Contrôle des connaissances du cours « Introduction aux méthodes de traitement des données » du DEA de Strasbourg.

Énoncé et Corrigé

Le jeudi 15 janvier 2004 de 14h à 16h.

Le contrôle est noté sur 20, le barème est indiqué à la fin du problème. Le polycopié et les notes de cours sont les seuls documents autorisés.

Estimation de paramètres.

Le but de cet exercice est de montrer comment la méthode standard d'estimation de paramètres doit être modifiée lorsque le nombre de photons détectés est faible.

Un détecteur forme le spectre d'un objet en produisant le nombre de photons détectés par canal d'énergie n_i, avec $1 \leq i \leq N$.

Ce nombre résulte de la convolution du spectre réel de la source avec la fonction d'appareil de l'instrument, qu'il peut être difficile, voire impossible d'inverser. On compare alors le résultat de l'observations aux prédictions d'un modèle spectral, c'est-à-dire à ce que l'on attendrait si le spectre de la source avait une forme donnée (par exemple une loi de puissance ou un spectre de corps noir), et on ajuste les paramètres de façon à obtenir le meilleur accord entre observations et prédiction du modèle.

Pour ce faire, on forme souvent la quantité :

$$S = \sum_{i=1}^{N} \frac{(n_i - e_i)^2}{e_i}$$

où les e_i, avec $1 \leq i \leq N$ sont les prédictions du modèle.

① A quelle loi devraient obéir les n_i ? On précisera les conditions de validité.

② Si $< n_i >$ est grand, donner une approximation de cette loi.

③ Dans ces conditions, donner la loi suivie par S sous l'hypothèse que le modèle est vrai.

En astrophysique des hautes énergies, il arrive fréquemment que les e_i soient petits, voire très petits devant 1.

④ Dire alors pourquoi S devient inadéquat pour estimer le meilleur ajustement. On pourra par exemple examiner le cas où on teste un modèle de spectre plat, avec tous les e_i égaux à une même valeur e

On utilise alors souvent une autre quantité :

$$C = 2(\sum_{i=1}^{N} e_i - \sum_{i=1}^{N} n_i \ln e_i)$$

⑤ Dans la limite où les n_i sont grands et l'ajustement est bon, quel est l'ordre de grandeur de $d_i = n_i - e_i$?

⑥ Montrer que si les n_i sont grands,

$$C = 2\sum_{i=1}^{N}(n_i - n_i \ln n_i) + \sum_{i=1}^{N} \frac{(n_i - e_i)^2}{e_i} + O(\frac{1}{\sqrt{n_i}})$$

⑦ Que peut-on dire dans ces conditions des minima de C et de S lorsqu'on fait varier les paramètres du modèle ?

⑧ Montrer que, indépendamment de la valeur de n_i, la probabilité d'obtenir le résultat $(n_1, \ldots n_N)$ est

$$P = \prod_{i=1}^{N} \frac{e_i^{n_i} e^{-e_i}}{n_i!}$$

A quelles conditions cela est-il vrai ?

⑨ Quel est alors le lien entre P et C ? A quoi correspond le minimum de C ?

($0.5 + 0.5 + 1 + 1 + 1 + 1 + 1 + 0.5 + 0.5 = 7$ points)

Corrigé.

❶ Sans autre information, on s'attend à ce que les n_i suivent une loi de poisson et soient indépendants. Il faut pour celà qu'on parle bien de photons (et non pas par exemple de photo-électrons), qu'il n'y ait pas de corrélation au niveau de la source (hypothèse très raisonnable), ni introduites par le détecteur lui même (par exemple par un long temps mort après la détection d'un photon, ce qui peut arriver).

❷ Si la moyenne $< n_i >$ est grande, la loi de Poisson peut être approximée par la loi normale, de moyenne $< n_i >$ et d'écart type $\sqrt{< n_i >}$.

❸ S suit alors la loi du χ^2 à N degrés de liberté

❹ Si les e_i sont très petits, alors S ne suit bien sûr plus la loi du χ^2. Ce qui est plus grave est que le minimum de S ne correspond plus au bon ajustement ; dans le cas où les n_i valent 0 ou 1, on favorisera de façon excessive les modèles prédisant e_i faible là où n_i vaut 0. Considérons par exemple le cas où le spectre est réellement plat et modélisé par $e_i = e$; si le nombre de photons détectés est αN, alors on voit facilement que

$$S = N[\frac{\alpha}{e} + e - 2\alpha]$$

qui est minimum pour $e = \sqrt{\alpha}$ au lieu de α comme on l'attendrait. L'écart peut être considérable.

❺ On s'attend à d_i de l'ordre de grandeur de l'erreur sur n_i, soit $d_i \sim \sqrt{n_i}$

❻ On peut écrire :

$$\begin{aligned} C &= 2\sum n_i - d_i - n_i \ln(n_i - d_i) \\ &= 2\sum n_i - d_i - n_i[\ln n_i - d_i/n_i - 1/2(d_i/n_i)^2 + O(n_i^{-3/2})] \end{aligned}$$

qui conduit au résultat demandé.

❼ On voit alors que $C \sim F(n_i) + S$; $F(n_i)$ ne varie pas quand on fait varier les paramètres du modèle, et donc les minima de S et C sont identiques.

❽ Ceci résulte de la première question (variables poissonniennes indépendantes).

❾ On voit que

$$-2\ln P = 2\sum(e_i - n_i \ln e_i) + \sum \ln n_i!$$

et que minimiser C revient à maximiser P (les n_i ne changent pas).

Composition de variables aléatoires

① Si X_1 et X_2 sont deux variables aléatoires indépendantes suivant des lois exponentielles de paramètres λ_1 et λ_2 respectivement, donner la densité de probabilité de la loi suivie par $X_1 + X_2$.

② Même question si X_1 et X_2, toujours indépendantes, suivent des lois du χ^2 à n_1 et n_2 degrés de liberté.

($1 + 1 = 2$ points)

Corrigé.

❶ Une application du cours donne

$$f(x) = \int_0^x f_1(t) f_2(x-t) dt$$

avec $f_1(t) = \lambda_1 e^{-\lambda_1 t}$ et $f_2(t) = \lambda_2 e^{-\lambda_2 t}$, lorsque x est positif, et 0 sinon. On trouve alors, toujours si $x \geq 0$:

$$f(x) = \frac{\lambda_1 \lambda_2}{\lambda_2 - \lambda_1}(e^{-\lambda_1 x} - e^{-\lambda_2 x})$$

si $\lambda_1 \neq \lambda_2$, et $f(x) = \lambda_1^2 x e^{-\lambda_1 x}$ sinon.

❷ C'est une simple question de cours. $X_1 + X_2$ suit une loi du χ^2 à $n_1 + n_2$ degrés de liberté.

Généralisation de la méthode des moments

Rappel : La méthode des moments résulte du remplacement de la fonction de répartition vraie F par la fonction de répartition empirique F_n dans une fonctionnelle faisant intervenir F. Prenons par exemple l'expression : $\int x^k dF$ qui est égale au moment non-centré μ'_k de la loi de fonction de répartition F. Remplaçons F par F_n, on obtient alors l'intégrale (de Stieljes) $\int x^k dF_n$. Lorsque F_n est une fonction en escalier, ce qui est bien le cas ici, cette intégrale est égale à une somme portant sur les valeurs de l'intégrand aux points de discontinuités de F_n pondérées par la valeur du saut. On obtient ici : $\int x^k dF_n = \sum_i x_i^k \frac{1}{n} = \frac{1}{n}\sum_i x_i^k$. Si le ou les paramètres de la loi sont désignés par θ on obtient une équation permettant de les déterminer en écrivant que le moment vrai est égal au moment empirique. Soit : $\mu'_k(\theta) = \frac{1}{n}\sum_i x_i^k$. La convergence de l'estimateur $\frac{1}{n}\sum_i X_i^k$ résulte de la loi des grands nombres généralisée aux fonctionnelles de F.

Mais ce qui a pu être fait pour l'intégrand x^k peut être aussi fait pour un autre intégrand bien choisi. Le but de cet exercice est de mettre en œuvre cette idée sur un exemple conduisant à des calculs relativement simples.

① On choisit comme intégrand la fonction « signe de x » $\operatorname{sgn}(x)$ ainsi définie :

$$\operatorname{sgn}(x) = \begin{cases} +1 & \text{si } x \geq 0\,; \\ -1 & \text{si } x < 0\,. \end{cases}$$

Dans l'exemple que nous traiterons, on négligera le cas $x = 0$ comme étant de probabilité nulle.
Soit F la fonction de répartition d'une variable aléatoire X continue et définie sur $(-\infty, +\infty)$, calculez $\int_{-\infty}^{+\infty} \operatorname{sgn}(x) dF$.

② Considérons un échantillon IID[1] de taille n : $(X_1, \ldots, X_n)$, ayant pour valeurs observées : $(x_1, \ldots, x_n)$. Calculez maintenant $\int_{-\infty}^{+\infty} \operatorname{sgn}(x) dF_n$.

③ On pose $S = \int_{-\infty}^{+\infty} \operatorname{sgn}(x) dF_n$, cette variable aléatoire est-elle discrète ou continue ? Quel est son domaine de définition ? On négligera, comme il est dit ci-dessus, le cas où une variable aléatoire X_i prend la valeur $x_i = 0$.

④ Quelle est la probabilité associée au cas où $S = k/n$? le symbole k désigne un entier tel que k/n appartient au domaine de définition. (nota : la suite de l'exercice ne dépend pas de la réponse à cette question)

⑤ On considère maintenant le cas de la loi uniforme entre a et b c'est-à-dire que toutes les variables aléatoires X_i possèdent la fonction de répartition :

$$F(x) = \begin{cases} 0 & \text{si } x < a\,; \\ \dfrac{x-a}{b-a} & \text{si } x \geq a, x \leq b\,; \\ 1 & \text{si } x > b\,. \end{cases}$$

On suppose que a est négatif et que b est positif. Quelle est la valeur de $F(0)$?

⑥ Trouver un estimateur de $\dfrac{a}{b}$ en écrivant que $\int_{-\infty}^{+\infty} \operatorname{sgn}(x) dF = \int_{-\infty}^{+\infty} \operatorname{sgn}(x) dF_n$.

⑦ Cet estimateur possède-t-il une moyenne ?

⑧ On prend maintenant l'exemple de la loi de Cauchy, de fonction de répartition :

$$\begin{aligned} F(x) &= \int_{-\infty}^{x} \frac{1}{\pi} \frac{1}{b} \frac{1}{1 + \left(\dfrac{u-a}{b}\right)^2} du\,, \\ &= \frac{1}{\pi} \arctan\left(\frac{x-a}{b}\right) + \frac{1}{2}\,. \end{aligned}$$

Par la même méthode, trouver un estimateur de $\dfrac{b}{a}$.

$(1,5 + 1,5 + 1,5 + 2 + 1 + 0,5 + 1 + 2 = 11$ points$)$

Corrigé.

❶ On a :

$$\begin{aligned} \int_{-\infty}^{+\infty} \operatorname{sgn}(x) dF &= \int_{-\infty}^{0} \operatorname{sgn}(x) dF + \int_{0}^{+\infty} \operatorname{sgn}(x) dF\,, \\ &= -\int_{-\infty}^{0} dF + \int_{0}^{+\infty} dF\,, \\ &= -(F(0) - 0) + (1 - F(0))\,, \\ &= 1 - 2F(0)\,. \end{aligned}$$

[1] IID veut dire « indépendant et identiquement réparti.»

❷ On a :

$$\int_{-\infty}^{+\infty} \operatorname{sgn}(x) dF_n = \sum_{i=1}^{n} \operatorname{sgn}(x_i) \frac{1}{n} ,$$
$$= \frac{1}{n} \sum_{i=1}^{n} \operatorname{sgn}(x_i) .$$

Ce sont les valeurs observées de la variable aléatoire : $\frac{1}{n} \sum_{i=1}^{n} \operatorname{sgn}(X_i)$.

❸ C'est une variable aléatoire discrète qui correspond à l'excédent d'un signe par rapport à l'autre le tout divisé par n. Il y a n cas possibles : de 0 valeurs positives à n valeurs positives par pas de 1. Les valeurs possibles de S vont donc de -1 (tous les X_i sont négatifs) à $+1$ (tous les X_i sont positifs) par pas de $2/n$.

❹ Si p désigne le nombre de valeur positives et q le nombre de valeurs négatives, on a : $p + q = n$ et $p - q = k$ d'où $k = 2q - n$. Pour un k donné on doit avoir exactement $q = \frac{1}{2}(k+n)$ valeurs positives ce qui arrive avec la probabilité donnée par la loi binômiale :

$$\Pr(S = k/n) = \mathrm{C}_n^q F(0)^q (1 - F(0))^{n-q} \quad \text{avec} \quad q = \tfrac{1}{2}(k+n) .$$

Où C_n^q désigne le nombre de façons de choisir q éléments parmi n (coefficient du binôme).

❺ On trouve $F(0) = -a/(b-a)$

❻ d'où $1 - 2F(0) = \dfrac{b+a}{b-a}$ et en posant $S = \frac{1}{n} \sum_{i=1}^{n} \operatorname{sgn}(X_i)$, on trouve :

$$\frac{b}{a} = \frac{S+1}{S-1} .$$

❼ Non, car pour $S = 1$, il prend la valeur $+\infty$ avec une probabilité donnée par la loi du binôme et qui n'est pas nulle. Ce résultat réduit considérablement l'utilité de cet estimateur dans ce cas particulier.

❽

$$F(0) = -\frac{1}{\pi} \arctan\Big(\frac{a}{b}\Big) + \frac{1}{2} .$$

On a : $1 - 2F(0) = \dfrac{2}{\pi} \arctan\Big(\dfrac{a}{b}\Big)$, d'où :

$$\frac{2}{\pi} \arctan\Big(\frac{a}{b}\Big) = \frac{1}{n} \sum_{i=1}^{n} \operatorname{sgn}(x_i) .$$

Ce qui conduit à l'estimateur :

$$\frac{a}{b} = \tan\Big(\frac{\pi}{2n} \sum_{i=1}^{n} \operatorname{sgn}(x_i)\Big) .$$

Expression qui pose encore des problèmes si $S = 1$.

Notez que l'exemple de la fonction $\operatorname{sgn}(x)$ donné dans cet exercice ne conduit pas à des estimateurs très utiles. Cette fonction a été choisie afin de simplifier les calculs mais le principe de la méthode reste intéressant.