Outils statistiques S3
Université de Strasbourg
Master Statistiques

Davide Giraudo

17 octobre 2023

Table des matières

1 Bootstrap

1.1 Principe

L'estimation statistique consiste à vouloir déterminer la valeur d'une statistique d'intérêt (moyenne, variance, quantile...) pour une loi inconnue. Les quantités mathématiques en jeu sont les suivantes :

— une loi inconnue P. On note $\mathbf{X} = (X_1, \ldots, X_n)$ le vecteur aléatoire contenant n tirages i.i.d. selon la loi P.

— une statistique d'intérêt θ qui dépend de P et que l'on chercher à estimer.

— une statistique $T(\mathbb{X})$ qui va servir à estimer θ et que l'on appelle estimateur de θ.

La seule connaissance que l'on peut extraire directement des données est un échantillon, c'est-à-dire une réalisation $x_1, \ldots, x_n$ de $\mathbb{X}$.

La valeur prise par T en l'échantillon recueilli est appelé estimation de θ et noté $\widehat{\theta}$:

$$\widehat{\theta} = T(x_1, \ldots, x_n). \tag{1.1.1}$$

Cette valeur n'est cependant qu'une approximation de la vraie valeur d'intérêt θ. Pour avoir une idée de l'erreur commise (via des intervalles de confiance) ou pour comparer des estimateurs entre eux, il faut être capable d'étudier les propriétés de T : son biais, sa variance, l'erreur quadratique moyenne, sa fonction de répartition, etc. Cependant, comme T dépend de P qui est inconnue, il est impossible d'accéder à sa loi de façon exacte. Il est par conséquent nécessaire de recourir à des hypothèses ou des approximations.

1.1.1 Option 1 : Hypothèses sur la forme de la loi

Une manière de contourner le problème peut être de faire une hypothèse sur la forme de la loi P. Une approche courante consiste à suppose qu'elle fait partie d'un famille paramétrée (par exemple, loi normale, exponentielle, de Poisson), et d'estimer les paramètres nécessaires (θ en fait souvent partie en pratique).

Exemple 1.1. Supposons qu'on cherche à estimer l'espérance d'une loi inconnue via l'estimateur de la moyenne, et qu'il est raisonnable de supposer que la variable d'intérêt suit une loi normale. La loi des grands nombres dit alors que l'estimateur est consistant et le théorème central limite dit que $\sqrt{n}\left(\widehat{\theta} - \theta\right)$ suit asymptotiquement une loi normale, ce qui permet d'accéder à la vitesse de convergence et des intervalles de confiance.

Dans ce cas, qui est la manière historique de faire de l'estimation, une approximation est faite via l'hypothèse d'appartenance de la loi inconnue à une famille paramétrée. Une fois cette hypothèse faite, on peut chercher à déterminer théoriquement les propriétés de l'estimateur dans ce cadre.

Cette approache a cependant deux limitations :

1. La crédibilité de l'hypothèse de forme (elle peut être levée dans certains cas, par exemple pour estimer l'espérance puisque le TCL est valable pour toute forme de la vraie loi).
2. La quantité à estimer : si on cherche à estimer la médiane plutôt que la moyenne ou des quantités dépendant de P de façon complexe, l'étude de leurs lois peut se révéler difficile voire impossible, même sous une hypothèse de forme.

1.1.2 Option 2 : Accès illimité (ou presque) aux données

Une manière différente de résoudre le problème serait d'avoir un très grand nombre d'estimations indépendantes, en d'autres termes un échantillon d'estimations $\left(\widehat{\theta}_1, \ldots, \widehat{\theta}_B\right)$. Une approximation possible de la loi de $T(\mathbb{X})$ est alors la loi empirique d'un tel échantillon quand B est grand.

Limitations :

1. Cette approche est en général impossible tout simplement en raison du nombre limité de données. En effet, il faudrait ici $n \times B$ mesures i.i.d. selon la loi inconnue, avec n et B les plus grands possibles.
2. Peut-on borner l'erreur faite en fonction de B et n ?

1.1.3 Option 3 : Bootstrap

L'explosion en termes de capacité de calcul ayant eu lieu les dernières décennies permet de faire un nombre très important de tirages i.i.d. selon toute loi simulable. Cela a ouvert la voie à une approche appelée Bootstrap qui se déroule en deux étapes :

1. On approxime la loi inconnue P par la loi empirique P_n de l'échantillon.
2. On génère un « grand »nombre B d'échantillons suivant P_n, appelés échantillons bootstrap. On obtient autant d'estimations de la grandeur d'intérêt, et leur loi empirique est utilisée pour approximer la loi de T.

En d'autres termes, on applique l'idée de l'option 2 en créant les données par simulation. Le prix à payer est qu'on simule suivant la loi P_n plutôt que la loi P.

Avantages :

1. on peut traiter n'importe quelle loi
2. on peut considérer n'importe quel estimateur, quelque soit sa complexité

Limitations :

1. Cette approche cumule deux approximations successives : remplacer P par P_n puis remplacer la loi de T sous P_n par la loi empirique de B tirages.
2. Peut-on borner l'erreur faite en fonction de B et n ?

1.2 Echantillon bootstrap

1.2.1 Produire des échantillons bootstrap

La procédure classique, intoduite par Efron [3] correspond à celle décrite dans le chapitre précédent, à savoir qu'un grand nombre B d'échantillons sont tirés suivant la loi empirique de l'échantillon $\mathbb{X} = (X_1, \ldots, X_n)$. Or tirer suivant cette loi revient à tirer uniformément un élément de l'échantillon. Le principe du bootstrap peut donc s'écrire simplement :

pour i de 1 à B
tirer n fois avec remise dans $\mathbb{X}$ pour obtenir un échantillon bootstrap $X^{*i} = (X_1^{*i}, \ldots, X_n^{*i})$
obtenir une estimation bootstrap $\widehat{\theta^{*i}} = T\left(X^{*i}\right)$.

Remarque 1.2. En pratique, l'échantillon $\mathbb{X}$ est une observation x, et ne comporte plus de caractère aléatoire. Le processus de tirage avec remise en réintroduit, si bien que les $\widehat{\theta^{*i}}$ sont différents les uns des autres et vont permettre d'appréhender la variabilité de la situation.

Exemple 1.3. On s'intéresse au salaire médian des étudiants un an après la validation de leur diplôme, en se basant sur les douze personnes ayant répondu au questionnaire

```
x = c(2200,1600,2400,1800,900,1600,3300,2100,1700,1200,2000,1600)
median(x)
## [1] 1750
```

Un échantillon bootstrap est obtenu en effectuant des tirages avec remise

```
xboot = sample(x,length(x),replace=TRUE)
xboot
##
[1] 1600 1600 2400 2200 1600 1700 900 1600 2100 2100 2400 1800
```

et il est aisé de produire ainsi un grand nombre d'échantillons et de déterminer les estimations correspondantes

```
median_boot <- c()
for (i in 1:1000){
xboot = sample(x,length(x),replace=TRUE)
median_boot <- c(median_boot,median(xboot))
}
boxplot(median_boot)
```

Variante : Le bootstrap paramétrique : Dans le cas où il semble raisonnable de supposer une forme de loi particulière pour P, il est possible d'incorporer cette hypothèse dans le processus du bootstrap. Dans ce cas, on tire les échantillons bootstrap suivant la loi paramétrique de paramètre θ.

Exemple 1.4. On considère le nombre annuel de morts par accident de la route de 2010 à 2019. Sous l'hypothèse (discutable) que l'échantillon est i.i.d., on peut supposer qu'il suit une loi de Poisson (modélisation des évènements rares). L'estimateur $\widehat{\lambda}$ du paramètre de la loi est alors simplement celui de la moyenne.

```
x=c(3994,3963,3653,3250,3384,3464,3477,3448,3488,3498)
lambda = mean(x)
lambda
## [1] 3561.9
```

Tirer un échantillon bootstrap revient alors à tirer un échantillon de même taille suivant la loi de Poisson de paramètre $\widehat{\lambda}$.

```
median_boot <- c()
for (i in 1:1000){
xboot = rpois(length(x),lambda=lambda)
median_boot <- c(median_boot,median(xboot))
}
boxplot(median_boot)
```

1.2.2 Estimateurs bootstrap

On s'intéresse à un paramètre inconnu θ dans un monde réel inaccessible. Le principe du bootstrap est de remplacer chaque quantité d'intérêt par son équivalent dans le monde bootstrap, où tout est connu ou approximable.

Monde réel	Monde du bootstrap
loi P inconnue	loi P_n connue (première approximation)
échantillon $X = (X_1, \ldots, X_n)$	échantillons $\mathcal{X}^* = \left(\mathcal{X}^*_{b,1}, \ldots, \mathcal{X}^*_{b,n}\right)$
estimateur $\widehat{\theta} = T(X)$	estimateurs $\widehat{\theta}^*_b = T(\mathcal{X}^*_b)$
loi de $\widehat{\theta} - \theta$ inconnue en absence d'hypothèses	loi de $\widehat{\theta}^* - \widehat{\theta}$ approximable car $\widehat{\theta}$ est connu et on peut approximer la loi de $\widehat{\theta}^*$ par la loi empirique de $\left(\widehat{\theta}^*_b\right)_{1 \leqslant b \leqslant B}$

La liste ci-dessous donne les exemples les plus courants d'estimations bootstrap de propriétés liées à un estimateur, mais n'est pas exhaustive.

1. La loi de $\widehat{\theta}$
   - Monde réel : on veut estimer $G(t) = \mathbb{P}\left(\widehat{\theta} \leqslant t\right)$.
   - Monde bootstrap :
     - première approximation : on approxime $G(t)$ par $G_n^*(t) = \mathbb{P}\left(\widehat{\theta^*} \leqslant t\right)$.
     - deuxième approximation : on approxime $G_n^*(t)$ par la loi empirique

$$G_{n,B}(t) = \frac{1}{B}\sum_{b=1}^{N} \mathbf{1}_{\widehat{\theta_b^*} \leqslant t}. \tag{1.2.1}$$

2. Estimation du biais de $\widehat{\theta}$
   - Monde réel : on veut estimer $\mathbb{E}_P\left(\widehat{\theta}\right) - \theta$.
   - Monde bootstrap :
     - première approximation : on approxime le biais par $\mathbb{E}_{P_n}\left(\widehat{\theta^*}\right) - \widehat{\theta}$.
     - deuxième approximation : on approxime la quantité précédente par

$$\frac{1}{B}\sum_{b=1}^{B} \widehat{\theta_b^*} - \widehat{\theta}. \tag{1.2.2}$$

     L'estimateur non biaisé est obtenu en retranchant l'estimation du biais à l'estimateur initial. On obtient donc l'estimateur

$$\widehat{\theta} - \left(\frac{1}{B}\sum_{b=1}^{B} \widehat{\theta_b^*} - \widehat{\theta}\right) = 2\widehat{\theta} - \frac{1}{B}\sum_{b=1}^{B} \widehat{\theta_b^*}. \tag{1.2.3}$$

3. Estimation de la variance de $\widehat{\theta}$.
   - Monde réel : on veut estimer $\mathbb{E}_P\left(\left(\widehat{\theta} - \mathbb{E}_P\left(\widehat{\theta}\right)\right)^2\right)$.
   - Monde bootstrap :
     - première approximation : on approxime par

$$\mathbb{E}_P\left(\left(\widehat{\theta^*} - \mathbb{E}_{P_n}\left(\widehat{\theta^*}\right)\right)^2\right)$$

       .
     - deuxième approximation : on approxime la quantité précédente par

$$\frac{1}{B}\sum_{b=1}^{B}\left(\widehat{\theta_b^*} - \frac{1}{B}\sum_{b'=1}^{B}\widehat{\theta_{b'}^*}\right)^2. \tag{1.2.4}$$

1.3 Intervalles de confiance

L'utilisation sans doute la plus fréquente du bootstrap est la production d'intervalles de confiance. En effet, ces intervalles sont cruciaux dans les applications pratiques de l'estimation et le bootstrap permet d'en générer de façon très simple quelque soit l'estimateur. Les assurances théoriques sur la qualité de l'approximation de l'intervalle réel par l'intervalle bootstrap sont cependant dépendantes du cadre d'application et de la méthode utilisée. Celles-ci sont présentées par ordre croissant de l'erreur entre le taux nominal et le taux réel.

On note

$$\widehat{\theta}^*_{(1)} \leqslant \ldots \leqslant \widehat{\theta}^*_{(B)} \tag{1.3.1}$$

le vecteur des B estimations bootstrap rangées par ordre croissant.

1.3.1 Intervalle de confiance des percentiles

L'une des manières les plus simples de fournir un intervalle de confiance est de négliger la première approximation et ainsi de considérer que les lois de $\widehat{\theta}$ et $\widehat{\theta}^*$ sont identiques.

L'intervalle de confiance au niveau $1-\alpha$ est alors

$$\mathrm{IC}_{\mathrm{perc}}\left(1-\alpha\right) = \left[\widehat{\theta}^*_{\lceil B\alpha/2\rceil}, \widehat{\theta}^*_{\lceil B(1-\alpha)/2\rceil}\right]. \tag{1.3.2}$$

1.3.2 Intervalle de confiance classique

Une alternative est de considérer la loi de l'erreur $\widehat{\theta}-\theta$, dont les réalisations bootstrap sont les $\theta^*_b - \theta$. Un intervalle de confiance de l'erreur commise est alors

$$\left[\widehat{\theta^*}_{\lceil B\alpha/2\rceil} - \widehat{\theta}, \widehat{\theta^*}_{\lceil B(1-\alpha)/2\rceil} - \widehat{\theta}\right] \tag{1.3.3}$$

ce qui donne l'intervalle de confiance

$$\mathrm{IC}\left(1-\alpha\right) = \left[2\widehat{\theta} - \widehat{\theta^*}_{\lceil B(1-\alpha)/2\rceil}, 2\widehat{\theta} - \widehat{\theta^*}_{\lceil B\alpha/2\rceil}\right] \tag{1.3.4}$$

1.3.3 Intervalle de confiance du bootstrap standardisé

Une dernière alternative est de considérer la loi de la statistique t (dite de Student)

$$S = \sqrt{n}\frac{\widehat{\theta}-\theta}{\sigma} \tag{1.3.5}$$

dont les réalisation bootstrap sont les

$$S^*_b = \sqrt{n}\frac{\widehat{\theta}^*_b - \widehat{\theta}}{\sigma\left(\mathcal{X}^*_b\right)}, \tag{1.3.6}$$

où $\mathcal{X}_b$ désigne le b-ième échantillon bootstrap. Si l'on dispose d'un estimateur $\widehat{\sigma}^2 = \sigma\left(F_n\right)^2$ de la variance asymptotique $\sigma^2\left(F\right)$, on peut prendre alors

$$\mathrm{IC}_t\left(1-\alpha\right) = \left[\widehat{\theta} - \frac{\widehat{\sigma}}{\sqrt{n}}S^*_{\lceil B(1-\alpha/2)\rceil}, \widehat{\theta} - \frac{\widehat{\sigma}}{\sqrt{n}}S^*_{\lceil B\alpha/2\rceil}\right]. \tag{1.3.7}$$

1.4 Fondements théoriques

1.4.1 Première approximation

Une étude de l'erreur faite lors de la première approximation des étapes du bootstrap est faite dans [4]. Elle s'appuie sur des extensions d'Edgeworth, qui sont des sortes de développement limités appliqués aux distributions en fonction de la taille n de l'échantillon.

Le résultat le plus remarquable est que si $\widehat{\theta}$ est asymptotiquement normal, c'est-à-dire si $S = \sqrt{n}\left(\widehat{\theta} - \theta\right)/\sigma(F)$ converge vers une loi normale centrée réduite, alors il existe un polynôme p tel que

$$\mathbb{P}(S \leqslant x) = \Phi(x) + \frac{1}{\sqrt{n}} p(x)\,\phi(x) + O\left(\frac{1}{n}\right), \tag{1.4.1}$$

où Φ désigne la fonction de répartition d'une loi normale centrée réduite et ϕ la densité d'une loi normale centrée réduite. Dans le monde bootstrap, on peut démontrer que si $S^* = \sqrt{n}\left(\widehat{\theta^*} - \widehat{\theta}\right)/\widehat{\sigma}$ converge vers une loi normale centrée réduite, alors il existe un polynôme p^* tel que

$$\mathbb{P}(S^* \leqslant x) = \Phi(x) + \frac{1}{\sqrt{n}} p^*(x)\,\phi(x) + O_{\mathbb{P}}\left(\frac{1}{n}\right). \tag{1.4.2}$$

De plus, $p - p^* = O_{\mathbb{P}}(n^{-1})$ ce qui implique que

$$\mathbb{P}(S \leqslant x) - \mathbb{P}(S^* \leqslant x) = O_{\mathbb{P}}\left(\frac{1}{n}\right). \tag{1.4.3}$$

L'approximation bootstrap de S est donc asymptotiquement meilleure que l'approximation normale habituellement faite à l'aide du TCL.

Exemple 1.5. Cela fonctionne pour la moyenne et la médiane.

Exemple 1.6. Cela ne fonctionne pas pour les extrêmes. Par exemple, pour une loi uniforme sur $[\theta, \theta+1]$, $n(\min_{1\leqslant i\leqslant n} X_i - \theta)$ converge vers une loi exponentielle.

1.4.2 Deuxième approximation

Des résultats de [2, 5] impliquent la borne suivante entre une loi théorique et celle d'un échantillon tiré suivant cette loi :

Théorème 1.7. *Soit $Y_1, \ldots, Y_B$ un échantillon i.i.d. tiré suivant une loi de fonction de répartition F et soit F_B la fonction de réparitition empirique associée. Alors*

$$\mathbb{P}\left(\sup_{t\in\mathbb{R}} |F(t) - F_B(t)| > \varepsilon\right) \leqslant 2\exp\left(-2\varepsilon^2 B\right). \tag{1.4.4}$$

Ce résultat appliqué à la loi P_n de l'échantillon observé et aux échantillons bootstrap permet de borner l'erreur faite lors de la deuxième approximation.

Remarque 1.8. Comme attendu, on voit que l'erreur dûe à la deuxième approximation ne dépend que de B, c'est-à-dire du nombre d'échantillons bootstrap générés. À condition de bénéficier de la puissance de calcul nécessaire, on peut donc la réduire autant que désiré.

Cependant, la première erreur dépend au contraire de n, qui est la taille de l'échantillon initial. Générer un grand nombre d'échatillons bootstrap n'influera pas sur cette erreur. On retrouve le comportement « habituel », à savoir que des observations plus nombreuses amènent une meilleure précision des estimations faites.

1.5 Test à l'aide de bootstrap

Pour clore ce chapitre, citons une autre application possible du bootstrap, qui est le fait de pouvoir mener des tests statistiques sans les hypothèses requises pour l'utilisation des tests standards.

Un test statistique repose sur :

1. le choix d'une statistique de test
2. le calcul de cette statistique sur l'échantillon observé
3. la détermination de la loi de cette statistique sous H_0.

Les points 2. et 3. permettent alors de déterminer la p-valeur, qui est la probabilité sous la loi du point 3. d'observer un résultat plus surprenant que la statistique observée du point 2. Dans l'approche classique des tests, on fait des hypothèses sur la loi des observations permettant un calcul théorique de la loi de la statistique sous H_0 et donc de la p-valeur. Une possibilité alternative est de générer, à partir des échantillons, des échantillons bootstrap qui vérifient H_0 et de déterminer la valeur de la statistique pour chacun de ces échantillons. On peut alors déterminer une p-valeur empirique en comparant la statistique observée à l'échantillon des statistiques bootstrap. On constatera que cette démarche implique à nouveau les deux approximations du bootstrap. Ce principe général peut s'adapter pour tout type de test mais nécessite de faire attention de trouver une manière de générer des échantillons bootstrap qui vérifient bien H_0 quand les échantillons de départ le vérifient. Les deux exemples ci-dessous, l'un simple et l'autre plus subtil, illustrent ce principe.

1.5.1 Test d'indépendance

On suppose que deux variables X et Y ont été mesurées sur n individus, donnant liex à un échantillon $(x, y) = ((x_1, y_1), \ldots, (x_n, y_n))$. On souhaite tester leur indépendance (test du Khi-deux d'indépendance sur des variables qualitatives ou test de corrélation sur des variables quantitatives par exemple).

Étape 1 On tire B échantillons bootstrap $\left(x^{*b}, y^{*b}\right)$, x^{*b} étant tiré comme un échantillon bootstrap dans x et y^{*b} étant tiré comme un échantillon bootstrap dans y. Les deux parties du tirage se faisant indépendamment l'une de l'autre, les x^{*b} et y^{*b} vivent bien sous H_0.

Étape 2 On détermine la valeur de la statistique s^{*b} sur chaque échantillon et on déduit la p-valeur empirique.

Par exemple, dans le cas d'un test du khi-deux,

$$\text{pval} = \frac{1}{B}\sum_{b=1}^{B} \mathbf{1}_{s^{*i} > s_{\text{obs}}}. \tag{1.5.1}$$

1.5.2 Test d'égalité de la médiane

On dispose de deux échantillons $x_1 = (x_{1,1}, \ldots, x_{1,m})$ et $x_2 = (x_{2,1}, \ldots, x_{2,n})$ de la même variable X mesurée dans deux populations différentes et on veut tester par exemple l'égalité de leurs médianes.

Attention ! On pourrait être tenté de créer des échantillons bootstrap x_1^{*b} de x_1 d'un côté et x_2^{*b} de x_2 de l'autre, puis calculer les statistiques de Wilcoxon issues de leur comparaison. Ce serait une erreur. En effet, les échantillons x_k^{*b} ont pour médiane théorique celle de x_k et à moins que la médiane de x_1 et x_2 ne soit la même, ces échantillons bootstrap ne vérifient pas H_0.

Pour palier ce problème, on note y l'échantillon obtenu en concaténant x_1 et x_2, autrement dit,

$$y = (x_{1,1}, \ldots, x_{1,m}, x_{2,1}, \ldots, x_{2,n}) \in \mathbb{R}^{m+n}. \tag{1.5.2}$$

Si H_0 est vraie, alors les médianes de x_1, x_2 et y sont identiques.

Étape 1 On tire B échantillons bootstrap à partir de y, qui sont notés $y^{*b}, b \in \{1, \ldots, B\}$.

Étape 2 On crée les échantillons x_1^{*b} et x_2^{*b} en prenant respectivement les m premières et n dernières valeurs de y^{*b}. Ces échantillons ont bien la même médiane théorique, qui est celle de z.

Étape 3 On détermine les B statistiques de Wilcoxon des couples d'échantillons bootstrap et on détermine la p-valeur empirique adaptée à la latéralité de test choisie.

1.6 Sources

En plus des articles cités dans le texte, ce chapitre été écrit en utilisant les slides d'Agathe Guilloux disponibles sur http://www.math-evry.cnrs.fr/_media/members/aguilloux/enseignements/bo et les notes de cours d'Étienne Birmelé.

2 Bootstrap : travaux dirigés

2.1 Exercice 1 : Implémentation et illustration des deux approximations

Dans cet exercice, on considère un échantillon x tiré suivant une loi exponentielle $\mathcal{E}(0.3)$ et on s'intéresse à l'estimation de sa médiane m.

1. Écrire des fonctions qui, à partir de x, déterminent les médianes de B échantillons bootstrap (à partir desquels on pourra estimer m et x un intervalle de confiance de m) dans les cas suivants :

(a) bootstrap non-paramétrique à la main (sans utiliser de fonction R pré-écrite)

(b) bootstrap non-paramétrique en utilisant la fonction `boot`.

2. (a) Comparer, à l'aide d'un boxplot, la distribution empirique de l'échantillon $(\widehat{m_1}, \ldots, \widehat{m_B})$ correspondant à $B = 500$ estimations obtenues sur des échantillons indépendants de taille $n = 20$, avec la distribution empirique de $(\widehat{m_1}, \ldots, \widehat{m_B})$ obtenu par bootstrap sur un échantillon.

   (b) Commentez. Quelle est l'approximation illustrée par ce graphique ? L'erreur diminue-t-elle lorqu'on augmente B ?

3. (a) Comparer les distributions empiriques des échantillons bootstrap pour $n = 100, 500, 1000$.

   (b) Commentez. Quelle est l'approximation illustrée par ce graphique ?

2.2 Exercice 2 : Bootstrap paramétrique vs bootstrap non paramétrique

On considère une variable X d'espérance μ et un échantillon x i.i.d. de loi X.

1. Écrire une procédure (soit à la main, soit en utilisant la fonction `bootstt`) qui donne un intervalle de confiance bootstrap de μ à un niveau $(1 - \alpha)$ dans les cas suivants :

   (a) bootstrap non-paramétrique

   (b) bootstrap paramétrique en supposant que X suit une loi exponentielle

   (c) bootstrap paramétrique en supposant que X suit une loi de Poisson

   (d) bootstrap paramétrique en supposant que X suit une loi normale

2. Écrire un programme qui génère 100 échantillons $(x_1, \ldots, x_{15})$ de loi de Poisson de paramètre $\lambda = 1.2$ et qui calcule sur chaque échantillon les bornes inférieures et supérieures des différents intervalles de confiance. Déterminer la largeur empirique moyenne de chaque type d'intervalle de confiance ainsi que la probabilité de couverture empirique de chaque type d'intervalle de confiance. Que constatez-vous ?

Recommencer avec une loi Gamma de paramètre de forme (shape) égal à 1.2 puis 0.8 pour un paramètre d'échelle (scale) égal à chaque fois à 1.

2.3 Exercice 3 : comparaison des espérances de vie entre fumeurs et non-fumeurs

On considère les échantillons suivants, qui donnent les âges de décès de vingt fumeurs et vingt non-fumeurs.

Non-fumeurs : 84, 73, 73, 83, 80, 67, 91, 76, 90, 70, 70, 81, 78, 68, 64, 82, 91, 72, 84, 66
Fumeurs : 94, 65, 71, 83, 47, 55, 96, 57, 57, 64, 77, 97, 51, 50, 48, 41, 71, 59, 86, 71

1. Tester l'égalité des espérances de vies entre les deux conditions à l'aide d'un test basé sur le bootstrap. On précisera les hypothèses, la latéralité et la statistique de test choisie.

On s'intéresse dorénavant au ration des espérances de vie, à savoir, si X_{F} et X_{NF} désignent les âges de décès chez les fumeurs et les non-fumeurs respectivement,

$$\theta = \frac{\mathbb{E}\left[X_{\mathrm{F}}\right]}{\mathbb{E}\left[X_{\mathrm{NF}}\right]}. \tag{2.3.1}$$

2. Déterminer une estimation bootstrap du biais et de la variance de θ.
3. Déterminer un intervalle de confiance à 95% de θ.

2.4 Exercice 4 : bootstrap et régression linéaire

On considère un problème de régression linéaire gaussienne

$$Y_i = \beta X_i + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}\left(0, \sigma_i^2\right). \tag{2.4.1}$$

On dispose d'un échantillon $((X_i, Y_i))_{i=1}^n$. On considère deux manières d'obtenir des intervalles de confiance pour l'estimation de β :

1. Considérer les X_i comme aléatoires et faire des échantillons bootstrap en tirant avec remise des couples (X_i, Y_i).
2. Considérer les X_i comme constants et seuls les résidus comme aléatoires. On peut alors estimer les résidus par $\widehat{\varepsilon_i} = Y_i - X_i\widehat{\beta}$, créer des échantillons bootstrap des résidus et définir $Y_i^{*b} = X_i^{*b}\widehat{\beta} + \varepsilon^{*b}$.

Dans les deux cas, chaque échantillon bootstrap ainsi construit contient n couples (X, Y) qui permettent d'obtenir une estimation bootstrap de β.

1. Charger les données airquality. Après avoir supprimé les lignes ayant une valeur manquante, estimer les paramètres d'une régression linéaire du taux d'ozone en fonction du rayonnement solaire, du vent et de la température. Donner un intervalle de confiance à 95% des coefficients de β avec chacune des méthodes précédentes.
2. Reprendre la question précédente sur des données homoscédastiques ($\sigma_i = \sigma$ pour tout i), puis sur des données fortement hétéroscédastiques (avec des σ_i d'un ordre de grandeur plus grand que $X\beta$). Commentez les résultats données par chacune des méthodes.

2.5 Exercice 5 : majoration de la première erreur du bootstrap

On considère une variable aléatoire X telle que $\mathbb{E}\left[|X|^4\right] < \infty$. On considère que le paramètre θ à estimer est la moyenne et on note

$$F_n(t) = \mathbb{P}\left(\sqrt{n}\left(\widehat{\theta}_n - \theta\right) \leqslant t\right), \quad \widehat{F}_n(t) = \mathbb{P}\left(\sqrt{n}\left(\widehat{\theta}_n^* - \widehat{\theta}\right) \leqslant t\right). \tag{2.5.1}$$

les probabilités d'avoir des erreurs d'estimation supérieures à $t/\sqrt{n}$ dans le monde réel et dans le monde bootstrap. On note Φ_σ la fonction de répartition de la loi normale centrée en 0 et d'écart-type σ. Le but de cet exercice est de montrer que

$$\sup_{t\in\mathbb{R}} \left|F_n(t) - \widehat{F}_n(t)\right| = O_{\mathbb{P}}\left(\frac{1}{\sqrt{n}}\right). \tag{2.5.2}$$

Pour cela, on admet le théorème de Berry-Esseen qui borne la convergence du TCL :

Théorème 2.1 (Berry-Esseen)**.** *Soit X une variable aléatoire d'espérance μ, de variance σ^2 et telle que $\mu^3 := \mathbb{E}\left[|X-\mu|^3\right] < \infty$. On considère un échantillon de n tirages indépendants, de moyenne $\overline{X_n}$. Alors*

$$\sup_{z\in\mathbb{R}}\left|\mathbb{P}\left(\sqrt{n}\frac{\overline{X_n}-\mu}{\sigma}\leqslant z\right)-\Phi_1(z)\right|\leqslant\frac{33}{4}\frac{\mu_3}{\sigma^3\sqrt{n}}. \tag{2.5.3}$$

1. Démontrer que
$$\sup_{t\in\mathbb{R}}|F_n(t)-\Phi_\sigma(t)| = O\left(\frac{1}{\sqrt{n}}\right). \tag{2.5.4}$$
2. Démontrer que
$$\sup_{t\in\mathbb{R}}\left|\widehat{F_n}(t)-\Phi_{\widehat{\sigma}}(t)\right| = O_{\mathbb{P}}\left(\frac{1}{\sqrt{n}}\right). \tag{2.5.5}$$
3. Démontrer que
$$\sup_{t\in\mathbb{R}}|\Phi_\sigma(t)-\Phi_{\widehat{\sigma}}(t)| = O_{\mathbb{P}}\left(\frac{1}{\sqrt{n}}\right). \tag{2.5.6}$$
On pourra commencer par montrer que $|\widehat{\sigma}-\sigma| = O_{\mathbb{P}}\left(n^{-1/2}\right)$.

3 Algorithme Expectation Maximization (EM)

Ce chapitre traite de l'algorithme Expectation-Maximisation, qui est un algorithme couramment utilisé lorqu'on modélise un phénomène à l'aide d'une variable catégorielle non-observée, par exemple un modèle de mélange.

3.1 Modèle de mélange gaussien

On considère une variable aléatoire X sur une population divisée en plusieurs catégories, de façon que la variable X suit une loi normale différente suivant la catégorie considérée.

Exemple 3.1. La taille dans une population humaine peut être considérée comme suivant une loi normale chez les femmes et les hommes, avec des paramètres différents suivant le sexe.

Dans le cas de variables catégorielles connues, l'estimation est simple, il suffit de traiter séparément chaque catégorie. Dans de nombreuses applications cependant, la variable catégorielle est non-observable (on parle de variable cachée ou de variable latente). Comment estimer alors les paramètres ?

3.1.1 Modèle

Le modèle de mélange gaussien est un modèle hiérarchique faisant intervenir une variable catégorielle Z à K classe modélisée par une loi multinomiale. L'individu i est de classe Z_i inconnue et une observation X_i avec

$$Z_i \sim \mathcal{M}(\alpha), \quad \alpha = (\alpha_1,\ldots,\alpha_K), \tag{3.1.1}$$

$$X_i \mid Z_i = k \sim \mathcal{N}(\mu_k, \sigma_k), \tag{3.1.2}$$

où $\sum_{k=1}^{K} \alpha_k = 1$ et $0 \leqslant \alpha_k \leqslant 1$.

3.1.2 Vraisemblance

On note $\Theta = (\alpha, \mu, \sigma) \in \mathbb{R}^{3K}$ l'ensemble des paramètres. De plus, on note $f_{\mu,\sigma}(x)$ la densité évaluée en x de la loi normale de moyenne μ et écart-type σ.

La vraisemblance conditionnelle sachant la variable Z_i est simplement la densité gaussienne, ce qui permet d'écrire la vraisemblance d'une observation

$$\mathcal{L}(X_i \mid \Theta) = \sum_{k=1}^{K} \mathcal{L}(X_i \mid Z_i = k, \Theta)\, \mathbb{P}(Z_i = k, \Theta) \tag{3.1.3}$$

$$= \sum_{k=1}^{K} \alpha_i f_{\mu_k,\sigma_k}(X_i). \tag{3.1.4}$$

Les observations étant indépendantes, la vraisemblance et log-vraisemblance complète s'écrivent alors

$$\mathcal{L}(\mathbf{X} \mid \Theta) = \prod_{i=1}^{n} \left(\sum_{k=1}^{K} \alpha_i f_{\mu_k,\sigma_k}(X_i) \right), \tag{3.1.5}$$

$$\log \mathcal{L}(\mathbf{X} \mid \Theta) = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} \alpha_i f_{\mu_k,\sigma_k}(X_i) \right). \tag{3.1.6}$$

La présence de la somme dans le logarithme empêche une écriture plus simple ainsi que la résolution théorique du problème d'optimisation. Le système d'équations résultant de l'annulation des dérivées partielles ne peut pas se résoudre à la main.

Une approche possible est de recourir à des algorithmes numériques de type Newton-Raphson pour résoudre le problème d'optimisation. Cependant, elle devient potentiellement très lourde lorsque la dimension du problème ou le nombre de classes augmente.

Remarque 3.2. Une autre méthode d'écriture de la vraisemblance est de la décomposer suivant l'ensemble du vecteur de variables latentes $\mathbf{Z}$:

$$\mathcal{L}(\mathbf{X} \mid \mathbf{Z}, \Theta) = \prod_{i=1}^{n} f_{\mu_{Z_i},\sigma_{Z_i}}(X_i) \tag{3.1.7}$$

et

$$\mathcal{L}(\mathbf{X} \mid \Theta) = \sum_{k_1,\ldots,k_n=1}^{K} \prod_{i=1}^{n} \alpha_{k_i} f_{\mu_{k_i},\sigma_{k_i}}(X_i). \tag{3.1.8}$$

Cette manière de procéder ne rend pas l'optimisation plus simple, au contraire. En effet, la somme fait intervenir K^n termes, ce qui rend même l'évaluation de la vraisemblance impossible en utilisation cette formule.

En revanche, on peut voir que dans le cas où les Z_i sont connus, l'estimation des paramètres μ et σ est aisée, étape sur laquelle reposera l'algorithme EM.

3.1.3 Loi a posteriori

Dans certaines applications des modèles à variables cachées, l'intérêt ne repose pas que dans l'estimation des paramètres des différentes lois, mais aussi dans le fait d'estimer la probabilité d'appartenance des différentes observations aux différentes classes. En termes de statistiques bayésiennes, on cherche à déterminer la loi de Z_i au vu des observations, c'est-à-dire sa loi a posteriori.

Ce calcul est simple à faire en utilisant la formule de Bayes une fois les paramètres du modèle estimés. Dans le cadre du mélange gaussien,

$$\mathbb{P}(Z_i = k \mid X_i) = \frac{\mathbb{P}(X_i \mid Z_i = k)\,\mathbb{P}(Z_i = k)}{\mathcal{L}(X_i)} \tag{3.1.9}$$

$$\propto \alpha_k f_{\mu_k,\sigma_k}(X_i)\,. \tag{3.1.10}$$

La somme des appartenances de Z_i à chacune des classes étant égale à 1, déterminer les numérateurs et utiliser cette contrainte permet de conclure.

3.2 Algorithme EM

3.2.1 Introduction

On se place dans le cas général d'un couple de variables (X, Z) tel que

— X est observée, et supposée suivre une loi paramétrique,

— Z est une variable qualitative latente.

On note θ le vecteur contenant les paramètres de X et les proportions de classes gouvernant Z.

On cherche à résoudre le problème d'optimisation

$$\widehat{\theta} = \text{argmax}_\theta\, \mathbb{P}(X \mid \theta) = \text{argmax}_\theta \sum_Z \mathbb{P}(X, Z \mid \theta)\,. \tag{3.2.1}$$

Dans de nombreux cas, par exemple le modèle de mélange gaussien, ce problème est difficile à résoudre alors que maximiser $\mathbb{P}(X, Z \mid \theta)$ est simple. Il est alors naturel de chercher à affecter des valeurs de Z pour estimer θ en s'intéressant à la loi a posteriori $\mathbb{P}(Z \mid X, \theta)$.

Cette dernière loi dépendant elle-même de θ, une telle approche entraîne la mise en place d'une procédure qui, partant d'une valeur arbitraire θ_0, va remettre à jour θ jusqu'à convergence.

Plusieurs stratégies sont possibles.

Stratégie 1 : MAP On peut assigner à chaque individu la valeur Z_i dont la probabilité a posteriori est la plus grande, ou MAP. Cette approche a l'avantage d'être très simple à mettre en oeuvre mais se paye par le fait d'une grande perte d'information pour les individus intermédiaires. Elle donne des résultats d'autant moins bons que les classes sont mélangées (proche ou de grande variance).

Dans ce cas, θ_m perme de déterminer le vecteur $\mathbf{Z}^{\text{MAP}}$ et

$$\theta_{m+1} = \text{argmax}_\theta \log \mathbb{P}\left(\mathbf{X}, \mathbf{Z}^{\text{MAP}} \mid \theta\right). \tag{3.2.2}$$

On notera que dans le cas de classes équiprobables, cette approche correspond à l'algorithme de classification des k-means (cf exercice 4.3).

Stratégie 2 : tirage au sort ou SEM Une autre approche est d'utiliser la loi a posteriori de Z_i pour tirer au sort l'affectation de l'individu i. Dans ce cas, θ_m permet de tirer au sort un vecteur $\mathbf{Z}^1$ et

$$\theta_{m+1} = \text{argmax}_\theta \log \mathbb{P}\left(\mathbf{X}, \mathbf{Z}^1 \mid \theta\right). \tag{3.2.3}$$

D'un point de vue intuitif, un individu dont on est sûr de la classe y restera avec grande probabilité mais un individu incertain en changera facilement.

Cette approche permet de perdre beaucoup moins d'information mais ne converge pas au sens où tout individu finit par rechanger de classe infiniment de fois. Elle converge cependant au sens markovien du term, à savoir qu'elle définit une chaîne de Markov convergente sur les Z, mais implique en pratique un temps d'exécution assez long.

Il est à noter qu'elle reste une option utilisée lorsque l'algorithme EM ne peut pas être mis en place sous l'appellation SEM pour stochastic EM.

Stratégie 3 : l'algorithme EM (Expectation-Maximisation) Dempster, Laird et Rubin, [1]

L'instabilité en termes de classes de l'approche SEM peut être réduite en ne tirant non pas une mais N valeurs de $\mathbf{Z}$ suivant $\mathbb{P}(Z \mid X, \theta_m)$ et en considérant la maximisation de la vraisemblance moyenne

$$\theta_{m+1} = \text{argmax}_\theta \frac{1}{N} \sum_{j=1}^{N} \log \mathbb{P}\left(X, Z^j \mid \theta\right). \tag{3.2.4}$$

En faisant tendre N vers l'infini,

$$\lim_{N\to\infty} \frac{1}{N} \sum_{j=1}^{N} \log \mathbb{P}\left(X, Z^j \mid \theta\right) = \int_Z \mathbb{P}(Z \mid X, \theta_m) \log \mathbb{P}(X, Z \mid \theta)\, dZ \tag{3.2.5}$$

$$= \mathbb{E}_Z \left[\log(X, Z \mid \theta) \mid X, \theta_m\right]. \tag{3.2.6}$$

Considérer

$$\theta_{m+1} = \text{argmax}_\theta \, \mathbb{E}_Z \left[\log(X, Z \mid \theta) \mid X, \theta_m\right] \tag{3.2.7}$$

permet d'éliminer l'incertitudue liée à l'échantillonage de SEM, tout en prenant tout la distribution a posteriori en compte.

3.3 Algorithme et preuve de convergence

3.3.1 Étapes de l'algorithme

1. On dispose d'observations i.i.d. $\mathbf{X} = (X_1, \ldots, X_n)$ de vraisemblance notée $\mathbb{P}(\mathbf{X} \mid \theta)$.
2. Maximiser $\log \mathbb{P}(\mathbf{X} \mid \theta)$ est impossible.
3. On considère des données cachées $\mathbf{Z} = (Z_1, \ldots, Z_n)$ dont la connaissance rendrait possible la maximisation de la log-vraisemblance des données complètes $\log \mathbb{P}(\mathbf{X}, \mathbf{Z} \mid \theta)$.

4. Comme on ne connaît pas ces données $\mathbf{Z}$, on estime la vraisemblance des données complètes en prenant en compte toutes les informations connues : l'estimateur est naturellement $\mathbb{E}_{\mathbf{Z}|\mathbf{X},\theta_m}\left[\log \mathbb{P}\left(\mathbf{X}, \mathbf{z} \mid \theta\right)\right]$ (étape « E »de l'algorithme).
5. On maximise cette vraisemblance estimée pour déterminer la nouvelle valeur du paramètre (étape « M »de l'algorithme).

Par conséquent, le passage de l'itération m à l'itération $m+1$ de l'algorithme consiste à déterminer

$$\theta_{m+1} = \operatorname{argmax}_\theta \mathbb{E}_{\mathbf{Z}|\mathbf{X},\theta_m}\left[\log \mathbb{P}\left(\mathbf{Z}, \mathbf{z} \mid \theta\right)\right]. \tag{3.3.1}$$

3.3.2 Démonstration de la croissance de la vraisemblance d'une itération à l'autre

À l'itération m, on dispose d'une valeur $\theta_m \in \mathbb{R}^d$ du vecteur de paramètres. On cherche une valeur θ, que l'on notera ensuite θ_{m+1}, qui augmente la vraisemblance, c'est-à-dire telle que

$$\Delta\left(\theta, \theta_m\right) := \log \mathbb{P}\left(\mathbf{X} \mid \theta\right) - \log \mathbb{P}\left(\mathbf{X} \mid \theta_m\right) \geqslant 0. \tag{3.3.2}$$

De plus, on cherche à rendre $\Delta\left(\theta, \theta_m\right)$ aussi grand que possible.

Malheureusement, on ne sait pas maximiser $\mathbb{P}\left(\mathbf{X} \mid \theta\right)$ et il en va de même pour $\Delta\left(\theta, \theta_m\right)$. Pour palier ce problème, on cherche une fonction $\theta \mapsto \delta\left(\theta, \theta_m\right)$ que l'on sait maximiser et qui vérifie

$$\forall \theta \in \mathbb{R}^d, \quad \Delta\left(\theta, \theta_m\right) \geqslant \delta\left(\theta \mid \theta_m\right) \tag{3.3.3}$$

$$\delta\left(\theta_m \mid \theta_m\right) = 0. \tag{3.3.4}$$

Si on trouve θ' qui maximise la fonction $\theta \mapsto \delta\left(\theta \mid \theta\right)$, alors on aura nécessairement $\Delta\left(\theta', \theta_m\right) \geqslant \delta\left(\theta_m \mid \theta_m\right) = 0$. Afin de trouver une fonction δ vérifiant (3.3.3) et (3.3.4), on représente la vraisemblance à l'aide des données cachées $\mathbf{Z} = (Z_1, \ldots, Z_n)$:

$$\mathbb{P}\left(\mathbf{X} \mid \theta\right) = \int \mathbb{P}\left(\mathbf{X}, \mathbf{z} \mid \theta\right) dZ\left(\mathbf{z}\right) = \int \mathbb{P}\left(\mathbf{X} \mid \mathbf{z}, \theta\right) \mathbb{P}\left(\mathbf{z} \mid \theta\theta\right) dZ\left(\mathbf{z}\right). \tag{3.3.5}$$

En utilisant $\int \mathbb{P}\left(\mathbf{z} \mid \mathbf{X}, \theta_m\right) dZ\left(\mathbf{z}\right) = 1$, on obtient

$$\Delta\left(\theta, \theta_m\right) = \log \mathbb{P}\left(\mathbf{X} \mid \theta\right) - \log \mathbb{P}\left(\mathbf{X} \mid \theta_m\right) \tag{3.3.6}$$

$$= \log\left(\int \mathbb{P}\left(\mathbf{X} \mid \mathbf{z}, \theta\right) \mathbb{P}\left(\mathbf{z} \mid \theta\right) dZ\left(\mathbf{z}\right)\right) - \int \mathbb{P}\left(\mathbf{z} \mid \mathbf{X}, \theta_m\right) dZ\left(\mathbf{z}\right) \log \mathbb{P}\left(\mathbf{X} \mid \theta_m\right). \tag{3.3.7}$$

On rappelle que l'inégalité de Jensen donne, pour toute variable aléatoire Y intégrable et toute fonction convexe $\phi\colon \mathbb{R} \to \mathbb{R}$,

$$\phi\left(\mathbb{E}\left[Y\right]\right) \leqslant \mathbb{E}\left[\phi\left(Y\right)\right]. \tag{3.3.8}$$

On peut écrire (3.3.6) sous la forme

$$\Delta\left(\theta, \theta_m\right) = \log\left(\int \frac{\mathbb{P}\left(\mathbf{X} \mid \mathbf{z}, \theta\right) \mathbb{P}\left(\mathbf{z} \mid \theta\right)}{\mathbb{P}\left(\mathbf{z} \mid \mathbf{X}, \theta_m\right)} \mathbb{P}\left(\mathbf{z} \mid \mathbf{X}, \theta_m\right) dZ\left(\mathbf{z}\right)\right)$$

$$- \int \mathbb{P}\left(\mathbf{z} \mid \mathbf{X}, \theta_m\right) dZ\left(\mathbf{z}\right) \log \mathbb{P}\left(\mathbf{X} \mid \theta_m\right). \tag{3.3.9}$$

En appliquant (3.3.8) à la fonction $\phi\colon y \mapsto -\log y$, on obtient

$$\begin{aligned}\Delta(\theta, \theta_m) &\geqslant \int \mathbb{P}(\mathbf{z} \mid \mathbf{X}, \theta_m) \log\left(\frac{\mathbb{P}(\mathbf{X} \mid \mathbf{z}, \theta)\,\mathbb{P}(\mathbf{z} \mid \theta)}{\mathbb{P}(\mathbf{z} \mid \mathbf{X}, \theta_m)}\right) dZ(\mathbf{z}) - \int \mathbb{P}(\mathbf{z} \mid \mathbf{X}, \theta_m)\, dZ(\mathbf{z}) \log \mathbb{P}(\mathbf{X} \mid \theta_m) \\ &= \int \mathbb{P}(\mathbf{z} \mid \mathbf{X}, \theta_m) \log\left(\frac{\mathbb{P}(\mathbf{X} \mid \mathbf{z}, \theta)\,\mathbb{P}(\mathbf{z} \mid \theta)}{\mathbb{P}(\mathbf{z} \mid \mathbf{X}, \theta_m)\,\mathbb{P}(\mathbf{X} \mid \theta_m)}\right) dZ(\mathbf{z}) \\ &= \int \mathbb{P}(\mathbf{z} \mid \mathbf{X}, \theta_m) \log\left(\frac{\mathbb{P}(\mathbf{X}, \mathbf{z} \mid \theta)}{\mathbb{P}(\mathbf{X}, \mathbf{z} \mid \theta_m)}\right) dZ(\mathbf{z})\end{aligned}$$

ce qui nous mène à la définition

$$\delta(\theta \mid \theta_m) := \int \mathbb{P}(\mathbf{z} \mid \mathbf{X}, \theta_m) \log\left(\frac{\mathbb{P}(\mathbf{X}, \mathbf{z} \mid \theta)}{\mathbb{P}(\mathbf{X}, \mathbf{z} \mid \theta_m)} dZ(\mathbf{z})\right). \tag{3.3.10}$$

On vient de voir que (3.3.3) est vérifiée, pour (3.3.4), c'est évident.

On pose alors

$$\theta_{m+1} = \operatorname{argmax}_\theta \delta(\theta \mid \theta_m) \tag{3.3.11}$$

$$= \operatorname{argmax}_\theta \int \mathbb{P}(\mathbf{z} \mid \mathbf{X}, \theta_m) \log\left(\frac{\mathbb{P}(\mathbf{X}, \mathbf{z} \mid \theta)}{\mathbb{P}(\mathbf{X}, \mathbf{z} \mid \theta_m)} dZ(\mathbf{z})\right) \tag{3.3.12}$$

$$= \operatorname{argmax}_\theta \int \mathbb{P}(\mathbf{z} \mid \mathbf{X}, \theta_m) \log\left(\mathbb{P}(\mathbf{X}, \mathbf{z} \mid \theta)\, dZ(\mathbf{z})\right) \tag{3.3.13}$$

$$= \operatorname{argmax}_\theta \mathbb{E}_{\mathbf{Z}|\mathbf{X},\theta_m}\left[\log \mathbb{P}(\mathbf{X}, \mathbf{z} \mid \theta)\right]. \tag{3.3.14}$$

La valeur θ_{m+1} est plus vraisemblable que θ_m, car

$$\log \mathbb{P}(\mathbf{X} \mid \theta_{m+1}) - \log \mathbb{P}(\mathbf{X} \mid \theta_m) = \Delta(\theta_{m+1} \mid \theta_m) \geqslant \delta(\theta_{m+1} \mid \theta_m) \geqslant \delta(\theta_m \mid \theta_m) = 0. \tag{3.3.15}$$

3.4 Commentaires

On sait que la suite $(\mathbb{P}(\mathbf{X} \mid \theta_m))_{m \geqslant 1}$ converge car elle est croissante et bornée. En revanche, il est possible que la convergence n'ait lieu que vers un maximum local. Le point de convergence dépend du point de départ. Il est préférable quand c'est possible de multiplier les points de départs, ou de ne pas le choisir arbitrairement.

4 Algorithme EM : travaux dirigés

4.1 Exercice 1 : Mélange gaussien à deux composantes

On considère un échantillon $X = (X_1, \ldots, X_n)$ tiré suivant le processus suivant :

$$Z_i \sim \text{Ber}(p)$$

$$X_i \mid Z_i = k \sim \mathcal{N}\left(\mu_k, \sigma_k^2\right).$$

On cherche à estimer $\theta = (p, \mu_0, \sigma_0, \mu_1, \sigma_1)$.

1. Déterminer $\delta(\theta \mid \theta_m)$.
2. Déterminer θ_{m+1}.
3. Implémenter l'algorithme correspondant. Le tester sur un jeu de données simulé.

4.2 Exercice 2 : Mélange gaussien à K composantes

Reprendre l'exercice précédent avec K composantes.

4.3 Exercice 3 : Algorithme des k-moyennes

On considère un ensemble de points $(X_1, \dots, X_n)$ de $\mathbb{R}^p$ que l'on chercher à répartir en k classes. Pour cela, l'algorithme des k-moyennes fonctionne ainsi :

1. K points $\mu_1, \dots, \mu_K$ de $\mathbb{R}^p$, appelées moyennes, sont choisis arbitrairement.
2. Chaque X_i est affecté à la classe k, où
$$k = \operatorname{argmin}_j \|X_i - \mu_j\|^2 . \tag{4.3.1}$$
3. Pour tout $1 \leqslant k \leqslant K$, on définit μ_k comme la moyenne des points qui ont été affectés à la classe k.
4. On répète les étapes 2 et 3 jusqu'à convergence.

Montrer que l'algorithme des k-moyennes revient à appliquer l'algorithme EM en supposant un mélange gaussien homoscédastique et en remplaçant l'étape E par la règle du maximum a posteriori avec des classes équiréparties a priori.

4.4 Exercice 4 : Mélange de lois exponentielles

Écrire un algorithme EM pour l'esitmation des paramètres d'un mélange de trois lois exponentielles. L'implémenter et l'illustrer sur des données simulées.

4.5 Exercice 5 : Détection d'une rupture dans un processus de comptage

On suppose qu'une suite de réels $(x_1, \dots, x_n)$ provient de réalisations de variables aléatoires $X_1, \dots, X_n$ telles qu'il existe un entier Z non observé tel que

$$\begin{aligned} Z &\sim \mathcal{U}(\{1, \dots, n\}), \\ X_i &\sim \mathcal{P}(\lambda_1) \text{ pour } i \leqslant Z, \\ X_i &\sim \mathcal{P}(\lambda_2) \text{ pour } i > Z. \end{aligned}$$

Autrement dit, les X_i suivent une loi de Poisson avec un changement de paramètre à un moment inconnu.

1. Écrire la vraisemblance d'une observation sous ce modèle.
2. Écrire l'étape E de l'algorithme EM.
3. Écrire l'étape M de l'algorithme EM.
4. Simuler un jeu de données avec $n = 100$, $Z = 40$, $\lambda_1 = 30$, $\lambda_2 = 60$.

4.6 Exercice 6 : Estimation en génétique

Le gène codant pour le groupe sanguin d'un individu a trois allèles possibles : A, B et O. Le type O est récessif, les types A et B sont dominants. On observe quatre groupes sanguins (ou phénotypes) : [A] (pour un génotype AA ou AO), [B] (pour un génotype BB ou BO), [AB] (pour un génotype AB) et [O] (pour un génotype OO). On souhaite estimer les probabilités qu'un chromosome porte l'allèle A, B ou O respectivement notées p_A, p_B et $p_O = 1 - p_A - p_B$.

On observe pour cela le phénotype de 521 personnes. Les données sont les suivantes :

n_A	n_B	n_{AB}	n_O
186	38	13	284

Déterminer une estimation de p_A, p_B et p_O en utilisant l'algorithme EM.

4.7 Exercice 7 : Estimation en épidémiologie

D'après les observations rapportées par McKendrick (1926), le choléra affectant les foyers indiens au début du 20ème siècle peut donner lieu à des infections sévères ou à des infections plus légèresqui sont alors asymptomatiques de sorte que l'observation des foyers sans cas est non fiable. Les malades asymptomatiques sont néanmoins transmetteurs de la maladie. McKendrick a recensé le nombre de cas par foyers dans une certaine région indienne au cours d'une épidémie donnée. Les données sont les suivantes :

nombre de cas par foyer	0	1	2	3	4	$\geqslant 5$
effectif	NA	32	16	6	1	0

En admettant que les comptages suivent une loi de Poisson, proposer une imputation du nombre de foyer sans cas en utilisant l'algorithme EM.

4.8 Exercice 8 : Etude démographique

Une étude démographique est réalisée afin d'étudier le nombre moyen d'enfants par foyer. Une campagne de recensement est effectuée. Pour des raisons techniques, les données suivantes sont recueillies :

nombre d'enfants par foyer	0	1ou 2	3 ou 4	5	$\geqslant 6$
effectifs	24	32	16	4	0

En admettant que la loi du nombre d'enfants par foyer est une loi de Poisson, proposer une estimation du paramètre λ de cette loi en utilisant l'algorithme EM.

Références

[1] A. P. Dempster, N. M. Laird, and D. B. Rubin, *Maximum likelihood from incomplete data via the EM algorithm*, J. Roy. Statist. Soc. Ser. B **39** (1977), no. 1, 1–38, With discussion. MR 501537

[2] A. Dvoretzky, J. Kiefer, and J. Wolfowitz, *Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator*, Ann. Math. Statist. **27** (1956), 642–669. MR 83864

[3] B. Efron, *Bootstrap methods : another look at the jackknife*, Ann. Statist. **7** (1979), no. 1, 1–26. MR 515681

[4] Peter Hall, *The bootstrap and Edgeworth expansion*, Springer Series in Statistics, Springer-Verlag, New York, 1992. MR 1145237

[5] P. Massart, *The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality*, Ann. Probab. **18** (1990), no. 3, 1269–1283. MR 1062069