Résumé du cours d'Outils statistiques S3
Université de Strasbourg
Master Statistiques

Davide Giraudo

8 septembre 2024

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 $\mathbb{X} = (X_1, \dots, X_n)$ 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 $\widehat{\theta}$:

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

1.1.1 Option 1 : Hypothèses sur la forme de la loi

On suppose que P 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).

Limitations :

1. La crédibilité de l'hypothèse de forme.
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

On aimerait disposer d'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. Il faudrait ici $n \times B$ mesures i.i.d. selon la loi inconnue, avec n et B les plus grands possibles : impossible en pratique.
2. Borne de l'erreur en fonction de B et n.

1.1.3 Option 3 : Bootstrap

Approche appelée Bootstrap 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.

Approche similaire à l'option 2 mais 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

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. Borne de l'erreur en fonction de B et n.

1.2 Echantillon bootstrap

1.2.1 Produire des échantillons bootstrap

Procédure introduite par Efron [3] : à un grand nombre B d'échantillons sont tirés suivant la loi empirique de l'échantillon $\mathbb{X} = (X_1, \ldots, X_n)$. Principe : 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 1.1. 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.

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 1.2. 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

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 $\widehat{\theta}^*_b = T\left(\mathcal{X}^*_b\right)$
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 $\left(\widehat{\theta}^*_b\right)_{1 \leqslant b \leqslant B}$

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

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

On considère que $\widehat{\theta} \approx \widehat{\theta^*}$. Ainsi

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

1.3.2 Intervalle de confiance classique

On considère la loi de l'erreur $\widehat{\theta} - \theta$. Intervalle de confiance de l'erreur commise

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

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^{e} 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.

Si $S = \sqrt{n}\left(\widehat{\theta} - \theta\right)/\sigma\left(F\right)$ converge vers une loi normale centrée réduite, alors il existe un polynôme p_n tel que

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

où Φ désigne la fonction de répartition d'une loi normale centrée réduite et ϕ la densité d'une loi normale centrée réduite. 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_n^* tel que

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

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

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

Exemple 1.3. Cela fonctionne pour la moyenne et la médiane.

Exemple 1.4. Cela ne fonctionne pas pour les extrêmes. Par exemple, pour une loi uniforme sur $[\theta, \theta + 1]$, $n\left(\min_{1 \leqslant i \leqslant n} X_i - \theta\right)$ converge vers une loi exponentielle.

1.4.2 Deuxième approximation

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

Théorème 1.5. *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 répartition empirique associée. Alors*

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

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 1.6. L'erreur due à la deuxième approximation ne dépend que du nombre d'échantillons bootstrap générés.

La première erreur dépend au contraire de la taille de l'échantillon initial. Générer un grand nombre d'échatillons bootstrap n'influera pas sur cette erreur.

1.5 Test à l'aide de bootstrap

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.

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.

1.5.1 Test d'indépendance

On suppose que deux variables X et Y ont été mesurées sur n individus, donnant lieu à un échantillon $(x, y) = ((x_1, y_1), \ldots, (x_n, y_n))$. On souhaite tester leur indépendance.

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 $y^{*b}, b \in \{1, \dots, B\}$.

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/bootstrap/slides.pdf et les notes de cours d'Étienne Birmelé.

2 Algorithme Expectation Maximization (EM)

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

2.1 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 2.1. 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 ?

2.1.1 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

$$Z_i \sim \mathcal{M}(\alpha), \quad \alpha = (\alpha_1, \dots, \alpha_K), \tag{2.1.1}$$

$$X_i \mid Z_i = k \sim \mathcal{N}(\mu_k, \sigma_k), \tag{2.1.2}$$

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

2.1.2 Vraisemblance

On note $\Theta = (\alpha, \mu, \sigma)$ 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.

$$\mathcal{L}(X_i \mid \Theta) = \sum_{k=1}^{K} \mathcal{L}(X_i \mid Z_i = k, \Theta)\, \mathbb{P}(Z_i = k, \Theta) \tag{2.1.3}$$

$$= \sum_{k=1}^{K} \alpha_i f_{\mu_k,\sigma_k}(X_i). \tag{2.1.4}$$

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

$$\mathcal{L}(\mathbf{X} \mid) = \prod_{i=1}^{n} \left(\sum_{k=1}^{K} \alpha_i f_{\mu_k,\sigma_k}(X_i) \right), \tag{2.1.5}$$

$$\log \mathcal{L}(\mathbf{X} \mid) = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} \alpha_i f_{\mu_k,\sigma_k}(X_i) \right). \tag{2.1.6}$$

Remarque 2.2. Une autre méthode d'écriture de la vraisemblance est de la décomposer suivant l'ensemble du vecteur de variables latentes $\mathbf{Z}$:

$$\mathcal{L}(\mathbf{X} \mid \mathbf{Z},) = \prod_{i=1}^{n} f_{\mu_{Z_i},\sigma_{Z_i}}(X_i) \tag{2.1.7}$$

et

$$\mathcal{L}(\mathbf{X} \mid) = \sum_{k_1,\ldots,k_n=1}^{K} \prod_{i=1}^{n} \alpha_{k_i} f_{\mu_{k_i},\sigma_{k_i}}(X_i). \tag{2.1.8}$$

Mais il y a K^n termes : évaluation de la vraisemblance impossible.

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.

2.1.3 Loi a posteriori

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.

Formule de Bayes dans le cadre du mélange gaussien,

$$\mathbb{P}(Z_i = k \mid X_i) = \frac{\mathbb{P}(X_i \mid Z_i = k)\, \mathbb{P}(Z_i = k)}{\mathcal{L}(X_i)} \tag{2.1.9}$$

$$\propto \alpha_k f_{\mu_k,\sigma_k}(X_i). \tag{2.1.10}$$

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.

2.2 Algorithme EM

2.2.1 Introduction

On se place dans le cas général d'une 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.

Pproblème d'optimisation

$$\widehat{\theta} = \text{argmax}_\theta \, \mathbb{P}(X \mid \theta) = \text{argmax}_\theta \sum_Z \mathbb{P}(X, Z \mid \theta). \tag{2.2.1}$$

D ifficile 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 : procédure qui, partant d'une valeur arbitraire θ_0, va remettre à jour θ jusqu'à convergence.

Plusieurs stratégies sont possibles.

Stratégie 1 : Max a posteriori (MAP) On peut assigner à chaque individu la valeur Z_i dont la probabilité a posteriori est la plus grande. Approche simple mais 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 permet de déterminer le vecteur $\mathbf{Z}^{\text{MAP}}$ et

$$\theta_{m+1} = \text{argmax}_\theta \log \mathbb{P}\left(\mathbf{X}, \mathbf{Z}^{\text{MAP}} \mid \theta\right). \tag{2.2.2}$$

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

$$\theta_{m+1} = \text{argmax}_\theta \log \mathbb{P}\left(\mathbf{X}, \mathbf{Z}^1 \mid \theta\right). \tag{2.2.3}$$

Convergence au sens markovien du terme, à savoir qu'elle définit une chaîne de Markov convergente sur les Z, temps d'exécution assez long.

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

$$\theta_{m+1} = \text{argmax}_\theta \, \frac{1}{N} \sum_{j=1}^{N} \log \mathbb{P}\left(X, Z^j \mid \theta\right). \tag{2.2.4}$$

En faisant tendre N vers l'infini,

$$\lim_{N\to\infty} \frac{1}{N} \sum_{j=1}^{N} \log \mathbb{P}\left(X, Z^j \mid \theta\right) = \int_Z \mathbb{P}(Z \mid X, \theta_m) \log \mathbb{P}(X, Z \mid \theta) \, dZ \tag{2.2.5}$$

$$= \mathbb{E}_Z \left[\log(X, Z \mid \theta) \mid X, \theta_m\right]. \tag{2.2.6}$$

Considérer

$$\theta_{m+1} = \operatorname{argmax}_\theta \mathbb{E}_Z \left[\log\left(X, Z \mid \theta\right) \mid X, \theta_m\right] \tag{2.2.7}$$

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

2.3 Algorithme et preuve de convergence

2.3.1 Étapes de l'algorithme

1. On dispose d'observations i.i.d. $\mathbf{X} = (X_1, \ldots, X_n)$ de vraisemblance notée $\mathbb{P}\left(\mathbf{X} \mid \theta\right)$.
2. Maximiser $\log \mathbb{P}\left(\mathbf{X} \mid \theta\right)$ 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 $\log \mathbb{P}\left(\mathbf{X}, \mathbf{Z} \mid \theta\right)$.
4. Comme on ne connaît pas ces donnés $\mathbf{Z}$, on estime la vraisemblance des données complètes en prenant en compte toutes les informations connues : l'estimateur est naturellement $\mathbb{E}_{\mathbf{Z}|\mathbf{X},\theta_m}\left[\log \mathbb{P}\left(\mathbf{X}, \mathbf{z} \mid \ \mathbf{theta}\right)\right]$ (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

$$\theta_{m+1} = \operatorname{argmax}_\theta \mathbb{E}_{\mathbf{Z}|\mathbf{X},\theta_m}\left[\log \mathbb{P}\left(\mathbf{Z}, \mathbf{z} \mid \theta\right)\right]. \tag{2.3.1}$$

2.3.2 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

$$\Delta\left(\theta, \theta_m\right) := \log \mathbb{P}\left(\mathbf{X} \mid \theta\right) - \log \mathbb{P}\left(\mathbf{X} \mid \theta\right) \geqslant 0. \tag{2.3.2}$$

On ne sait pas maximiser $\mathbb{P}\left(\mathbf{X} \mid \theta\right)$ et il en va de même pour $\Delta\left(\theta, \theta_m\right) : \theta \mapsto \delta\left(\theta, \theta_m\right)$ que l'on sait maximiser et qui vérifie

$$\forall \theta \in \mathbb{R}^d, \quad \Delta\left(\theta, \theta_m\right) \geqslant \delta\left(\theta \mid \theta_m\right) \tag{2.3.3}$$

$$\delta\left(\theta_m \mid \theta_m\right) = 0. \tag{2.3.4}$$

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 un fonction δ vérifiant (2.3.3) et (2.3.4), on représente la vraisemblance à l'aide des données cachées $\mathbf{Z} = (Z_1, \ldots, Z_n)$:

$$\mathbb{P}\left(\mathbf{X} \mid \theta\right) = \int \mathbb{P}\left(\mathbf{X}, \mathbf{z} \mid \theta\right) dZ\left(\mathbf{z}\right) = \int \mathbb{P}\left(\mathbf{X} \mid \mathbf{z}, \theta\right) \mathbb{P}\left(\mathbf{z}, \theta\right) dZ\left(\mathbf{z}\right). \tag{2.3.5}$$

En utilisant $\int \mathbb{P}(\mathbf{z} \mid \mathbf{X}, \theta_m)\, dZ(\mathbf{z}) = 1$ et l'inégalité de Jensen on obtient

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

ce qui nous mène à la définition

$$\delta(\theta \mid \theta_m) := \int \mathbb{P}(\mathbf{z} \mid \mathbf{X}, \theta_m) \log\left(\frac{\mathbb{P}(\mathbf{X}, \mathbf{z} \mid \theta)}{\mathbb{P}(\mathbf{X}, \mathbf{z} \mid \theta_m)}\right) dZ(\mathbf{z}). \tag{2.3.6}$$

On vient de voir que (2.3.3) est vérifiée, pour (2.3.4), c'est évident.

On pose alors

$$\theta_{m+1} = \operatorname{argmax}_\theta \mathbb{E}_{\mathbf{Z}|\mathbf{X},\theta_m}[\log \mathbb{P}(\mathbf{X}, \mathbf{z} \mid \theta)]. \tag{2.3.7}$$

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

$$\log \mathbb{P}(\mathbf{X} \mid \theta_{m+1}) - \log \mathbb{P}(\mathbf{X} \mid \theta_m) = \Delta(\theta_{m+1} \mid \theta_m) \geqslant \delta(\theta_{m+1} \mid \theta_m) \geqslant \delta(\theta_m \mid \theta_m) = 0. \tag{2.3.8}$$

2.4 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.

2.5 Sources

Les notes de cours d'Etienne Birmelé et de Frédéric Santos.

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