on peut simplifier cet algorithme en ne simulant que la composante V_j dont on a vraiment besoin :

1. Tirer l'étiquette $U \sim \sum_{j=1}^{m} \pi_j \delta_{\{j\}}$.
2. Tirer une réalisation de la U-ième composante : $V_U \sim h_{\phi_U}$.
3. Poser $X = V_U$.

TABLE 3.3 – Composition des 12 mélanges i.

(a)	Densité normale standard	$\mathcal{N}(0,1)$
(b)	Densité unimodale asymétrique	$\frac{1}{5}\mathcal{N}(0,1) + \frac{1}{5}\mathcal{N}(\frac{1}{2},(\frac{2}{3})^2) + \frac{3}{5}\mathcal{N}(\frac{13}{15},(\frac{5}{9})^2)$
(c)	Densité fortement asymétrique	$\sum_{k=0}^{7} \frac{1}{8}\mathcal{N}(3((\frac{2}{3})^k - 1),(\frac{2}{3})^{2k})$
(d)	Densité unimodale leptocurtique	$\frac{2}{3}\mathcal{N}(0,1) + \frac{1}{3}\mathcal{N}(0,(\frac{1}{10})^2)$
(e)	Densité avec outlier	$\frac{1}{10}\mathcal{N}(0,1) + \frac{9}{10}\mathcal{N}(0,(\frac{1}{10})^2)$
(f)	Densité bimodale	$\frac{1}{2}\mathcal{N}(-1,(\frac{2}{3})^2) + \frac{1}{2}\mathcal{N}(1,(\frac{2}{3})^2)$
(g)	Densité bimodale séparée	$\frac{3}{4}\mathcal{N}(-\frac{3}{2},(\frac{1}{2})^2) + \frac{1}{4}\mathcal{N}(\frac{3}{2},(\frac{1}{2})^2)$
(h)	Densité bimodale asymétrique	$\frac{3}{4}\mathcal{N}(0,1) + \frac{1}{4}\mathcal{N}(\frac{3}{2},(\frac{1}{3})^2)$
(i)	Densité trimodale	$\frac{9}{20}\mathcal{N}(-\frac{6}{5},(\frac{3}{5})^2) + \frac{9}{20}\mathcal{N}(\frac{6}{5},(\frac{3}{5})^2) + \frac{1}{10}\mathcal{N}(0,(\frac{1}{4})^2)$
(j)	Griffe	$\frac{49}{100}\mathcal{N}(-1,(\frac{2}{3})^2) + \frac{49}{100}\mathcal{N}(1,(\frac{2}{3})^2) + \sum_{k=0}^{6} \frac{1}{350}\mathcal{N}((k-3)/2,(\frac{1}{100})$
(k)	Griffe asymétrique	$\frac{1}{2}\mathcal{N}(0,1) + \sum_{k=-2}^{2} \frac{2^{1-k}}{31}\mathcal{N}(k+\frac{1}{2},(2^{-k}/10)^2)$
(l)	Peigne	$\sum_{k=0}^{5} \frac{2^{5-k}}{63}\mathcal{N}((65 - 96(\frac{1}{2})^k)/21,(\frac{32}{63})^2/2^{2k})$

Alternativement, on peut utiliser la loi latente $\mathbb{Q}$ dans la simulation d'un mélange. Ceci revient à remplacer le tirage de la variable latente U par la génération d'une réalisation de la variable aléatoire W. Autrement dit, on tire directement le paramètre ϕ :

1. Tirer le paramètre $W \sim \sum_{j=1}^{m} \pi_j \delta_{\{\phi_j\}}$.
2. Tirer une réalisation V de la loi h_W.
3. Poser $X = V$.

L'intérêt du dernier algorithme est qu'il s'applique également à la simulation d'un mélange **continu**.

3.2.5 Nouvelles classes de lois de probabilité

On peut constater que les modèles de mélange définissent des nouvelles classes de lois de probabilité qui sont très grandes et parfois très intéressantes en elles-mêmes. En d'autres termes, on peut utiliser un modèle de mélange même si le phénomène observé ne permet pas de parler de groupes ou sous-populations. Dans ce cas, un modèle de mélange est juste utilisé pour approcher la loi du phénomène observé, quand les familles de lois classiques ne sont pas appropriées.

Pour illustrer la diversité des mélanges de lois normales, nous en donnons quelques exemples : Les mélanges présentés dans la Table 3.3 sont représentés graphiquement dans les Figures 3.3 et 3.4. Le nombre de classes varie entre 2 et 9. Ces figures nous donnent une idée de la grande variété de lois de probabilité que l'on peut obtenir par des mélanges gaussiens.

En fait, il est possible d'approcher toute densité continue à l'aide d'un mélange gaussien, au sens de la norme L_1 ou uniformément sur tout compact.

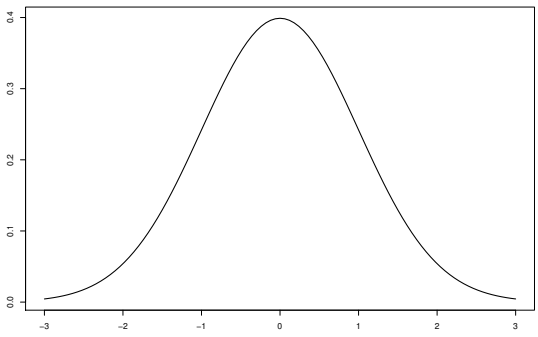
Théorème 6. *Soit g une densité continue.*

(i) Pour tout $\varepsilon > 0$ il existe un mélange gaussien fini de densité $\bar{g}$ donnée par

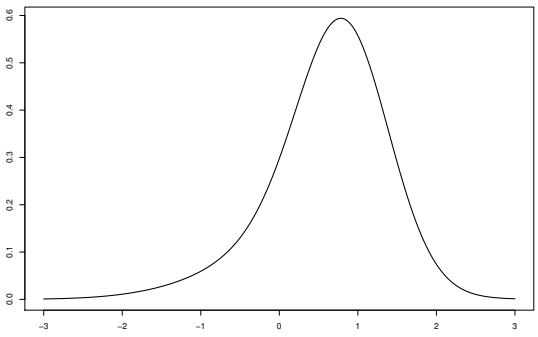
$$\bar{g}(x) = \sum_{j=1}^{m} \pi_j f_{\mathcal{N}(\mu_j,\sigma_j^2)}(x),$$

avec $\mu_j \in \mathbb{R}$, $\sigma_j > 0$, $\pi_j > 0$ tel que $\sum_{j=1}^{m} \pi_j = 1$, tel que

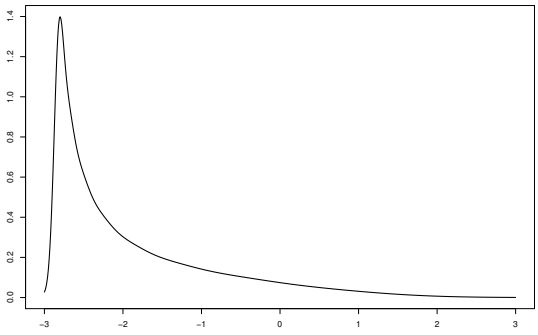
$$\|g - \bar{g}\|_1 := \int_{\mathbb{R}} |g(x) - \bar{g}(x)| \mathrm{d}x < \varepsilon.$$



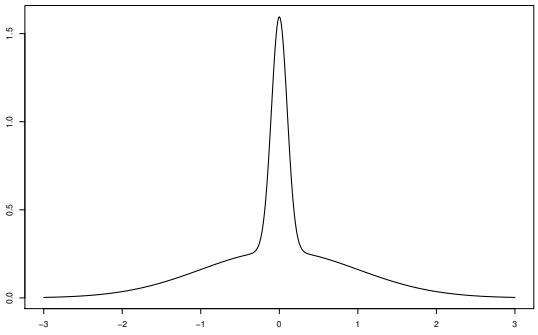
(a) Densité normale standard



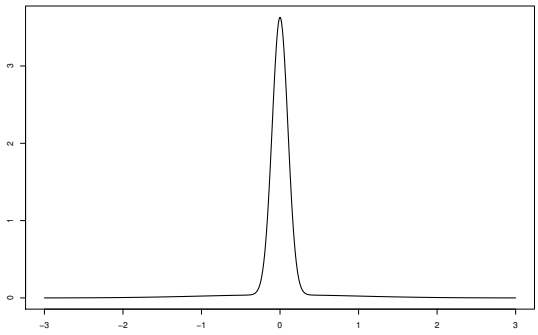
(b) Densité unimodale asymétrique



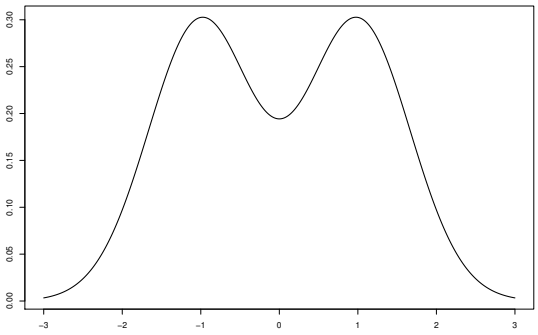
(c) Densité fortement asymétrique



(d) Densité unimodale leptocurtique

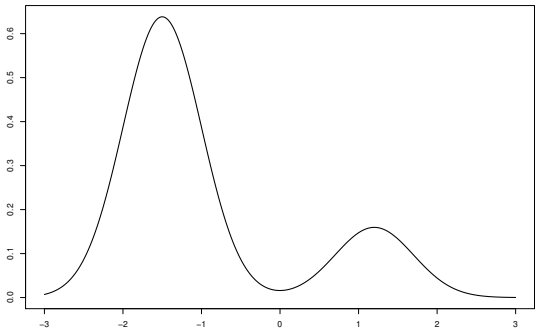


(e) Densité avec outlier

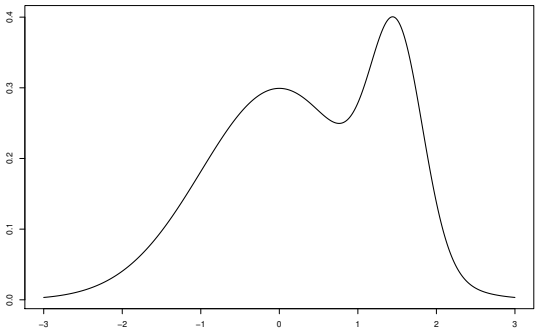


(f) Densité bimodale

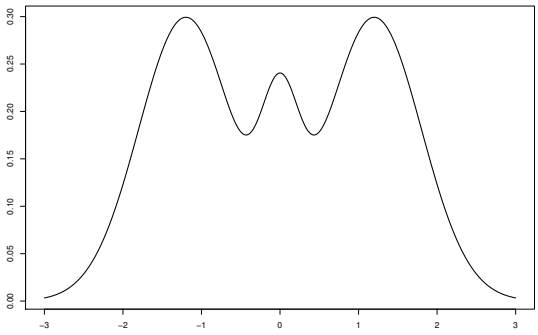
FIGURE 3.3 – Graphes de densités de mélanges gaussiens.



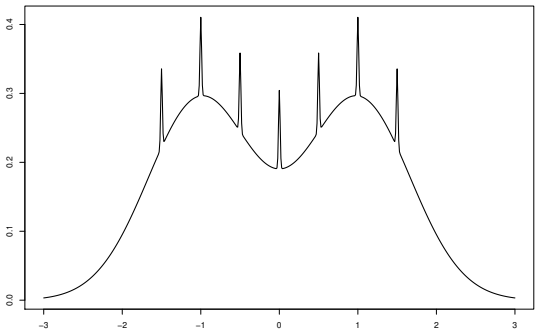
(g) Densité bimodale séparée



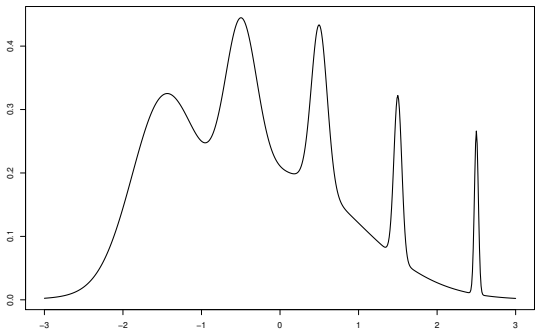
(h) Densité bimodale asymétrique



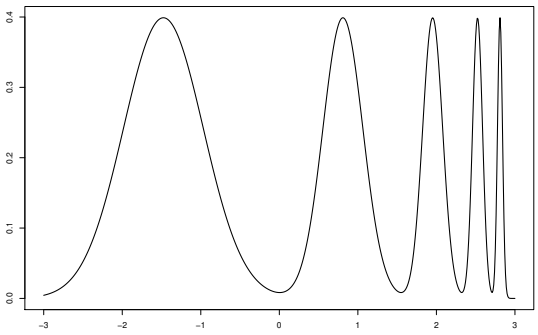
(i) Densité trimodale



(j) Griffe



(k) Griffe asymétrique



(l) Peigne

FIGURE 3.4 – Graphes de densités de mélanges gaussiens (suite).

(ii) De même, pour tout $\varepsilon > 0$ et tout compact $\mathcal{K}$, il existe un mélange gaussien fini de densité $\bar{g}$ tel que

$$\sup_{x\in\mathcal{K}} |g(x) - \bar{g}(x)| < \varepsilon\,.$$

Démonstration. Supposons qu'on a démontré le point (ii) et démontrons (i). Soit $\varepsilon > 0$ fixé. Puisque g est intégrable, il existe un compact $\mathcal{K}$ tel que

$$\int_{\mathbb{R}\setminus\mathcal{K}} g(x)\mathrm{d}x < \frac{\varepsilon}{4}.$$

D'après (ii), il existe un mélange gaussien $\bar{g}$ tel que

$$\sup_{x\in\mathcal{K}} |g(x) - \bar{g}(x)| < \frac{\varepsilon}{4|\mathcal{K}|},$$

où $|\mathcal{K}|$ désigne la mesure de Lebesgue de $\mathcal{K}$. Donc, $\int_{\mathcal{K}} |g(x) - \bar{g}(x)|\mathrm{d}x < \frac{\varepsilon}{4}$. Or,

$$\|g - \bar{g}\|_1 = \int_{\mathbb{R}} |g(x) - \bar{g}(x)|\mathrm{d}x = \int_{\mathbb{R}\setminus\mathcal{K}} |g(x) - \bar{g}(x)|\mathrm{d}x + \int_{\mathcal{K}} |g(x) - \bar{g}(x)|\mathrm{d}x.$$

On constate que

$$\begin{aligned}
\int_{\mathbb{R}\setminus\mathcal{K}} |g(x) - \bar{g}(x)|\mathrm{d}x &\leq \int_{\mathbb{R}\setminus\mathcal{K}} g(x)\mathrm{d}x + \int_{\mathbb{R}\setminus\mathcal{K}} \bar{g}(x)\mathrm{d}x \\
&= \int_{\mathbb{R}\setminus\mathcal{K}} g(x)\mathrm{d}x + 1 - \left[\int_{\mathcal{K}} (\bar{g}(x) - g(x))\mathrm{d}x + \int_{\mathcal{K}} g(x)\mathrm{d}x\right] \\
&= 2\int_{\mathbb{R}\setminus\mathcal{K}} g(x)\mathrm{d}x - \int_{\mathcal{K}} (\bar{g}(x) - g(x))\mathrm{d}x \\
&\leq 2\int_{\mathbb{R}\setminus\mathcal{K}} g(x)\mathrm{d}x + \int_{\mathcal{K}} |\bar{g}(x) - g(x)|\mathrm{d}x \\
&< 2\frac{\varepsilon}{4} + \frac{\varepsilon}{4} = \frac{3}{4}\varepsilon,
\end{aligned}$$

ce qui termine la preuve de (i).

Pour montrer (ii), soient $\varepsilon > 0$ et $\mathcal{K}$ un compact fixé. Par densité par rapport à la norme uniforme sur $\mathcal{K}$, on peut supposer que g est lipschitzienne et bornée (sinon on remplace g par une densité lipschitzienne et bornée $\tilde{g}$ telle que la norme uniforme sur $\mathcal{K}$ de $g - \tilde{g}$ est plus petite que $\varepsilon/2$ et on cherche à approcher $\tilde{g}$ par un mélange gaussien avec une précision $\varepsilon/2$, ce qui concluera). Considérons la convolution g_σ de g par un noyau gaussien $f_{\mathcal{N}(0,\sigma^2)}$ de variance $\sigma^2 > 0$:

$$g_\sigma(x) := g \star f_{\mathcal{N}(0,\sigma^2)} = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} g(x + \sigma y) e^{-\frac{y^2}{2}}\mathrm{d}y = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{\mathbb{R}} g(y) e^{-\frac{(x-y)^2}{2\sigma^2}}\mathrm{d}y$$

(qui est bien une densité de probabilité, à savoir celle de la loi de $X + \sigma Y$ avec X de loi g et $Y \sim \mathcal{N}(0,1)$). Remarquons que g_σ est un mélange (infini) de lois gaussiennes $\{\mathcal{N}(m,\sigma^2), m \in \mathbb{R}\}$ (dont la densité de mélange est g). Il faudra donc dans un deuxième temps l'approcher par un mélange fini. Mais montrons déjà que, pour $\sigma \to 0$, g_σ converge uniformément vers g. En notant L la constante de Lipschitz de g,

$$|g(x) - g_\sigma(x)| \leqslant \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} |g(x) - g(x + \sigma y)| e^{-\frac{y^2}{2}}\mathrm{d}y \leqslant L\sigma \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} |y| e^{-\frac{y^2}{2}}\mathrm{d}y \leqslant L\sigma\,.$$

Pour $\sigma = \varepsilon/(2L)$, on a donc $\sup_{x\in\mathcal{K}} |g(x) - g_\sigma(x)| \leqslant \varepsilon/2$. Il reste à approcher g_σ par une densité de mélange gaussien fini $\bar{g}$, uniformément sur $\mathcal{K}$, à un niveau de précision $\varepsilon/2$, puisqu'on aura alors

$$\sup_{x\in\mathcal{K}} |g(x) - \bar{g}(x)| \;\leqslant\; \sup_{x\in\mathcal{K}} |g(x) - g_\sigma(x)| + \sup_{x\in\mathcal{K}} |\bar{g}(x) - g_\sigma(x)| \;\leqslant\; \frac{1}{2}\varepsilon + \frac{1}{2}\varepsilon = \varepsilon\,.$$

Pour $M > 0$ et $n \in \mathbb{N}_*$ à préciser plus tard, on considère une subdivision de $[-M, M]$ en n intervalles de la forme $[x_i, x_{i+1}]$ avec $x_i = -M + 2Mi/n$ pour $i = 0, \ldots, n$. On ajoute $x_0 = -\infty$, $x_{n+1} = +\infty$ et on pose $y_i = (x_i + x_{i+1})/2$ pour $i = 0, \ldots, n$. Ainsi, pour $i = 1, .., n$,

$$\pi_i = \int_{y_{i-1}}^{y_i} g(y)\mathrm{d}y \qquad \Rightarrow \qquad \sum_{i=1}^{n} \pi_i = \int_{-\infty}^{+\infty} g(y)\mathrm{d}y = 1\,.$$

On pose finalement

$$\bar{g}(x) = \sum_{i=1}^{n} \frac{\pi_i}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-x_i)^2}{2\sigma^2}}\,.$$

Pour tout $x \in \mathcal{K}$,

$$|g_\sigma(x) - \bar{g}(x)| \;\leqslant\; \frac{1}{\sqrt{2\pi\sigma^2}} \sum_{i=1}^{n} \int_{y_{i-1}}^{y_i} \left| e^{-\frac{(x-x_i)^2}{2\sigma^2}} - e^{-\frac{(x-y)^2}{2\sigma^2}} \right| g(y)\mathrm{d}y$$

On commence par choisir M suffisamment grand pour que les deux intervalles extrêmes ($i = 1$ et n) soit de masse suffisamment petite, autrement dit on choisit M tel que

$$\int_{|x|\geqslant M-1} g(y)\mathrm{d}(y) \;<\; \frac{\varepsilon}{8}\sqrt{2\pi\sigma^2}\,.$$

On choisit aussi M assez grand pour que $\mathcal{K} \subset [-M, M]$. En utilisant que $(-\infty, y_1] \cup [y_n, +\infty) \subset \{x \in \mathbb{R},\ |x| \geqslant M - 1\}$ pour tout choix de $n \geqslant M$ (on suppose donc $n \geqslant M$ dans la suite), on obtient

$$\frac{1}{\sqrt{2\pi\sigma^2}} \sum_{i\in\{1,n\}} \int_{y_{i-1}}^{y_i} \left| e^{-\frac{(x-x_i)^2}{2\sigma^2}} - e^{-\frac{(x-y)^2}{2\sigma^2}} \right| g(y)\mathrm{d}y \;\leqslant\; \frac{2}{\sqrt{2\pi\sigma^2}}(\pi_1 + \pi_n) \;\leqslant\; \frac{\varepsilon}{4}\,.$$

Pour les intervalles finis ($i \in \{2, \ldots, n-1\}$), $|y - x_i| \leqslant M/n$ pour tout $y \in [y_{i-1}, y_i]$. En utilisant que la fonction $s \mapsto e^{-s}$ est 1-Lipschitz sur $\mathbb{R}_+$ et que la fonction $s \mapsto s^2$ est $4M$-Lipschitz sur $[-2M, 2M]$, et le fait que $x - y \in [-2M, 2M]$ si $x \in \mathcal{K} \subset [-M, M]$ et $y \in [-M, M]$, on obtient

$$\left| e^{-\frac{(x-x_i)^2}{2\sigma^2}} - e^{-\frac{(x-y)^2}{2\sigma^2}} \right| \;\leqslant\; \frac{4M}{2\sigma^2}|x_i - y| \;\leqslant\; \frac{2M^2}{n\sigma^2}$$

pour tout $x \in \mathcal{K}$, $i \in \{2, \ldots, n\}$ et $y \in [y_{i-1}, y_i]$. On choisit donc n assez grand pour avoir

$$\frac{2M^2}{n\sigma^2\sqrt{2\pi\sigma^2}} \;\leqslant\; \frac{\varepsilon}{4}\,,$$

ce choix impliquant que pour tout $x \in \mathcal{K}$,

$$\frac{1}{\sqrt{2\pi\sigma^2}} \sum_{i=2}^{n-1} \int_{y_{i-1}}^{y_i} \left| e^{-\frac{(x-x_i)^2}{2\sigma^2}} - e^{-\frac{(x-y)^2}{2\sigma^2}} \right| g(y)\mathrm{d}y \;\leqslant\; \frac{\varepsilon}{4} \sum_{i=2}^{n-1} \int_{y_{i-1}}^{y_i} g(y)\mathrm{d}y \;\leqslant\; \frac{\varepsilon}{4}\,,$$

et donc finalement $\sup_{x\in\mathcal{K}} |\bar{g}(x) - g_\sigma(x)| \leqslant \varepsilon/2$, ce qui conclut la preuve. □

3.2.6 Identifiabilité

En général, on dit qu'un modèle statistique $\{P_\theta, \theta \in \Theta\}$ est **identifiable** si et seulement si

$$\forall \theta, \theta' \in \Theta, \quad \mathbb{P}_\theta = \mathbb{P}_{\theta'} \Longrightarrow \theta = \theta'.$$

Sans identifiabilité du modèle il est impossible de définir (et donc estimer) de manière unique une "vraie" valeur du paramètre θ à partir des observations de la loi $\mathbb{P}_\theta$.

Le modèle de mélange ainsi défini n'est pas identifiable. Il est clair que tout mélange de m classes peut être représenté par un mélange de $m+1$ classes, soit en ajoutant une population avec poids $\pi_j = 0$, soit en coupant une sous-population en deux avec le même comportement, c'est-à-dire avec le même paramètre ϕ_j.

Voyons un petit exemple : le mélange suivant de deux gaussiennes

$$0.3 f_{\mathcal{N}(0,1)} + 0.7 f_{\mathcal{N}(1,2)} \tag{3.4}$$

peut s'écrire comme un mélange de trois gaussiennes avec deux composantes identiques :

$$0.1 f_{\mathcal{N}(0,1)} + 0.2 f_{\mathcal{N}(0,1)} + 0.7 f_{\mathcal{N}(1,2)}$$

ou encore comme un mélange de trois composantes différentes dont une de poids 0 :

$$0.3 f_{\mathcal{N}(0,1)} + 0.7 f_{\mathcal{N}(1,2)} + 0 f_{\mathcal{N}(100,100)}.$$

Pour obtenir un modèle de mélange avec exactement m populations différentes, il faut ajouter les contraintes suivantes sur les paramètres : $\pi_j > 0$ pour tout j, et $\phi_j \neq \phi_{j'}$ pour tout $j \neq j'$.

Cependant, même sous ces contraintes, le modèle n'est toujours pas identifiable, car il est possible de permuter les indices. Plus précisément, on peut aussi bien associer le couple de paramètres (ϕ_1, π_1) au comportement des femmelles que les paramètres (ϕ_2, π_2), car il n'y a pas de règle pour définir quelle population est la premier, deuxième... composante du mélange.

Par exemple, le mélange gaussien en (3.4) est le même que le suivant où on a permuté les composantes :

$$0.7 f_{\mathcal{N}(1,2)} + 0.3 f_{\mathcal{N}(0,1)}.$$

Ce problème d'identifiabilité du modèle, où la permutation des composantes ne change pas la loi de mélange, est connu comme le problème du **label switching**.

Si $\Phi \subset \mathbb{R}$, on peut obtenir l'identifiabilité du modèle de mélange par les deux contraintes suivantes

$$\phi_1 < \phi_2 < \dots \phi_m, \quad \text{et} \quad \pi_j > 0 \quad \text{pour } j = 1, \dots, m.$$

Sous ces contraintes la forme de l'ensemble de paramètres Θ devient compliquée, ce qui peut entraîner des problèmes sérieux au niveau du calcul d'estimateurs. En effet, certains algorithmes d'optimisation ne peuvent pas prendre en compte de telles contraintes sur l'espace des paramètres.

En général, on se contente d'une notion d'identifiabilité plus faible. On dit que la famille de lois de mélange

$$\mathcal{G} = \left\{ g(\cdot) = \sum_{j=1}^{m} \pi_j h_{\phi_j}(\cdot), \pi_j > 0, \sum_{j=1}^{m} \pi_j = 1, \phi_j \in \Phi \right\}$$

est **identifiable à une permutation des paramètres près** si et seulement si pour tout $g = \sum_{j=1}^{m} \pi_j h_{\phi_j} \in \mathcal{G}$ et $g' = \sum_{j=1}^{m} \pi'_j h_{\phi'_j} \in \mathcal{G}$, on a

$$g = g' \Longrightarrow \mathbb{Q} = \mathbb{Q}',$$

où $\mathbb{Q}$ est $\mathbb{Q}'$ sont les lois latentes respectives des mélanges définis par g et g', à savoir

$$\mathbb{Q} = \sum_{j=1}^{m} \pi_j \delta_{\{\phi_j\}} \qquad \text{et} \qquad \mathbb{Q}' = \sum_{j=1}^{m} \pi'_j \delta_{\{\phi'_j\}}.$$

La plupart de mélange de lois usuelles (gaussien, Poisson, Beta, Gamma) sont bien identifiable à une permutation des paramètres près. Un contre-exemple est le mélange de lois Bernoulli ou de lois uniformes (cf. TD).

3.2.7 Estimation de paramètres

Revenons aux exemples des longueurs d'aile des passereaux et du chlorure dans le sang. Notons $\mathbf{x} = (x_1, \ldots, x_n)$ un échantillon i.i.d. d'un modèle de mélange avec deux populations ($m = 2$), où chaque composante suit une loi normale. La densité de mélange f_θ s'écrit donc comme

$$\begin{aligned} f_\theta(x) &= p f_{\mathcal{N}(\mu_1,\sigma_1^2)} + (1-p) f_{\mathcal{N}(\mu_2,\sigma_2^2)} \\ &= \frac{p}{\sqrt{2\pi\sigma_1^2}} \exp\left\{-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right\} + \frac{1-p}{\sqrt{2\pi\sigma_2^2}} \exp\left\{-\frac{(x-\mu_2)^2}{2\sigma_2^2}\right\}, \end{aligned} \tag{3.5}$$

avec $p \in]0,1[$. On cherche à estimer les paramètres inconnus $\theta = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, p)$.

Pour la méthode du maximum de vraisemblance, on calcule la fonction de vraisemblance par

$$\mathcal{L}(\mathbf{x};\theta) = \prod_{i=1}^{n} f_\theta(x_i) = \frac{1}{(2\pi)^{n/2}} \prod_{i=1}^{n} \left(\frac{p}{\sigma_1} \exp\left\{-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right\} + \frac{1-p}{\sigma_2} \exp\left\{-\frac{(x-\mu_2)^2}{2\sigma_2^2}\right\} \right).$$

et la fonction de log-vraisemblance

$$\begin{aligned} \ell(\theta) &= \sum_{i=1}^{n} \log f_\theta(x_i) \\ &= -\frac{n}{2}\log(2\pi) + \sum_{i=1}^{n} \log\left(\frac{p}{\sigma_1} \exp\left\{-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right\} + \frac{1-p}{\sigma_2} \exp\left\{-\frac{(x-\mu_2)^2}{2\sigma_2^2}\right\} \right). \end{aligned}$$

Or,

$$\begin{aligned} \frac{\partial}{\partial p}\ell(\theta) &= \sum_{i=1}^{n} \frac{\frac{1}{\sigma_1}\exp\left\{-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right\} - \frac{1}{\sigma_2}\exp\left\{-\frac{(x-\mu_2)^2}{2\sigma_2^2}\right\}}{\frac{p}{\sigma_1}\exp\left\{-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right\} + \frac{1-p}{\sigma_2}\exp\left\{-\frac{(x-\mu_2)^2}{2\sigma_2^2}\right\}} \\ \frac{\partial}{\partial \mu_1}\ell(\theta) &= \sum_{i=1}^{n} \frac{\frac{p}{\sigma_1^3}(x_i-\mu_1)\exp\left\{-\frac{(x_i-\mu_1)^2}{2\sigma_1^2}\right\}}{\frac{p}{\sigma_1}\exp\left\{-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right\} + \frac{1-p}{\sigma_2}\exp\left\{-\frac{(x-\mu_2)^2}{2\sigma_2^2}\right\}} \\ \frac{\partial}{\partial \mu_2}\ell(\theta) &= \ldots \end{aligned}$$

Il est clair que l'équation $\nabla\ell(\theta) = 0$ n'admet pas de solution explicite. En effet, le fait que la fonction de vraisemblance s'écrit comme un produit de sommes rend sa maximisation assez compliquée. Le calcul de l'estimateur du maximum de vraisemblance dans de modèles de mélange nécessite généralement des méthodes numériques.

3.2.8 Modèles à variables latentes

Les modèlse de mélange font partie d'une famille de modèles plus large, qui sont ceux qui font intervenir des variables cachées comme l'étiquette U dans le modèle de mélange (la variable qui désigne l'appartenance de groupe). On parle aussi de **variables latentes** ou de **variables manquantes**, quand il y a des variables du modèle qui ne sont pas observées, et on appelle ces modèles des **modèles à variables latentes**.

On peut citer comme autre exemple de modèle à variables latentes les **données tronquées** où les observations sont données par $X_i = \min\{Z_i, c\}$, où $c \in \mathbb{R}$ est une constante et les Z_i sont des variables i.i.d. de loi f_θ. Ainsi, si $Z_i > c$, on n'observe pas la valeur de Z_i.

Un autre exemple similaire sont les **données censurées**. Soient Z_i des variables aléatoires i.i.d. de loi f_θ et W_i des variables aléatoires i.i.d. de loi g_λ. Supposons que les $(Z_i)_i$ et $(W_i)_i$ sont indépendantes. On observe pour $i = 1, \dots, n$ les variables δ_i et X_i définies comme

$$\delta_i = \mathbb{1}_{\{W_i \leq Z_i\}} \quad \text{et} \quad X_i = \min\{Y_i, W_i\} = \begin{cases} Y_i, & \text{si } \delta_i = 0 \\ W_i, & \text{si } \delta_i = 1 \end{cases}$$

Dans la suite nous utiliserons les notations suivantes. Soit $\mathbf{x}$ un échantillon i.i.d. de densité p_{θ_0} dans le modèle statistique $\{p_\theta, \theta \in \Theta\}$ avec $\Theta \subset \mathbb{R}^d$. On dit que $\mathbf{x}$ sont les **données incomplètes** du modèle. On dénote $\mathbf{u}$ les variables latentes du modèle. On dit que $(\mathbf{x}, \mathbf{u})$ sont les **données complètes** du modèle. On considère $(\mathbf{x}, \mathbf{u})$ comme une réalisation de densité q_{θ_0} dans un modèle statistique $\{q_\theta, \theta \in \Theta\}$. Ainsi, p_θ est la loi marginale de q_θ, à savoir

$$p_\theta(\mathbf{x}) = \int_{\mathbf{z}} q_\theta(\mathbf{x}, \mathbf{z}) \mu(\mathrm{d}\mathbf{z}).$$

Typiquement, un modèle de données incomplètes $\{p_\theta, \theta \in \Theta\}$ est très compliqué de sorte que les estimateurs classiques (EMM ou EMV) ne sont pas calculables explicitement. L'objectif de l'introduction de variables latentes dans le modèle est de passer à un modèle pour lequel les calculs se passent mieux. En effet, on se rend compte facilement dans l'exemple d'un mélange gaussien que l'EMV serait explicite si l'on avait observé les données complètes $(\mathbf{x}, \mathbf{u})$.

Dans le paragraphe suivant nous présenterons une méthode numérique pour approcher l'EMV dans des modèles à variables latentes, qui exploite justement le fait que l'EMV dans le modèle des données complètes est abordable.

3.3 Algorithme EM

L'algorithme EM de Dempster, Laird et Rubin (1977) est une procédure itérative pour approcher l'EMV dans des modèles à variables latentes.

3.3.1 Contexte d'application

Soit $(\mathbf{u}, \mathbf{x})$ un échantillon de la densité q_{θ_0} dans le modèle statistique $\{q_\theta, \theta \in \Theta\}$ avec $\Theta \subset \mathbb{R}^d$. On observe $\mathbf{x}$, mais pas $\mathbf{u}$. Nous supposons que le modèle est tel que la maximisation de la log-vraisemblance $\theta \mapsto \log p_\theta(\mathbf{x})$ des données incomplète n'a pas de solution explicite, alors que la fonction de log-vraisemblance $\theta \mapsto \log q_\theta(\mathbf{x}, \mathbf{u})$ des données complètes est facile à maximiser.

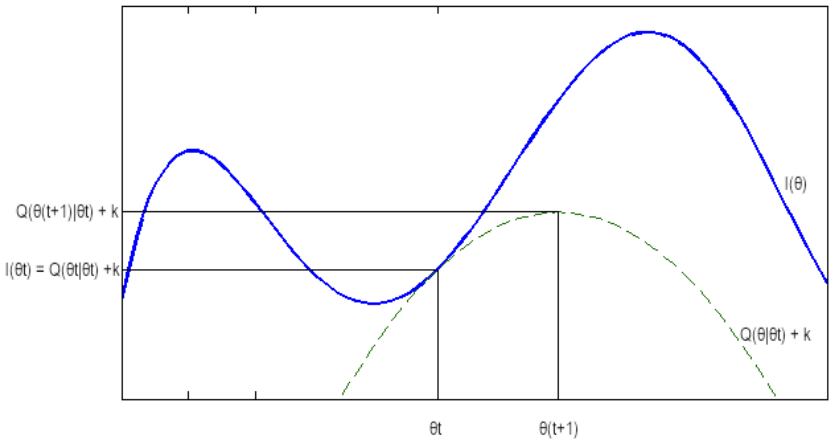


FIGURE 3.5 – Illustration d'une itération de l'algorithme EM. La courbe solide représente la log-vraisemblance $\theta \mapsto \ell(\theta) = \log p_\theta(\mathbf{x})$, qui est minorée par $\theta \mapsto Q(\theta|\theta^{(t)}) + k$ (courbe hachurée). Les deux courbes se croisent en la valeur actuelle $\theta^{(t)}$ de θ. En maximisant $\theta \mapsto Q(\theta|\theta^{(t)})$, on augmente $\theta \mapsto \ell(\theta)$.

3.3.2 L'ALGORITHME EM

L'idée de l'algorithme EM est d'utiliser l'espérance conditionnelle de la log-vraisemblance des données complètes sachant les observations $\mathbf{x}$ comme approximation de la log-vraisemblance des données incomplètes. Ainsi, on "estime" les variables latentes $\mathbf{u}$ et on exploite le fait que la log-vraisemblance $\theta \mapsto \log q_\theta(\mathbf{x}, \mathbf{u})$ est facile à maximiser. Or, afin de calculer cette espérance conditionnelle, il faut connaître la loi conditionnelle des variables latentes $\mathbf{U}$ sachant $\mathbf{X} = \mathbf{x}$ sous la loi $\mathbb{P}_\theta$. Il faudrait alors connaitre la vraie valeur du paramètre θ, ce qui n'est évidemment pas le cas. On procède alors de façon itérative : en partant d'une valeur initiale du paramètre $\theta^{(0)}$, on évalue l'espérance conditionnelle de la log-vraisemblance et ensuite on met à jour le paramètre $\theta^{(t)}$ en maximisant cette espérance conditionnelle. Concrètement, l'itération t de l'**algorithme EM** consiste consiste en les étapes suivantes :

Étape E. Calculer la fonction

$$\theta \mapsto Q(\theta|\theta^{(t-1)}) \qquad \text{avec} \qquad Q(\theta|\theta') := \mathbb{E}_{\theta'}[\log q_\theta(\mathbf{X}, \mathbf{U})|\mathbf{X} = \mathbf{x}],$$

et où $\theta^{(t-1)}$ est le résultat de l'itération précédente.

Étape M. Maximiser la fonction $\theta \mapsto Q(\theta|\theta^{(t-1)})$. Plus précisément, calculer la nouvelle valeur de θ par

$$\theta^{(t)} = \arg\max_{\theta \in \Theta} Q(\theta|\theta^{(t-1)}).$$

La première étape est dite étape d'*espérance*, la deuxième étape de *maximisation*, d'où le nom de l'algorithme EM (en anglais *expectation-maximisation*).

3.3.3 PROPRIÉTÉS DE L'ALGORITHME EM

L'objectif de l'algorithme EM est d'approcher l'EMV. Bien qu'on ne puisse pas garantir que ce but soit toujours atteint, on peut montrer des propriétés importantes de cet algorithme. Notamment, à chaque itération la log-vraisemblance $\ell(\theta) = \log p_\theta(\mathbf{x})$ est augmentée.

Théorème 7. *Soit $(\theta^{(t)})_{t\geq 1}$ une suite obtenue par l'algorithme EM. La log-vraisemblance $\ell(\theta)$ des données incomplètes vérifie, pour tout t,*

$$\ell(\theta^{(t+1)}) \geq \ell(\theta^{(t)}).$$

Démonstration. Il suffit de montrer que, pour tout θ, il existe une constante k_θ telle que

(i) la log-vraisemblance $\ell(\theta)$ est minorée par $Q(\theta|\theta') + k_{\theta'}$ pour tout θ, θ' et

(ii) $Q(\theta|\theta) + k_\theta = \ell(\theta)$ pour tout θ.

En effet, avec ces deux propriétés de $Q(\theta|\theta')$ on a, pour tout t

$$\begin{aligned}
\ell(\theta^{(t)}) &\overset{(ii)}{=} Q(\theta^{(t)}|\theta^{(t)}) + k_{\theta^{(t)}} \\
&\leq \max_{\theta\in\Theta} Q(\theta|\theta^{(t)}) + k_{\theta^{(t)}} \\
&= Q(\theta^{(t+1)}|\theta^{(t)}) + k_{\theta^{(t)}} \\
&\overset{(i)}{\leq} \ell(\theta^{(t+1)}).
\end{aligned}$$

Autrement dit, en maximisant l'espérance conditionnelle $Q(\theta|\theta^{(t)})$ dans l'étape M, on obtient forcément une valeur $\theta^{(t+1)}$ où la log-vraisemblance est plus élevée qu'au point $\theta^{(t)}$, c'est-à-dire $\ell(\theta^{(t+1)}) \geq \ell(\theta^{(t)})$. Ce phénomène est illustré graphiquement dans la Figure 3.5.

Montrons donc (i) et (ii). Ce résultat repose essentiellement sur l'inégalité de Jensen, selon laquelle pour toute fonction h concave et toute v.a. Z on a $\mathbb{E}[h(Z)] \leq h(\mathbb{E}[Z])$. En effet, soient p et q deux densités par rapport à une même mesure μ, alors on obtient par la concavité du logarithme

$$\begin{aligned}
\mathbb{E}_p[\log q(Z)] - \mathbb{E}_p[\log p(Z)] &= \mathbb{E}_p\left[\log \frac{q(Z)}{p(Z)}\right] \\
&\leq \log\left(\mathbb{E}_p\left[\frac{q(Z)}{p(Z)}\right]\right) = \log\left(\int \frac{q(z)}{p(z)} p(z)\mu(\mathrm{d}z)\right) = 0,
\end{aligned}$$

d'où $\mathbb{E}_p[\log q(Z)] \leq \mathbb{E}_p[\log p(Z)]$. Nous appliquons cette inégalité aux densités conditionnelles des données complètes sachant les données observées, i.e. on pose

$$p(\mathbf{u}) = \frac{q_{\theta'}(\mathbf{x},\mathbf{u})}{p_{\theta'}(\mathbf{x})} \quad \text{et} \quad q(\mathbf{u}) = \frac{q_\theta(\mathbf{x},\mathbf{u})}{p_\theta(\mathbf{x})}.$$

On obtient alors

$$\begin{aligned}
Q(\theta|\theta') - \ell(\theta) &= \mathbb{E}_{\theta'}[\log q_\theta(\mathbf{x},\mathbf{U})|\mathbf{X}=\mathbf{x}] - \log p_\theta(\mathbf{x}) \\
&= \mathbb{E}_{\theta'}\left[\log \frac{q_\theta(\mathbf{x},\mathbf{U})}{p_\theta(\mathbf{x})}\middle|\mathbf{X}=\mathbf{x}\right] \\
&\leq \mathbb{E}_{\theta'}\left[\log \frac{q_{\theta'}(\mathbf{x},\mathbf{U})}{p_{\theta'}(\mathbf{x})}\middle|\mathbf{X}=\mathbf{x}\right] \\
&= Q(\theta'|\theta') - \ell(\theta'),
\end{aligned}$$

avec égalité pour $\theta = \theta'$. Par conséquent, $\theta \mapsto Q(\theta|\theta') - Q(\theta'|\theta') + \ell(\theta')$ est un minorant de $\theta \mapsto \ell(\theta)$, et en posant $k_{\theta'} = -Q(\theta'|\theta') + \ell(\theta')$ on obtient (i) et (ii). □

Bien qu'on ait montré que l'algorithme EM augmente la log-vraisemblance à chaque itération, cela ne garantit évidemment pas sa convergence vers le maximum global. On peut montrer sous des conditions assez générales que l'algorithme EM converge toujours vers un point critique de la log-vraisemblance $\ell(\theta)$, mais il est possible que ce soit seulement un maximum local ou un point de selle.

3.3.4 Aspects pratiques

Initialisation

Comme pour de nombreux algorithmes d'optimisation, la limite de l'algorithme EM dépend de son initialisation. Autrement dit, en fonction du choix de $\theta^{(0)}$, l'algorithme peut converger vers l'EMV ou non. Il n'y a pas de règle générale pour un choix convenable de $\theta^{(0)}$, il faut regarder au cas par cas. On peut parfois initialiser avec un estimateur simple (et pas très précis) de θ_0. Si la convergence de l'algorithme n'est pas trop lente, il est envisageable de lancer l'algorithme plusieurs fois avec différents points initiaux $\theta^{(0)}$ choisis au hasard. On choisit ensuite comme estimateur de θ_0 le résultat avec la plus grande vraisemblance.

Critère d'arrêt

Après combien d'itérations convient-t-il d'arrêter l'algorithme EM ? Comment savoir si l'algorithme a convergé ? Il n'y a pas de nombre d'itérations qui soit convenable pour tous les modèles. Dans certains cas l'algorithme converge rapidement, dans d'autres très lentement. De manière générale, plus le nombre de variables latentes est important, plus la convergence est lente.

On peut observer la suite $(\theta^{(t)})_{t\geq 1}$ et arrêter l'algorithme quand la différence entre deux $\theta^{(t)}$ successifs est petits, c'est-à-dire lorsque $\|\theta^{(t+1)} - \theta^{(t)}\| < \varepsilon$, où $\|\cdot\|$ désigne ou la norme euclidienne ou la norme sup et $\varepsilon > 0$ est un seuil fixé par avance. Comme les éléments de θ ne sont pas forcément tous du même ordre de grandeur, il vaut mieux considérer l'erreur relative, à savoir $\|\theta^{(t+1)} - \theta^{(t)}/\theta^{(t)}\|$, où la division est élément par élément.

Autre alternative : on peut fonder le critère d'arrêt sur la fonction de log-vraisemblance, plus précisément sur la convergence de la suite $(\ell(\theta^{(t)}))_{t\geq 1}$. On arrête dès que l'augmentation entre deux itérations est trop faible, c'est-à-dire quand $\ell(\theta^{(t+1)}) - \ell(\theta^{(t)}) < \varepsilon$ pour un seuil $\varepsilon > 0$ donné.

Dans les deux cas, il est possible que qu'on s'arrête trop tôt, quand l'algorithme n'a pas encore convergé. Cela arrive si la convergence est très lente ou si la fonction de vraisemblance est très plate.

Grâce à son caractère général, l'algorithme EM s'applique à des problèmes très variés et son utilisation est très répandue en pratique. Au fil du temps, de nombreuses variantes de cet algorithme sont nées. Dans ce cours nous nous contenterons d'étudier le cadre classique du modèle de mélange.

3.3.5 Exemple : Mélange gaussien

Reprenons l'exemple d'un mélange de deux lois normales, dont la densité de mélange f_θ est donnée par (3.5). Autrement dit, $\mathbf{x} = (x_1, \ldots, x_n)$ est un échantillon i.i.d. de la variable aléatoire X définie par

$$X = \mathbb{1}_{\{U=1\}} V_1 + \mathbb{1}_{\{U=2\}} V_2,$$

où U, V_1, V_2 sont des variables aléatoires indépendantes telles que V_j suit la loi $\mathcal{N}(\mu_j, \sigma_j^2)$ pour $j = 1, 2$ et U vérifie $\mathbb{P}(U = 1) = 1 - \mathbb{P}(U = 2) = p$. L'étiquette U n'est pas observée, il s'agit de la variable latente du modèle. Notons u_i la réalisation de U associée à l'observation x_i, et $\mathbf{u} = (u_1, \ldots, u_n)$.

Il est clair que la loi conditionnelle de X sachant que $U = u$ est la loi normale $\mathcal{N}(\mu_u, \sigma_u^2)$,

i.e.

$$f_{X|U}(x|u) = f_{\mathcal{N}(\mu_u,\sigma_u^2)}(x), \quad \text{pour tout } x \in \mathbb{R}, u \in \{1,2\}.$$

Or, la densité jointe $p_{(X,U)}$ de (X,U) par rapport à la mesure $\nu = \lambda \otimes \delta_{\{1,2\}}$, où λ désigne la mesure de Lebesgue sur $\mathbb{R}$ et $\delta_{\{1,2\}}$ la mesure de comptage sur $\{1,2\}$, est donnée par

$$p_{(X,U)}(x,u) = f_{X|U}(x|u)p_U(u) = p^{\mathbb{1}\{u=1\}}(1-p)^{\mathbb{1}\{u=2\}} f_{\mathcal{N}(\mu_u,\sigma_u^2)}(x), \quad u \in \{1,2\}, x \in \mathbb{R}.$$

La densité jointe q_θ de l'échantillon $(\mathbf{x},\mathbf{u}) = (x_1,\dots,x_n,u_1,\dots,u_n)$ est donc

$$q_\theta(\mathbf{x},\mathbf{u}) = \prod_{i=1}^n p_{(X,U)}(x_i,u_i) = p^{\sum_{i=1}^n \mathbb{1}\{u_i=1\}}(1-p)^{n-\sum_{i=1}^n \mathbb{1}\{u_i=1\}} \prod_{i=1}^n f_{\mathcal{N}(\mu_{u_i},\sigma_{u_i}^2)}(x_i).$$

On en déduit la fonction Q de l'algorithme EM

$$\begin{aligned}
Q(\theta|\theta') &= \mathbb{E}_{\theta'}\left[\log q_\theta(\mathbf{x},\mathbf{U})|\mathbf{X}=\mathbf{x}\right] \\
&= \mathbb{E}_{\theta'}\left[\sum_{i=1}^n \mathbb{1}_{\{U_i=1\}}\log p + \left(n - \sum_{i=1}^n \mathbb{1}_{\{U_i=1\}}\right)\log(1-p) + \sum_{i=1}^n \log\left(f_{\mathcal{N}(\mu_{U_i},\sigma_{U_i}^2)}(x_i)\right)\middle|\mathbf{X}=\mathbf{x}\right] \\
&= \log p \sum_{i=1}^n \mathbb{E}_{\theta'}\left[\mathbb{1}_{\{U=1\}}\middle|X=x_i\right] + \log(1-p)\left(n - \sum_{i=1}^n \mathbb{E}_{\theta'}\left[\mathbb{1}_{\{U=1\}}\middle|X=x_i\right]\right) + \\
&\quad + \sum_{i=1}^n \mathbb{E}_{\theta'}\left[\log\left(f_{\mathcal{N}(\mu_U,\sigma_U^2)}(x_i)\right)\middle|X=x_i\right].
\end{aligned}$$

D'une part, $\mathbb{E}_{\theta'}\left[\mathbb{1}_{\{U=1\}}\middle|X=x_i\right] = \mathbb{P}_{\theta'}(U=1|X=x_i) =: \pi_{\theta'}(x_i)$. D'autre part,

$$\begin{aligned}
\mathbb{E}_{\theta'}\left[\log\left(f_{\mathcal{N}(\mu_U,\sigma_U^2)}(x_i)\right)\middle|X=x_i\right] &= \mathbb{E}_{\theta'}\left[\log\left(\frac{1}{\sqrt{2\pi}\sigma_U}\exp\left\{-\frac{(x_i-\mu_U)^2}{2\sigma_U^2}\right\}\right)\middle|X=x_i\right] \\
&= -\log\sqrt{2\pi} - \mathbb{E}_{\theta'}[\log(\sigma_U)|X=x_i] - \mathbb{E}_{\theta'}\left[\frac{(x_i-\mu_U)^2}{2\sigma_U^2}\middle|X=x_i\right] \\
&= -\log\sqrt{2\pi} - [\log(\sigma_1)\pi_{\theta'}(x_i) + \log(\sigma_2)(1-\pi_{\theta'}(x_i))] \\
&\quad - \frac{1}{2}\left\{\frac{(x_i-\mu_1)^2}{\sigma_1^2}\pi_{\theta'}(x_i) + \frac{(x_i-\mu_2)^2}{\sigma_2^2}(1-\pi_{\theta'}(x_i))\right\}.
\end{aligned}$$

Il suffit de déterminer les probabilités conditionnelles $\pi_{\theta'}(x_i)$ pour tout $i = 1,\dots,n$, pour connaitre entièrement la fonction $Q(\theta|\theta')$. Or,

$$\pi_{\theta'}(x) = \mathbb{P}_{\theta'}(U=1|X=x) = \frac{p_{(X,U)}(x,1)}{f_X(x)} = \frac{p' f_{\mathcal{N}(\mu_1',\sigma_1'^2)}(x)}{p' f_{\mathcal{N}(\mu_1',\sigma_1'^2)}(x) + (1-p') f_{\mathcal{N}(\mu_2',\sigma_2'^2)}(x)}. \tag{3.6}$$

Notons que $0 < \pi_{\theta'}(x) < 1$ pour tout x et θ'.

Enfin, on obtient

$$\begin{aligned}
Q(\theta|\theta') &= \log p \sum_{i=1}^n \pi_{\theta'}(x_i) + \log(1-p)\left(n - \sum_{i=1}^n \pi_{\theta'}(x_i)\right) + \\
&\quad - n\log\sqrt{2\pi} - \frac{1}{2}\log(\sigma_1^2)\sum_{i=1}^n \pi_{\theta'}(x_i) - \frac{1}{2}\log(\sigma_2^2)\sum_{i=1}^n (1-\pi_{\theta'}(x_i)) \\
&\quad - \frac{1}{2\sigma_1^2}\sum_{i=1}^n \pi_{\theta'}(x_i)(x_i-\mu_1)^2 - \frac{1}{2\sigma_2^2}\sum_{i=1}^n (1-\pi_{\theta'}(x_i))(x_i-\mu_2)^2
\end{aligned}$$

Pour la maximisation de la fonction $\theta \mapsto Q(\theta|\theta')$ dans l'étape M, il est utile de constater que l'on peut décomposer la maximisation en trois problèmes indépendants, car $Q(\theta|\theta')$ s'écrit comme la somme de trois fonctions : $Q(\theta|\theta') = Q_1(p|\theta') + Q_2(\mu_1, \sigma_1|\theta') + Q_3(\mu_2, \sigma_2|\theta')$. On obtient pour les dérivées partielles,

$$\begin{aligned}
\frac{\partial}{\partial p} Q(\theta|\theta') &= \frac{1}{p}\sum_{i=1}^{n} \pi_{\theta'}(x_i) - \frac{1}{1-p}\left(n - \sum_{i=1}^{n} \pi_{\theta'}(x_i)\right) \\
\frac{\partial^2}{\partial^2 p} Q(\theta|\theta') &= -\frac{1}{p^2}\sum_{i=1}^{n} \pi_{\theta'}(x_i) - \frac{1}{(1-p)^2}\left(n - \sum_{i=1}^{n} \pi_{\theta'}(x_i)\right) < 0 \\
\frac{\partial}{\partial \mu_1} Q(\theta|\theta') &= \frac{1}{\sigma_1^2}\sum_{i=1}^{n} (x_i - \mu_1)\pi_{\theta'}(x_i) = \frac{1}{\sigma_1^2}\left[\sum_{i=1}^{n} x_i \pi_{\theta'}(x_i) - \mu_1 \sum_{i=1}^{n} \pi_{\theta'}(x_i)\right] \\
\frac{\partial^2}{\partial^2 \mu_1} Q(\theta|\theta') &= -\frac{1}{\sigma_1^2}\sum_{i=1}^{n} \pi_{\theta'}(x_i) < 0 \\
\frac{\partial}{\partial \sigma_1^2} Q(\theta|\theta') &= -\frac{1}{2\sigma_1^2}\sum_{i=1}^{n} \pi_{\theta'}(x_i) + \frac{1}{2\sigma_1^4}\sum_{i=1}^{n} \pi_{\theta'}(x_i)(x_i - \mu_1)^2 \\
\frac{\partial^2}{\partial^2 \sigma_1^2} Q(\theta|\theta') &= \frac{1}{2\sigma_1^4}\sum_{i=1}^{n} \pi_{\theta'}(x_i) - \frac{1}{\sigma_1^6}\sum_{i=1}^{n} \pi_{\theta'}(x_i)(x_i - \mu_1)^2
\end{aligned}$$

et similaire pour les dérivées partielles par rapport à μ_2 et σ_2^2.

Les fonctions $p \mapsto Q(\theta|\theta')$ et $\mu_1 \mapsto Q(\theta|\theta')$ étant strictement concave, on trouve leur maxima par les points critiques qui sont uniques. Ainsi

$$\begin{aligned}
\frac{\partial}{\partial p} Q(\theta|\theta') = 0 &\iff p = \frac{1}{n}\sum_{i=1}^{n} \pi_{\theta'}(x_i) \\
\frac{\partial}{\partial \mu_1} Q(\theta|\theta') = 0 &\iff \mu_1 = \frac{\sum_{i=1}^{n} x_i \pi_{\theta'}(x_i)}{\sum_{i=1}^{n} \pi_{\theta'}(x_i)} =: \hat{\mu}_1.
\end{aligned}$$

Maintenant, maximisant $\sigma_1^2 \mapsto Q(\theta|\theta')\big|_{\mu_1=\hat{\mu}_1}$. D'abord on trouve pour le point critique

$$\frac{\partial}{\partial \sigma_1^2} Q(\theta|\theta')\bigg|_{\mu_1=\hat{\mu}_1} = 0 \quad \iff \quad \sigma_1^2 = \frac{\sum_{i=1}^{n} \pi_{\theta'}(x_i)(x_i - \hat{\mu}_1)^2}{\sum_{i=1}^{n} \pi_{\theta'}(x_i)} =: \hat{\sigma}_1^2,$$

ensuite on vérifie qu'on a bien

$$\frac{\partial^2}{\partial^2 \sigma_1^2} Q(\theta|\theta')\bigg|_{\mu_1=\hat{\mu}_1, \sigma_1=\hat{\sigma}_1^2} = -\frac{\left(\sum_{i=1}^{n} \pi_{\theta'}(x_i)\right)^2}{2\left(\sum_{i=1}^{n} \pi_{\theta'}(x_i)(x_i - \hat{\mu}_1)^2\right)^3} < 0.$$

Donc, il s'agit bien d'un maximum local. De plus, étant l'unique point critique, il s'agit du maximum global de la fonction.

Les calculs pour μ_2 et σ_2^2 sont analogues.

En conclusion, l'algorithme EM consiste à calculer successivement pour tout $t = 1, 2, \ldots$

$$\begin{aligned}
p^{(t+1)} &= \frac{\sum_{i=1}^{n} \pi_{\theta^{(t)}}(x_i)}{n}, \quad \mu_1^{(t+1)} = \frac{\sum_{i=1}^{n} x_i \pi_{\theta^{(t)}}(x_i)}{\sum_{i=1}^{n} \pi_{\theta^{(t)}}(x_i)}, \quad \mu_2^{(t+1)} = \frac{\sum_{i=1}^{n} x_i (1 - \pi_{\theta^{(t)}}(x_i))}{n - \sum_{i=1}^{n} \pi_{\theta^{(t)}}(x_i)} \\
\sigma_1^{2(t+1)} &= \frac{\sum_{i=1}^{n} \pi_{\theta^{(t)}}(x_i)(x_i - \mu_1^{(t+1)})^2}{\sum_{i=1}^{n} \pi_{\theta^{(t)}}(x_i)}, \quad \sigma_2^{2(t+1)} = \frac{\sum_{i=1}^{n} (1 - \pi_{\theta^{(t)}}(x_i))(x_i - \mu_2^{(t+1)})^2}{\sum_{i=1}^{n} (1 - \pi_{\theta^{(t)}}(x_i))},
\end{aligned}$$

où les $\pi_{\theta^{(t)}}(x_i)$ sont donnés par (3.6).

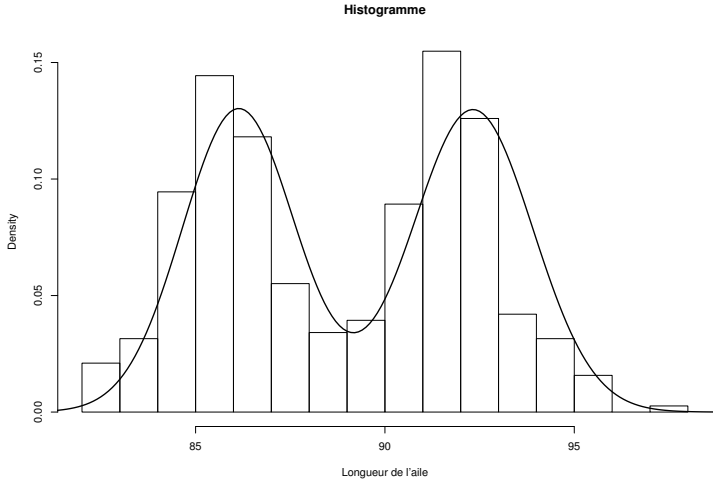
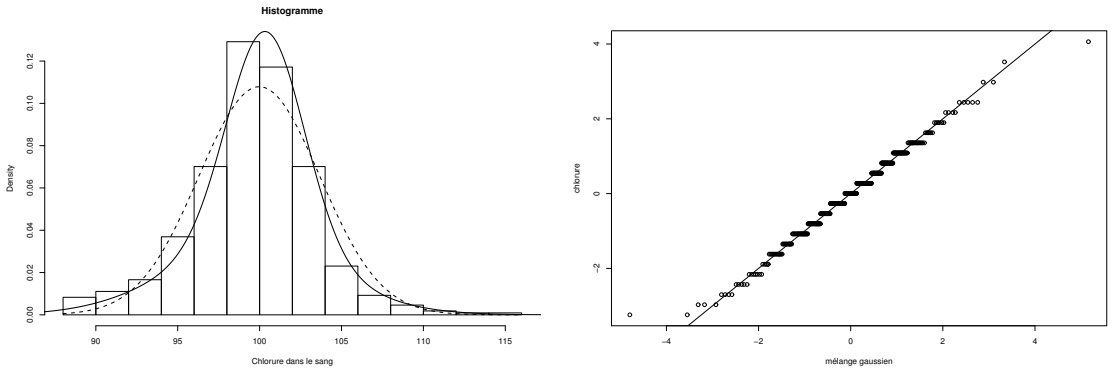


FIGURE 3.6 – Histogramme des longueurs des ailes et densité d'un mélange de deux lois normales dont les paramètres sont estimés par l'algorithme EM.



(a) (b)

FIGURE 3.7 – (a) Histogramme des données de chlorure dans le sang, la densité de la loi normale $\mathcal{N}(\hat{\mu}, \hat{\sigma}^2)$ où $\hat{\mu}$ et $\hat{\sigma}^2$ sont l'EMV (en trait pointillé) et la densité d'un mélange de deux lois normales dont les paramètres sont estimés par l'algorithme EM. (b) QQ-plot des données standardisées en comparaison au mélange gaussien estimé.

Revenons aux exemples des passereaux et du chlorure pour illustrer cet algorithme. Les Figures 3.6 et 3.7 (a) montrent les densités d'un mélange de deux lois normales où les paramètres ont été calculé par l'algorithme EM décrit ci-dessus. On observe la bonne adéquation des densités estimées avec les données. Dans le cas du chlorure, nous voyons que les points du QQ-plot de la Figure 3.7 (b) qui compare les données au mélange gaussien estimé donnent un meilleur résultat que dans le cas d'une simple loi normale (Figure 3.2 (b)).

3.4 Échantillonneur de Gibbs

Dans cette dernière partie du cours nous verrons une méthode alternative d'estimation d'un modèle de mélange. Il s'agit d'un estimateur Bayésien que l'on calcule par un échantillonneur de Gibbs.

3.4.1 Approche bayésienne

En statistique il y a deux grandes écoles : l'approche fréquentiste et l'approche bayésienne. Nous présentons dans ce paragraphe les fondamentaux de la statistique bayésienne en comparaison à la statistique fréquentiste.

Modélisation

Approche fréquentiste C'est l'approche que nous avons considéré jusqu'ici dans lequel on se donne un modèle statistique $\{\mathbb{P}_\theta, \theta \in \Theta\}$ et on observe les variables aléatoires $X_i, i = 1, \dots, n$ qui sont de loi $\mathbb{P}_{\theta_0}$ avec $\theta_0 \in \Theta$ et dans notre cas i.i.d..

Approche bayésienne On considère que le paramètre θ est lui-même une variable aléatoire ! Plus précisément, en plus du modèle $\{\mathbb{P}_\theta, \theta \in \Theta\}$, on introduit une loi de probabilité π sur Θ, que l'on appelle **loi a priori**. Ainsi, on suppose le modèle hiérarchique suivant :

(i) Le paramètre θ est une variable aléatoire de loi π que l'on n'observe pas.

(ii) Conditionnellement à θ, les variables aléatoires $X_i, i = 1, \dots, n$ sont i.i.d. de loi $\mathbb{P}_\theta$.

Estimation

Approche fréquentiste On estime le paramètre θ_0 à partir de réalisations $\mathbf{x} = (x_1, \dots, x_n)$ de $\mathbf{X} = (X_1, \dots, X_n)$, par exemple par le maximum de vraisemblance.

Approche bayésienne Tout d'abord, on cherche à estimer la *loi* conditionnelle de θ sachant les observations $\mathbf{x} = (x_1, \dots, x_n)$, à savoir

$$p_{\theta|\mathbf{X}}(\theta|\mathbf{x}) = \frac{p_{\mathbf{X},\theta}(\mathbf{x}, \theta)}{p_{\mathbf{X}}(\mathbf{x})},$$

que l'on appelle **loi a posteriori**.

Dans la notation usuelle en bayésien on omet les indices et on note

— $\pi(\theta|\mathbf{x})$ pour la loi a posteriori $p_{\theta|\mathbf{X}}(\theta|\mathbf{x})$,

— $\pi(\mathbf{x}, \theta)$ pour la loi jointe $p_{\mathbf{X},\theta}(\mathbf{x}, \theta)$ de $(\mathbf{X}, \theta)$ et
— $\pi(\mathbf{x})$ pour la loi marginale des observations $p_{\mathbf{X}}(\mathbf{x})$.

On construit des estimateurs de la valeur de θ avec laquelle les observations $\mathbf{x}$ ont été générées, à partir de la loi a posteriori $\pi(\theta|\mathbf{x})$. Ainsi, un estimateur de θ est la **moyenne a posteriori** définie comme

$$\hat{\theta} = \mathbb{E}[\theta|\mathbf{X} = \mathbf{x}] = \int_{\Theta} \theta\pi(\theta|\mathbf{x})\mathrm{d}\theta.$$

Un autre estimateur est donné par le **maximum a posteriori** (MAP) défini par

$$\hat{\theta}^{MAP} = \arg\max_{\theta\in\Theta} \pi(\theta|\mathbf{x}).$$

Exemple. (Loi de Poisson avec loi Gamma comme loi a priori) Considérons le modèle où les $X_i|\lambda, i = 1, \ldots, n$ sont i.i.d. de loi de Poisson Poi(λ), et le paramètre λ est de loi Gamma $\Gamma(\alpha, 1)$. On dit que α est le hyperparamètre du modèle (que l'on suppose connu/choisi par l'utilisateur). Autrement dit, la densité a priori de λ s'écrit

$$\pi(\lambda) = \frac{1}{\Gamma(\alpha)}\lambda^{\alpha-1}\mathrm{e}^{-\lambda}, \quad \lambda > 0,$$

et la densité conditionnelle de $X_i, i = 1, \ldots, n$ sachant λ est donnée par

$$\pi(\mathbf{x}|\lambda) = \prod_{i=1}^{n} \pi(x_i|\lambda) = \prod_{i=1}^{n} \left[\frac{\lambda^{x_i}}{x_i!}\mathrm{e}^{-\lambda}\right] = \frac{\lambda^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!}\mathrm{e}^{-n\lambda}.$$

Ainsi, la loi jointe de $\mathbf{X}$ et λ est donnée par

$$\begin{aligned}\pi(\mathbf{x}, \lambda) = \pi(\mathbf{x}|\lambda)\pi(\lambda) &= \frac{\lambda^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!}\mathrm{e}^{-n\lambda} \times \frac{1}{\Gamma(\alpha)}\lambda^{\alpha-1}\mathrm{e}^{-\lambda} \\ &= \frac{\lambda^{\sum_{i=1}^n x_i - \alpha - 1}}{\Gamma(\alpha)\prod_{i=1}^n x_i!}\mathrm{e}^{-(n+1)\lambda}.\end{aligned}$$

On en déduit la loi a posteriori de λ sachant les observations $\mathbf{x}$:

$$\begin{aligned}\pi(\lambda|\mathbf{x}) &= \frac{\pi(\mathbf{x}, \lambda)}{\pi(\mathbf{x}} \\ &= \text{cste} \times \lambda^{\sum_{i=1}^n x_i - \alpha - 1}\mathrm{e}^{-(n+1)\lambda} \\ &\propto \pi(\mathbf{x}, \lambda).\end{aligned}$$

La notation $\propto$ ("proportionnel à") est courante en statistique bayésienne, car en général les constantes de normalisation sont inutiles. Néanmoins, l'usage de ce symbole conduit vite à des erreurs. Il faut alors l'utiliser avec beaucoup de prudence.

On vient de montrer que la loi a posteriori de λ est une loi Gamma, à savoir

$$\lambda|\mathbf{x} \sim \Gamma\left(\sum_{i=1}^{n} x_i - \alpha, n + 1\right).$$

Rappelons que si $U \sim \Gamma(\alpha, \beta)$, alors $\mathbb{E}[U] = \alpha/\beta$. Ainsi, l'estimateur de λ par moyenne a posteriori est donné par

$$\hat{\lambda}^{\text{moyAP}} = E[\lambda|\mathbf{X} = \mathbf{x}] = \frac{\sum_{i=1}^n x_i - \alpha}{n + 1}.$$

On voit bien que plus la taille n d'échantillon est grande, moins la valeur de l'hyperparamètre α influence l'estimateur $\hat{\lambda}$.

Rappelons que l'estimateur du maximum de vraisemblance de λ dans le modèle fréquentiste est la moyenne empirique $\bar{x}_n$. Ainsi, pour n assez grand, la différence entre l'estimateur fréquentiste et l'estimateur bayésien $\hat{\lambda}^{\text{moyAP}}$ est négligeable.

Modèle de mélange

Approche fréquentiste La densité d'une observation X d'un modèle de mélange est de la forme

$$p_\theta(x) = \sum_{k=1}^{K} \pi_k h(\cdot, \varphi_k),$$

et on note $\theta = (\boldsymbol{\pi}, \boldsymbol{\varphi}) = (\pi_1, \ldots, \pi_K, \varphi_1, \ldots, \varphi_K)$ le vecteur de paramètres du modèle. Des variables aléatoires $X_i, i = 1, \ldots, n$ issues du mélange p_θ sont définies à l'aide de variables latentes Z_i, à savoir

(i) Les variables aléatoires $Z_i, i = 1, \ldots, n$ sont i.i.d. de loi $\sum_{k=1}^{K} \pi_k \delta_{\{k\}}$. Elles ne sont pas observées.

(ii) Conditionnellement aux $Z_i, i = 1, \ldots, n$, les variables aléatoires $X_i, i = 1, \ldots, n$ sont indépendantes de loi $h(\cdot, \varphi_{Z_i})$.

L'estimateur du maximum de vraisemblance de θ est approché par l'algorithme EM.

Approche bayésienne Pour un modèle bayésien de mélange, il faut introduire une loi a priori π sur Θ. On définit des variables aléatoires $X_i, i = 1, \ldots, n$ d'un mélange bayésien de façon suivante :

(i) Le paramètre θ est une variable aléatoire de loi π que l'on n'observe pas.

(ii) Conditionnellement à θ, les variables aléatoires $Z_i, i = 1, \ldots, n$ sont i.i.d. de loi $\sum_{k=1}^{K} \pi_k \delta_{\{k\}}$. Elles ne sont pas observées.

(iii) Conditionnellement aux $Z_i, i = 1, \ldots, n$ et à θ, les variables aléatoires $X_i, i = 1, \ldots, n$ sont indépendantes de loi $h(\cdot, \varphi_{Z_i})$.

La loi conditionnelle de $\mathbf{X} = (X_1, \ldots, X_n)$ sachant les variables latentes $\mathbf{Z} = (Z_1, \ldots, Z_n)$ et θ s'écrit

$$\pi(\mathbf{x}|\mathbf{z}, \theta) = \prod_{i=1}^{n} h(x_i, \varphi_{z_i}).$$

La loi conditionnelle de $\mathbf{Z}$ sachant θ est donnée par

$$\pi(\mathbf{z}|\theta) = \prod_{i=1}^{n} \pi_{z_i}.$$

Pour la densité a posteriori de θ sachant $\mathbf{x}$ on trouve

$$\begin{aligned}
\pi(\theta|\mathbf{x}) &= \frac{\pi(\mathbf{x}, \theta)}{\pi(\mathbf{x})} \\
&\propto \pi(\mathbf{x}, \theta) \\
&= \int_{\mathbf{z}} \pi(\mathbf{x}, \mathbf{z}, \theta) \mu(\mathrm{d}\mathbf{z}) \\
&= \int_{\mathbf{z}} \pi(\mathbf{x}|\mathbf{z}, \theta) \pi(\mathbf{z}|\theta) \pi(\theta) \mu(\mathrm{d}\mathbf{z}) \\
&= \int_{\mathbf{z}} \left\{ \prod_{i=1}^{n} h(x_i, \varphi_{z_i}) \times \prod_{i=1}^{n} \pi_{z_i} \right\} \mu(\mathrm{d}\mathbf{z}) \times \pi(\theta) \\
&= \prod_{i=1}^{n} \left(\int_{z_i} h(x_i, \varphi_{z_i}) \pi_{z_i} \mu(\mathrm{d}z_i) \right) \times \pi(\theta) \\
&= \prod_{i=1}^{n} \left(\sum_{k=1}^{K} h(x_i, \varphi_k) \pi_k \right) \times \pi(\theta).
\end{aligned}$$

On voit bien que la loi a posteriori est difficile à manipuler. En particulier, la densité ne se factorise pas (on n'arrive pas à séparer les paramètres en différents termes). Par conséquent, conditionnellement aux observations $\mathbf{x}$, les paramètres $\pi_1, \ldots, \pi_K, \varphi_1, \ldots, \varphi_K$ ne sont pas indépendantes.

Cela pose des problèmes sérieux pour l'estimation de la valeur de θ. En effet, la moyenne a posteriori $\mathbb{E}[\theta|\mathbf{x}] = \int_\Theta \theta\pi(\theta|\mathbf{x})\mathrm{d}\theta$ n'a pas d'expression analytique. D'ailleurs, aucun autre estimateur bayésien est explicite. Par ailleurs, la structure compliquée de dépendance des paramètres implique que la simulation de réalisations de la loi a posteriori $\pi(\theta|\mathbf{x})$ n'est pas évidente. Donc, il n'est pas non plus envisageable de faire des simulations de Monte-Carlo pour approcher la moyenne a posteriori. Nous verrons que la solution ici sera une méthode MCMC.

3.4.2 Rappel : Metropolis-Hastings

L'objectif est de simuler d'une loi π sur un ensemble E. L'algorithme de Metropolis-Hastings consiste à construire une chaîne de Markov $(X_k)_{k\geq 1}$ dont la probabilité invariante est π. Par le théorème ergodique ponctuel, on a pour toute fonction f intégrable

$$\frac{1}{n}\sum_{k=1}^{n} f(X_k) \xrightarrow{p.s.} \mathbb{E}_\pi[f(X)], \quad n \to \infty.$$

Algorithme

Soit P une matrice stochastique de proposition ou de transition telle que

$$\forall x, y \in E, \quad P(x,y) > 0 \Longleftrightarrow P(y,x) > 0.$$

Notons $\rho(\cdot,\cdot)$ le rapport de Metropolis-Hastings définit par

$$\rho(x,y) = \min\left\{\frac{\pi(y)P(y,x)}{\pi(x)P(x,y)}, 1\right\}.$$

À l'instant k, la valeur actuelle de x est x_k, et

(i) on choisit un voisin y de x_k avec probabilité $P(x_k, y)$, et

(ii) on pose

$$x_{k+1} = \begin{cases} y, & \text{avec probabilité } \rho(x_k,y) \text{ (on accepte } y) \\ x_k, & \text{avec probabilité } 1-\rho(x_k,y) \text{ (on refuse } y) \end{cases}$$

Un échantillonneur de Gibbs est un algorithme de Metropolis-Hastings avec

$$\rho(x,y) = 1, \quad \forall x, y \in E.$$

Autrement dit, toutes les propositions sont acceptées.

3.4.3 Échantillonneur de Gibbs

Soit $\pi(x)$ la loi cible que l'on ne sait pas simuler directement. Or, soit Y une variable aléatoire quelconque. Alors, $\pi(x)$ est la loi marginale de la loi jointe $\pi(x,y)$ de X et Y :

$$\pi(x) = \int_y \pi(x,y)\mu(\mathrm{d}y).$$

Si on sait simuler des réalisations (x_k, y_k) de la loi jointe $\pi(x, y)$, alors les valeurs simulées x_k sont des réalisations de la loi marginale $\pi(x)$. Mais comment simuler de la loi jointe $\pi(x, y)$ s'il est déjà difficile de simuler de la loi marginale ?

Souvent, on trouve une variable Y telle que la simulation des deux *lois conditionnelles* $\pi(x|y)$ et $\pi(y|x)$ est facile. On peut utiliser ces lois conditionnelles pour construire un noyau de transition P d'une chaîne de Markov dont la loi invariante est $\pi(x, y)$. Pour cela, il faut alterner la simulation des x et y selon les lois conditionnelles $\pi(x|y)$ et $\pi(y|x)$, où on conditionne toujours par la valeur actuelle de l'autre variable.

Algorithme : Échantillonneur de Gibbs ou Two-stage Gibbs sampler

On choisit un point initial $y^{(0)}$.

À l'itération t, la valeur actuelle de y est $y^{(t)}$, et

(i) on génère $x^{(t+1)} \sim \pi(x|y^{(t)})$,

(ii) on génère $y^{(t+1)} \sim \pi(y|x^{(t+1)})$.

Justification

On peut montrer que les deux suites $(x^{(t)})_t$ et $(y^{(t)})_t$ générées par l'échantillonneur de Gibbs sont des chaînes de Markov de loi invariante $\pi(x)$ et $\pi(y)$, respectivement. Par ailleurs, la distribution limite de $(x^{(t)}, y^{(t)})_t$ est $\pi(x, y)$, *si* la chaîne de Markov est irréductible (c'est-à-dire si tous les points du support de $\pi(x, y)$ peuvent être atteints en un nombre fini de pas).

3.4.4 Échantillonneur de Gibbs pour le modèle de mélange

Maintenant, notre loi cible est la loi a posteriori $\pi(\theta|\mathbf{x}) = \prod_{i=1}^{n} \left(\sum_{k=1}^{K} h(x_i, \varphi_k)\pi_k \right) \times \pi(\theta)$ dont on veut approcher les moments, notamment la moyenne a posteriori. Nous allons mettre en œuvre un échantillonneur de Gibbs afin de construire une chaîne de Markov dont la loi invariante est $\pi(\theta|\mathbf{x})$.

Pour le modèle de mélange, il est naturelle d'augmenter les données par les variables latentes $\mathbf{Z} = (Z_1, \ldots, Z_n)$, à savoir les étiquettes ou appartenances de groupes des observations. Les $\mathbf{Z}$ jouent alors le rôle de la variable Y du Paragraphe 3.4.3, et la variable conditionnelle $\theta|\mathbf{x}$ correspond à la variable X du Paragraphe 3.4.3.

Pour l'échantillonneur de Gibbs, il est essentiel de savoir simuler des lois conditionnelles $\pi(\mathbf{z}|\mathbf{x}, \theta)$ et $\pi(\theta|\mathbf{x}, \mathbf{z})$. Or,

$$\begin{aligned}
\pi(\mathbf{z}|\mathbf{x}, \theta) = \frac{\pi(\mathbf{x}, \mathbf{z}, \theta)}{\pi(\mathbf{x}, \theta)} &= \frac{\pi(\mathbf{x}|\mathbf{z}, \theta)\pi(\mathbf{z}|\theta)\pi(\theta)}{\pi(\mathbf{x}, \theta)} \\
&\propto \pi(\mathbf{x}|\mathbf{z}, \theta)\pi(\mathbf{z}|\theta) \\
&= \prod_{i=1}^{n} h(x_i, \varphi_{z_i}) \times \prod_{i=1}^{n} \pi_{z_i} \\
&= \prod_{i=1}^{n} \pi_{z_i} h(x_i, \varphi_{z_i}).
\end{aligned}$$

Comme la loi conditionnelle $\pi(\mathbf{z}|\mathbf{x}, \theta)$ se factorise, les Z_i sont, conditionnellement à $\mathbf{x}$ et

θ, indépendantes avec

$$\pi(z_i|\mathbf{x},\theta) \propto \pi_{z_i} h(x_i, \varphi_{z_i}).$$

Autrement dit, il s'agit d'une loi discrète à valeurs dans $\{1,\dots,K\}$ de probabilités

$$\mathbb{P}(Z_i = k|\mathbf{x},\theta) = \frac{\pi_k h(x_i,\varphi_k)}{\sum_{l=1}^K \pi_l h(x_i,\varphi_l)}, \quad k = 1,\dots,K. \tag{3.7}$$

Nous constatons que cette une loi très simple à simuler. Sous `R` on peut le faire avec la fonction `sample()`.

Quant à la loi conditionnelle $\pi(\theta|\mathbf{x},\mathbf{z})$, on trouve

$$\begin{aligned}\pi(\theta|\mathbf{x},\mathbf{z}) &= \frac{\pi(\mathbf{x},\mathbf{z},\theta)}{\pi(\mathbf{x},\mathbf{z})} = \frac{\pi(\mathbf{x}|\mathbf{z},\theta)\pi(\mathbf{z}|\theta)\pi(\theta)}{\pi(\mathbf{x},\mathbf{z})} \\ &\propto \pi(\mathbf{x}|\mathbf{z},\theta)\pi(\mathbf{z}|\theta)\pi(\theta) \\ &= \pi(\theta)\prod_{i=1}^n \pi_{z_i} h(x_i,\varphi_{z_i}).\end{aligned}$$

On voit que la loi conditionnelle $\pi(\theta|\mathbf{x},\mathbf{z})$ dépend de la loi a priori $\pi(\theta)$ sur le paramètre θ, que l'on doit choisir en sorte que la simulation de la loi conditionnelle $\pi(\theta|\mathbf{x},\mathbf{z})$ est faisable. Il semble naturel d'utiliser une loi a priori factorisée pour les deux parties $\boldsymbol{\pi}$ et $\boldsymbol{\varphi}$ du paramètre $\theta = (\boldsymbol{\pi},\boldsymbol{\varphi}) = (\pi_1,\dots,\pi_K,\varphi_1,\dots,\varphi_K)$. Ainsi, la loi conditionnelle $\pi(\theta|\mathbf{x},\mathbf{z})$ se factorise également, ce qui est toujours avantageux en vue de la simulation. Plus précisément, si $\pi(\theta) = \pi(\boldsymbol{\pi})\pi(\boldsymbol{\varphi})$, on obtient

$$\pi(\theta|\mathbf{x},\mathbf{z}) \propto \left(\pi(\boldsymbol{\pi})\prod_{i=1}^n \pi_{z_i}\right) \times \left(\pi(\boldsymbol{\varphi})\prod_{i=1}^n h(x_i,\varphi_{z_i})\right).$$

Autrement dit,

$$\pi(\theta|\mathbf{x},\mathbf{z}) = \pi(\boldsymbol{\pi}|\mathbf{x},\mathbf{z})\pi(\boldsymbol{\varphi}|\mathbf{x},\mathbf{z})$$

avec

$$\pi(\boldsymbol{\pi}|\mathbf{x},\mathbf{z}) \propto \pi(\boldsymbol{\pi})\prod_{i=1}^n \pi_{z_i} \quad \text{et} \quad \pi(\boldsymbol{\varphi}|\mathbf{x},\mathbf{z}) \propto \pi(\boldsymbol{\varphi})\prod_{i=1}^n h(x_i,\varphi_{z_i}).$$

Loi a priori $\pi(\boldsymbol{\pi})$ sur $\boldsymbol{\pi}$

Notons que le paramètre $\boldsymbol{\pi} = (\pi_1,\dots,\pi_K)$ appartient au $K-1$–simplex

$$\mathcal{S}_{K-1} = \left\{(p_1,\dots,p_K) \in [0,1]^K : \sum_{k=1}^K p_k = 1\right\}.$$

Nous utilisons comme loi a priori $\pi(\boldsymbol{\pi})$ la **loi de Dirichlet** $\mathcal{D}(\gamma_1,\dots,\gamma_K)$, qui est une loi continue à valeurs dans $\mathcal{S}$. La densité $\pi(\boldsymbol{\pi})$ est alors donnée par

$$\pi(\boldsymbol{\pi}) = \frac{\Gamma\left(\sum_{k=1}^K \gamma_k\right)}{\prod_{k=1}^K \gamma_k} \prod_{k=1}^K \pi_k^{\gamma_k - 1}.$$

La densité de la loi de Dirichlet est illustrée dans la Figure 3.8. Elle est unimodale si $\gamma_k > 1$ pour tout $k = 1,\dots,K$, et plus ou moins symétrique et pointue (concentrée) selon le choix

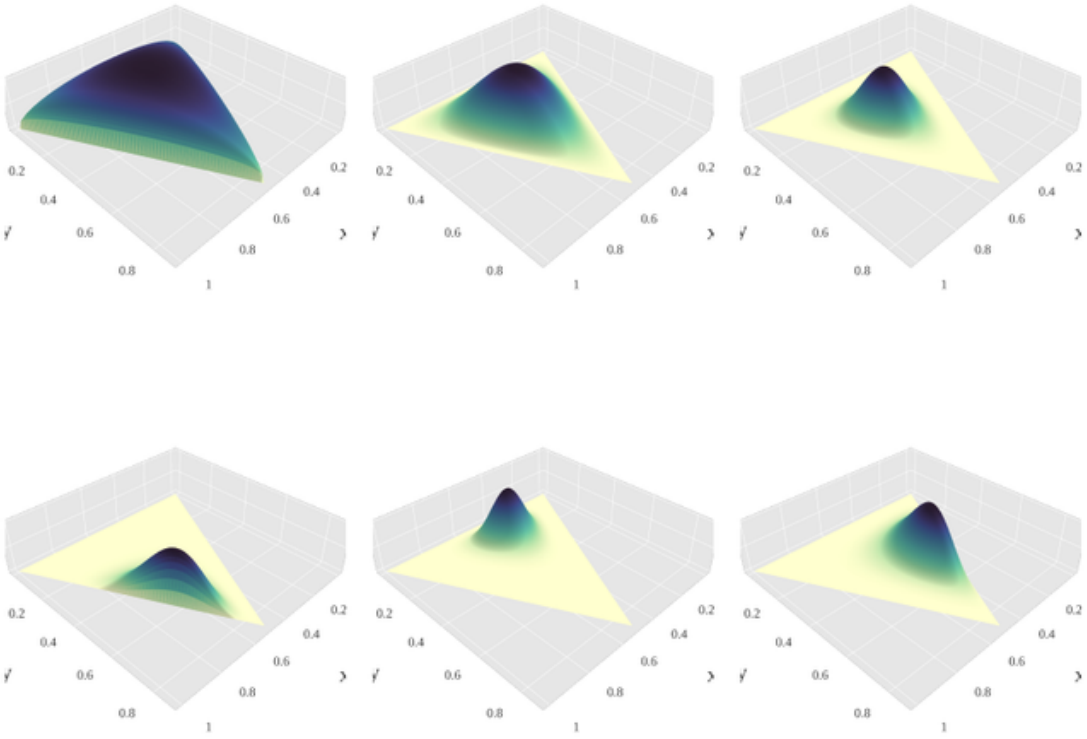


FIGURE 3.8 – Densité de la loi de Dirichlet sur le simplex $\mathcal{S}_2$ avec les paramètres $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \gamma_3)$ suivants : première ligne de gauche à droite : $(1.3, 1.3, 1.3)$, $(3, 3, 3)$, $(7, 7, 7)$, deuxième ligne de gauche à droite : $(6, 2, 6)$, $(14, 9, 5)$, $(2, 6, 11)$. Source : Wikipedia, Dirichlet distribution.

des paramètres. Remarquons quand $\gamma_k = 1$ pour tout $k = 1, \ldots, K$, la loi de Dirichlet est la loi uniforme sur le simplex $\mathcal{S}_{K-1}$.

Pour la loi conditionnelle $\pi(\boldsymbol{\pi}|\mathbf{x}, \mathbf{z})$ on obtient

$$\begin{aligned}\pi(\boldsymbol{\pi}|\mathbf{x}, \mathbf{z}) &\propto \pi(\boldsymbol{\pi}) \prod_{i=1}^{n} \pi_{z_i} \\ &\propto \prod_{k=1}^{K} \pi_k^{\gamma_k - 1} \times \prod_{i=1}^{n} \pi_{z_i} \\ &= \prod_{k=1}^{K} \pi_k^{\sum_{i=1}^{n} \mathbb{1}\{z_i = k\} + \gamma_k - 1}.\end{aligned}$$

On observe que la loi conditionnelle $\pi(\boldsymbol{\pi}|\mathbf{x}, \mathbf{z})$ est la loi de Dirichlet sur le simplex $\mathcal{S}_{K-1}$ de paramètres $(\gamma_1 + n_1, \ldots, \gamma_K + n_K)$, où n_k désigne le nombre de variables latentes à valeur k défini comme $n_k = \#\{i : z_i = k\} = \sum_{i=1}^{n} \mathbb{1}\{z_i = k\}$ pour $k = 1, \ldots, K$. Autrement dit, pour simuler la loi conditionnelle $\pi(\boldsymbol{\pi}|\mathbf{x}, \mathbf{z})$, il faut générer des réalisations de la loi de Dirichlet

$$\mathcal{D}(\gamma_1 + n_1, \ldots, \gamma_K + n_K),$$

ce qui est facile sous `R` grâce à la fonction `rdirichlet` du package `gtools`.

Loi a priori $\pi(\boldsymbol{\varphi})$ sur $\boldsymbol{\varphi}$

Pour $\{h(\cdot, \varphi), \varphi \in \Phi\}$ nous considérons une **famille exponentielle**. Si $\Phi \subset \mathbb{R}$, cela veut dire qu'il existe des fonctions g, R et Ψ telles que les densités $h(\cdot, \varphi)$ s'écrivent sous la forme

$$h(x, \varphi) = g(x) \exp\{\varphi R(x) - \Psi(\varphi)\}, \quad \forall x \in \mathbb{R}, \forall \varphi \in \Phi.$$

La plupart de familles de lois usuelles comme la loi gaussienne, exponentielle, Gamma, Beta, khi-deux, Poisson ou Bernoulli sont des familles exponentielles.

Dans ce cas, un bon choix de la loi a priori sur le paramètres $\boldsymbol{\varphi} = (\varphi_1, \ldots, \varphi_K)$ d'un mélange est l'utilisation de lois a priori indépendantes de la forme

$$\pi(\boldsymbol{\varphi}) = \prod_{k=1}^{K} \pi(\varphi_k) \quad \text{avec} \quad \pi(\varphi_k) \propto \mathrm{e}^{\varphi_k \xi_k - \lambda_k \Psi(\varphi_k)}, \tag{3.8}$$

où ξ_k et λ_k sont les hyperparamètres de la loi a priori. Pour la loi conditionnelle $\pi(\boldsymbol{\varphi}|\mathbf{x}, \mathbf{z})$ on obtient

$$\begin{aligned}\pi(\boldsymbol{\varphi}|\mathbf{x}, \mathbf{z}) &\propto \pi(\boldsymbol{\varphi}) \prod_{i=1}^{n} h(x_i, \varphi_{z_i}) \\ &\propto \prod_{k=1}^{K} \pi(\varphi_k) \times \prod_{k=1}^{K} \prod_{i:z_i=k} h(x_i, \varphi_k) \\ &= \prod_{k=1}^{K} \underbrace{\pi(\varphi_k) \prod_{i:z_i=k} h(x_i, \varphi_k)}_{\propto \pi(\varphi_k|\mathbf{x}, \mathbf{z})}.\end{aligned}$$

On voit que les φ_k sont indépendants conditionnellement aux $\mathbf{x}$ et $\mathbf{z}$. Or,

$$\begin{aligned}\pi(\varphi_k|\mathbf{x},\mathbf{z}) &\propto \pi(\varphi_k)\prod_{i:z_i=k} h(x_i,\varphi_k)\\ &\propto \mathrm{e}^{\varphi_k\xi_k-\lambda_k\Psi(\varphi_k)}\prod_{i:z_i=k}\mathrm{e}^{\varphi_k R(x_i)-\Psi(\varphi_k)}\\ &= \exp\left\{\varphi_k\left(\xi_k+\sum_{i:z_i=k}R(x_i)\right)-\Psi(\varphi_k)\left(\lambda_k+n_k\right)\right\},\end{aligned} \tag{3.9}$$

où $n_k=\sum_{i:z_i=k}\mathbb{1}\{z_i=k\}$.

Pour aller plus loin, il faut choisir un cadre précis. Voyons dans le cas d'un mélange gaussien à variances connues.

Exemple. Mélange gaussien (à variances connues)

Considérons le mélange gaussien de densité

$$\pi(x|\theta)=\sum_{k=1}^{K}\pi_K f_{\mathcal{N}(\mu_k,1)}(x),$$

avec paramètre $\theta=(\pi_1,\ldots,\pi_K,\mu_1,\ldots,\mu_K)$. Montrons d'abord que la famille des lois normales $\mathcal{N}(\mu,1)$ forme une famille exponentielle. On a

$$\begin{aligned}h(x|\mu)=f_{\mathcal{N}(\mu,1)}(x)&=\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}(x-\mu)^2\right\}\\ &=\underbrace{\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-x^2/2}}_{=g(x)}\exp\left\{\underbrace{x}_{=R(x)}\mu-\underbrace{\frac{\mu^2}{2}}_{=\Psi(\mu)}\right\}.\end{aligned}$$

D'après (3.8), on choisit la loi a priori $\pi(\mu_k)$ de μ_k comme

$$\begin{aligned}\pi(\mu_k)&\propto\exp\left\{\mu_k\xi_k-\lambda_k\frac{\mu^2}{2}\right\}\\ &=\exp\left\{-\frac{\lambda_k}{2}\left(\mu_k^2-\frac{2\mu_k\xi_k}{\lambda_k}\right)\right\}\\ &\propto\exp\left\{-\frac{\lambda_k}{2}\left(\mu_k-\frac{\xi_k}{\lambda_k}\right)^2\right\}\\ &\propto f_{\mathcal{N}(\xi_k/\lambda_k,1/\lambda_k)}(\mu_k).\end{aligned}$$

Par (3.9), la loi conditionnelle $\pi(\mu_k|\mathbf{x},\mathbf{z})$ est de la forme

$$\begin{aligned}\pi(\mu_k|\mathbf{x},\mathbf{z})&\propto\exp\left\{\mu_k\left(\xi_k+\sum_{i:z_i=k}x_i\right)-\frac{\mu_k^2}{2}\left(\lambda_k+n_k\right)\right\}\\ &=\exp\left\{-\frac{\lambda_k+n_k}{2}\left[\mu_k^2-\frac{2\mu_k}{\lambda_k+n_k}\left(\xi_k+\sum_{i:z_i=k}x_i\right)\right]\right\}\\ &\propto\exp\left\{-\frac{\lambda_k+n_k}{2}\left[\mu_k-\frac{1}{\lambda_k+n_k}\left(\xi_k+\sum_{i:z_i=k}x_i\right)\right]^2\right\},\end{aligned}$$

où on reconnait la densité normale

$$\mathcal{N}\left(\frac{\xi_k + \sum_{i:z_i=k} x_i)}{\lambda_k + n_k}, \frac{1}{\lambda_k + n_k}\right),$$

qui est une loi que l'on sait simuler facilement.

Pour résumer, pour l'échantillonneur de Gibbs pour un mélange gaussien à variances connues, nous utilisons les lois a priori

$$\boldsymbol{\pi} \sim \mathcal{D}(\gamma_1, \ldots, \gamma_K) \quad \text{et} \quad \mu_k \sim \mathcal{N}\left(\frac{\xi_k}{\lambda_k}, \frac{1}{\lambda_k}\right), \quad k = 1, \ldots, K,$$

où les $\mu_1, \ldots, \mu_K$ sont mutuellement indépendants. Les hyperparamètres $\gamma_k, \xi_k, \lambda_k$ pour $k = 1, \ldots, K$ sont à choisir par l'utilisateur.

L'algorithme a besoin d'un point initial $\theta^{(0)} = (\boldsymbol{\pi}^{(0)}, \mu_1^{(0)}, \ldots, \mu_K^{(0)})$.

À l'itération t, la valeur actuelle de θ est noté $\theta^{(t-1)}$, et on procède comme suit :

(i) Pour $i = 1, \ldots, n$ générer des $z_i^{(t)}$: Pour rappel, d'après (3.7) les $z_i^{(t)}$ prennent leurs valeurs dans $\{1, \ldots, K\}$ avec probabilité

$$\mathbb{P}(Z_i^{(t)} = k | \mathbf{x}, \theta^{(t-1)}) = \frac{\pi_k^{(t-1)} h(x_i, \varphi_k^{(t-1)})}{\sum_{l=1}^K \pi_l^{(t-1)} h(x_i, \varphi_l^{(t-1)})}, \quad k = 1, \ldots, K.$$

(ii) Générer $\boldsymbol{\pi}^{(t)}$ selon la loi de Dirichlet

$$\mathcal{D}\left(\gamma_1^{(t)} + n_1^{(t)}, \ldots, \gamma_K^{(t)} + n_K^{(t)}\right),$$

où $n_k^{(t)} = \sum_{i=1}^n \mathbb{1}\{z_i^{(t)} = k\}$ pour $k = 1, \ldots, K$.

(iii) Pour $k = 1, \ldots, K$ générer

$$\mu_k^{(t)} \sim \mathcal{N}\left(\frac{\xi_k + \sum_{i:z_i^{(t)}=k} x_i}{\lambda_k + n_k^{(t)}}, \frac{1}{\lambda_k + n_k^{(t)}}\right).$$

En sortie, cet algorithme renvoie la suite

$$\left(\mathbf{z}^{(t)}; \theta^{(t)}\right)_{t\geq 1} = \left(z_1^{(t)}, \ldots, z_n^{(t)}; \boldsymbol{\pi}^{(t)}; \mu_1^{(t)}, \ldots, \mu_K^{(t)}\right)_{t\geq 1},$$

qui est par construction une chaîne de Markov de loi invariante $\pi(\mathbf{z}, \theta|\mathbf{x})$. La partie $(\theta^{(t)})_{t\geq 1}$ est une chaîne de Markov de loi invariante $\pi(\theta|\mathbf{x})$ et peut être utilisé pour estimer la moyenne a posteriori comme estimateur de θ. L'autre partie $(\mathbf{z}^{(t)})_{t\geq 1}$ est une chaîne de Markov de loi invariante $\pi(\mathbf{z}|\mathbf{x}, \theta)$ et on peut l'utiliser pour estimer l'appartenance de groupe de chacune des n observations.

3.4.5 Aspects de mise en œuvre

Initialisation

D'un point de vue théorique, on peut initialiser l'algorithme comme on veut, car si la chaîne de Markov est irréductible, tout l'espace sera exploré. Néanmoins, en pratique, on ne prend jamais en compte les premières itérations afin d'"oublier" le point initial, c'est le *burn-in time*. Autrement dit, on jette par exemple les 500 premières itérations.

Critères d'arrêt

Normalement, il faut que l'algorithme explore suffisamment bien tout l'espace pour avoir une bonne approximation de la loi a posteriori. En pratique, il est impossible de savoir d'avance quand ce sera le cas. Il y a plusieurs outils graphiques qui permettent d'analyser la convergence de l'algorithme (cf. TP).

Problème du label switching

L'échangeabilité des composantes de mélange a des conséquences problématiques. On dit que les lois a priori $\pi(\boldsymbol{\pi})$ et $\pi(\boldsymbol{\varphi})$ sont **échangeables** si pour toute permutation τ des valeurs $\{1, 2, \ldots, K\}$, la loi de $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$ est la même que celle de $(\pi_{\tau(1)}, \ldots, \pi_{\tau(K)})$, et, de même, $\pi(\varphi_1, \ldots, \varphi_K) = \pi(\varphi_{\tau(1)}, \ldots, \varphi_{\tau(K)})$.

Proposition 3. *(i) Soit $\pi(\boldsymbol{\pi})$ une loi a priori échangeable. Alors, les lois marginales $\pi(\pi_k)$ pour $k = 1, \ldots, K$ sont les mêmes, et $\mathbb{E}_{\boldsymbol{\pi}}[\pi_k] = \frac{1}{K}$ pour $k = 1, \ldots, K$.*

(ii) Si, de plus, $\pi(\boldsymbol{\varphi}) = \pi\left(\prod_{k=1}^{K} \pi(\varphi_k)\right)$ est une loi a priori échangeable, alors
- *les lois marginales conditionnelles $\pi(\pi_k|\mathbf{x})$ pour $k = 1, \ldots, K$ sont les mêmes, et $\mathbb{E}_{\boldsymbol{\pi}}[\pi_k|\mathbf{x}] = \frac{1}{K}$ pour $k = 1, \ldots, K$.*
- *les lois marginales conditionnelles $\pi(\varphi_k|\mathbf{x})$ pour $k = 1, \ldots, K$ sont les mêmes.*

Démonstration. En utilisant l'échangeabilité de $\pi(\boldsymbol{\pi})$, on trouve

$$\begin{aligned}\pi(\pi_1) &= \underbrace{\int \ldots \int}_{K-1 \text{ integrals}} \pi(\pi_1, \pi_2, \ldots, \pi_K) \mathrm{d}\pi_2 \ldots \pi_K \\ &= \int \ldots \int \pi(\pi_{\tau(1)}, \pi_{\tau(2)}, \ldots, \pi_{\tau(K)}) \mathrm{d}\pi_2 \ldots \pi_K \\ &= \pi(\pi_{\tau(1)}),\end{aligned}$$

pour toute permutation τ des valeurs de $\{1, 2, \ldots, K\}$. Cela implique que les lois $\pi(\pi_k), k = 1, \ldots, K$ sont toutes les mêmes.

Or, on a $\sum_{k=1}^{K} \pi_k = 1$ *p.s.*, ce qui entraîne que

$$1 = \mathbb{E}\left[\sum_{k=1}^{K} \pi_k\right] = \sum_{k=1}^{K} \mathbb{E}[\pi_k] = K\mathbb{E}[\pi_1].$$

D'où $\mathbb{E}[\pi_1] = \mathbb{E}[\pi_k] = 1/K$.

Pour la loi a posteriori on trouve,

$$\begin{aligned}\pi(\pi_1|\mathbf{x}) &\propto \pi(\pi_1, \mathbf{x}) \\ &= \underbrace{\int \ldots \int}_{2K-1 \text{ integrals}} \pi(\pi_1, \pi_2, \ldots, \pi_K, \varphi_1, \ldots, \varphi_K, \mathbf{x}) \mathrm{d}\pi_2 \ldots \pi_K \mathrm{d}\varphi_1 \ldots \mathrm{d}\varphi_K \\ &= \int \ldots \int \pi(\mathbf{x}|\theta)\pi(\boldsymbol{\pi})\pi(\boldsymbol{\varphi}) \mathrm{d}\pi_2 \ldots \pi_K \mathrm{d}\varphi_1 \ldots \mathrm{d}\varphi_K.\end{aligned}$$

Or, pour toute permutation τ on a

$$\pi(\mathbf{x}|\theta) = \sum_{k=1}^{K} \pi_k h(x_i, \varphi_k) = \sum_{k=1}^{K} \pi_{\tau(k)} h(x_i, \varphi_{\tau(k)}).$$

D'où

$$\begin{aligned}\pi(\pi_1|\mathbf{x}) &\propto \int\dots\int \sum_{k=1}^{K} \pi_{\tau(k)} h(x_i, \varphi_{\tau(k)})\pi(\boldsymbol{\pi})\pi(\boldsymbol{\varphi})\mathrm{d}\pi_2\dots\pi_K\mathrm{d}\varphi_1\dots\mathrm{d}\varphi_K \\ &= \int\dots\int \sum_{k=1}^{K} \pi_{\tau(k)} h(x_i, \varphi_{\tau(k)})\pi(\underbrace{\pi_{\tau(1)},\dots,_{\tau(K)}}_{=:\boldsymbol{\pi}_\tau})\pi(\underbrace{\varphi_{\tau(1)},\dots,\varphi_{\tau(K)}}_{=:\boldsymbol{\varphi}_\tau})\mathrm{d}\pi_2\dots\pi_K\mathrm{d}\varphi_1\dots\mathrm{d}\varphi_K \\ &\propto \pi(\pi_{\tau(1)}|\mathbf{x}),\end{aligned}$$

par échangeabilité des lois a priori.

Quant à la loi conditionnelle de φ_1, on trouve de la même façon que

$$\begin{aligned}\pi(\varphi_1|\mathbf{x}) &\propto \pi(\varphi_1, \mathbf{x}) \\ &= \int\dots\int \pi(\theta, \mathbf{x})\mathrm{d}\pi_1\dots\pi_K\mathrm{d}\varphi_2\dots\mathrm{d}\varphi_K \\ &= \int\dots\int \pi(\mathbf{x}|\theta)\pi(\boldsymbol{\pi})\pi(\boldsymbol{\varphi})\mathrm{d}\pi_1\dots\pi_K\mathrm{d}\varphi_2\dots\mathrm{d}\varphi_K \\ &= \int\dots\int \pi(\mathbf{x}|\theta_\tau)\pi(\boldsymbol{\pi}_\tau)\pi(\boldsymbol{\varphi}_\tau)\mathrm{d}\pi_1\dots\pi_K\mathrm{d}\varphi_2\dots\mathrm{d}\varphi_K \\ &\propto \pi(\varphi_{\tau(1)}|\mathbf{x}).\end{aligned}$$

□

La conséquence de cette proposition est que les estimateurs bayésiens, en particulier les moyennes a posteriori, sont les mêmes !! Par exemple, on a $\mathbb{E}_{\boldsymbol{\pi}}[\pi_k|\mathbf{x}] = \frac{1}{K}$ pour $k = 1,\dots,K$. Cela n'a pas de sens du point de vue de l'estimation.

Afin de remédier à ce problème, on pourrait avoir l'idée d'introduire des contraintes d'identifiabilité (du type $\varphi_1 < \varphi_2 < \dots < \varphi_K$). De telles contraintes de troncature ont un impact sur les lois a priori et a posteriori, mais pas forcément dans le bon sens. Il est généralement déconseillé de le faire.

Une meilleure solution consiste à faire tourner l'échantillionneur de Gibbs comme décrit ci-dessus, et c'est *après* que l'on ordonne les valeurs des $(\pi_k^{(t)}, \varphi_k^{(t)})$ et des $z_i^{(t)}$ pour tout t, par exemple selon une condition du type $\varphi_1^{(t)} < \varphi_2^{(t)} < \dots < \varphi_K^{(t)}$. Autrement dit, la première composante est toujours la composante avec le plus petit paramètre $\varphi_k^{(t)}$, et on permute les $\pi_k^{(t)}$ et les $z_i^{(t)}$ dans le même sens.

En revanche, assez souvent en pratique, l'échantillonneur de Gibbs n'explore pas tout l'espace, mais juste *un* mode, et donc le problème ne se produit pas. En effet, dans le cas d'un modèle de mélange d'ordre K, la densité a posteriori $\pi(\theta|\mathbf{x})$ est multimodale. Par échangeabilité des composantes, $\pi(\theta|\mathbf{x})$ a au moins $K!$ maximums locaux. En pratique, l'échantillonneur de Gibbs ne parvient pas à explorer tous les modes, car pour passer de l'un mode à l'autre, il faut "traverser une vallée" ce qui est difficile pour l'algorithme. En général, il reste coincé dans le mode le plus près du point initial. En quelque sorte, c'est le même type de problème que l'on a vu pour l'algorithme EM. Au fait, pour passer à un autre mode, il faut que beaucoup de variables $z_i^{(t+1)}$ changent de valeur d'une itération à l'autre, ce qui n'arrive que rarement.

3.5 Comparaison de Gibbs et EM pour mélanges

Tout d'abord, on remarque que les deux algorithmes s'inscrivent dans des approches statistiques assez différentes, l'EM dans une approche statistique fréquentiste où il est question d'approcher l'estimateur du maximum de vraisemblance, alors que l'échantillonneur de Gibbs est une méthode de la statistique bayésienne. Plus précisément, l'algorithme EM est une méthode numérique pour résoudre un problème d'optimisation, alors que l'échantillonneur de Gibbs a pour objectif de générer une chaîne de Markov de la loi a posteriori.

Malgré ses différences fondamentales, les deux algorithmes ont beaucoup de points en commun. Tous les deux sont des méthodes itératives. En plus, ils dépendent de l'initialisation car les deux méthodes n'arrivent pas à explorer l'espace entier ce qui est dû au problème d'échangeabilité des composantes du mélange. Par ailleurs, tous les deux se servent du même astuce : utiliser le concept des variables latentes pour transformer un problème très difficile (calcul de l'EMV pour l'un, calcul de la loi a priori pour l'autre) en un problème soluble. En effet, au final, la structure des deux algorithmes est très similaire :

L'itération t consiste à

(i) estimer les variables latentes Z_i. Plus précisément,
   — pour EM : estimer les probabilités conditionnelles $\mathbb{P}_{\theta^{(t)}}(Z_i = k|\mathbf{x})$.
   — pour Gibbs : générer des $z_i^{(t+1)}$ selon la loi conditionnelle actuelle $\pi(z_i|\mathbf{x}, \theta^{(t)})$.

(ii) mettre à jour le paramètre θ. Plus précisément,
   — pour EM : en maximisant l'espérance conditionnelle $\theta \mapsto \mathbb{E}_{\theta^{(t)}}[\log q_\theta(\mathbf{x}, \mathbf{Z})|\mathbf{x}]$ de la log-vraisemblance des données complètes.
   — Gibbs : générer $\theta^{(t+1)}$ selon la loi conditionnelle $\pi(\theta|\mathbf{x}, \mathbf{z}^{(t+1)})$.

Quelque part le principe des deux algorithmes est le même, seulement l'algorithme EM est un procédé déterministe et Gibbs est une version aléatoire d'une même idée d'algorithme. Finalement, on peut observer que le comportement et la performance des deux méthodes sont assez similaires.

Bibliographie

Davison, A. C. and D. V. Hinkley (1997). *Bootstrap Methods and Their Application*. Cambridge University Press.

Droesbeke, J., G. Saporta, and C. Thomas-Agnan (2013). *Modèles à variables latentes et modèles de mélange*. Editions Technip.

Efron, B. and R. J. Tibshirani (1993). *An Introduction to the Bootstrap*. Monographs on Statistics and Applied Probability. Chapman & Hall/CRC.

Guyader, A. (2017). Statistique – Partie 1. Notes de cours, disponible sur le site `http://www.lsta.upmc.fr/guyader/statM1.html`.

Lafaye de Micheaux, P., R. Drouilhet, and B. Liquet (2010). *Le logiciel R : Maîtriser le langage - Effectuer des analyses statistiques*. Statistique et probabilités appliquées. Springer Paris.

Lejeune, M. (2004). *Statistique : la théorie et ses applications*. Collection Statistique et probabilités appliquées. Springer.

Marin, J. and C. Robert (2007). *Bayesian Core : A Practical Approach to Computational Bayesian Statistics*. Springer Texts in Statistics. Springer New York.

McLachlan, G. and T. Krishnan (2008). *The EM algorithm and extensions* (2. ed ed.). Wiley series in probability and statistics. Wiley.

McLachlan, G. and D. Peel (2000). *Finite Mixture Models*. Wiley Series in Probability and Statistics. Wiley-Interscience.

Rebafka, T. (2017). Statistique – Partie 2. Notes de cours, disponible sur le site `https://www.lpsm.paris//pageperso/rebafka`.

Rivoirard, V. and G. Stoltz (2012). *Statistique mathématique en action : cours, problèmes d'application corrigés et mises en action concrètes*. Mathématiques concrètes. Vuibert.

Robert, C. P. and G. Casella (2004). *Monte Carlo Statistical Methods*. New York : Springer.