ESTIMATEUR PAR HISTOGRAMME : RISQUE

Sous de conditions de régularité de la densité f, et en supposant de définir ν en fonction de la taille de l'échantillon N de sorte que $\nu_N \longmapsto 0$ quand $N \longmapsto +\infty$, nous pouvons démontrer le résultat asymptotique suivant :

$$\mathrm{MISE}_f(\nu) = \underbrace{\frac{\nu^2}{12}\int_{\mathrm{support}} \left(f'(x)\right)^2 dx + \frac{1}{N\nu}}_{\textbf{Terme principale du risque}} + \underbrace{\mathcal{O}\left(\nu^2\right) + \mathcal{O}\left(\frac{1}{N}\right)}_{\textbf{Terme résiduel}} \quad \text{lorsque} \quad N \to \infty$$

Le terme principale du risque est minimisé pour : $$\boxed{\nu_N^{\mathrm{opt}} = \left(\frac{N}{6}\int_{\mathrm{support}} \left(f'(x)\right)^2 dx\right)^{-1/3}}$$

Même si cette fenêtre optimale ne peut pas être déterminée précisément (car f est inconnue), ce résultat nous permet de conclure que lorsque N est grand, la fenêtre optimale ν_N doit être de l'ordre de $N^{-1/3}$

ESTIMATEUR PAR HISTOGRAMME : FENÊTRE OPTIMALE PAR VALIDATION CROISÉE

Nous cherchons maintenant à estimer la fenêtre optimale de notre histogramme, en estimant le risque uniquement à partir des observations.

Nous cherchons à définir un estimateur $\hat{J}$ de $MISE_f(\nu) - ||f||_2^2$ qu'il soit sans biais. Pour toute densité f et tout ν. Soit :

$$\begin{aligned} J(\nu, x_1, \ldots, x_N) &= \mathrm{MISE}_f(\nu) - ||f||_2^2 \\ &= \frac{1}{N\nu} - \frac{N+1}{N\nu} \sum_{j=1}^{b} p_j^2 \end{aligned}$$

Afin que $\hat{J}$ estimateur de J soit sans biais, il suffit d'avoir un estimateur sans biais de p_j^2 pour tout j, en se rappelant que p_j corresponds à a la fréquence théorique des observations se situant dans le j-ème intervalle.

ESTIMATEUR PAR HISTOGRAMME : FENÊTRE OPTIMALE PAR VALIDATION CROISÉE

Un approche naïf consiste à estimer p_j^2 par $\hat{p}_j^2$, où $\hat{p}_j$ est la fréquence empirique, c'est-à-dire $\hat{p}_j \coloneqq C_j/N$, en suivant les notations vu plus tôt dans ce cours.

EXERCICE. Calculer le biais de $\hat{p}_j^2$

Rappel : $C_j \sim Binom(N, p_j)$

ESTIMATEUR PAR HISTOGRAMME : FENÊTRE OPTIMALE PAR VALIDATION CROISÉE

SOLUTION

Etant donné que $C_j \sim Binom(N, p_j)$, il en découle :

$$\mathbb{E}[\hat{p}_j] = p_j\,;\, \mathrm{Var}[\hat{p}_j] = \frac{p_j(1-p_j)}{N}$$

Et par conséquent :

$$\mathbb{E}[\hat{p}_j^2] = \mathrm{Var}[\hat{p}_j] + (\mathbb{E}[\hat{p}_j])^2 = p_j^2\left(1 - \frac{1}{N}\right) + \frac{p_j}{N}$$

QUESTION. L'estimateur $\hat{p}_j^2$ est-il un estimateur sans biais de de p_j^2 ?

ESTIMATEUR PAR HISTOGRAMME : FENÊTRE OPTIMALE PAR VALIDATION CROISÉE

Nous pouvons en déduire les observations suivantes :

- $\hat{p}_j^2$ est un estimateur biaisé de p_j^2
- $\hat{p}_j^2 - \hat{p}_j/N$ est un estimateur sans biais de $p_j^2(1-1/N)$
- D'où :

$$\tilde{p}_j^2 := \frac{\hat{p}_j^2 - \hat{p}_j/N}{1-1/N} = \frac{N}{N-1}\hat{p}_j^2 - \frac{1}{N-1}\hat{p}_j$$

est un estimateur sans biais de $\hat{p}_j^2$.

Ceci nous permet enfin de proposer l'estimateur suivant (rappel : $\sum_{j=1}^b \hat{p}_j = 1$) :

$$\hat{J}(\nu, x_1, \ldots, x_N) = \frac{2}{(N-1)\nu} - \frac{N+1}{(N-1)\nu}\sum_{j=1}^b \hat{p}_j^2, \quad \hat{p}_j = C_j/N$$

ESTIMATEUR PAR HISTOGRAMME : FENÊTRE OPTIMALE PAR VALIDATION CROISÉE

L'estimateur $\hat{J}$ peut être utilisé pour déterminer automatiquement la fenêtre optimale ν par la méthode de validation croisée :

1. Poser : $m = \min_i x_i; l = \max_i x_i - m, i = 1, \ldots, N$
2. Initialiser : $b_{CV} = 1, \hat{J}_{CV} = \hat{J}(l/1, x_1, \ldots, x_N)$
3. Tant que $b < N$:
   - Calculer J $\coloneqq \hat{J}(l/b, x_1, \ldots, x_N)$
   - Si $j < \hat{J}$:
     - $b_{CV} = b$
     - $\hat{J}_{CV} = J$
4. Et enfin : $\nu_{CV} = l/b_{CV}$

→ Reprendre le TP, **Partie II.**