Introduction à la statistique non paramétrique

Catherine MATIAS

CNRS, Laboratoire Statistique & Génome, Évry

http://stat.genopole.cnrs.fr/∼cmatias

Atelier SFDS
27/28 septembre 2012

Partie 3 - Estimation de densité : histogrammes, noyaux et projections

Plan partie 3

Plan partie 3

Contexte de l'estimation de densité (univariée)

- Observations : $X_1, \ldots, X_n$ v.a. i.i.d. réelles de fdr F et admettant une densité $f = F'$.
- But : estimer (à partir des observations) f en faisant le moins d'hypothèses possibles sur cette densité.
- Typiquement, on supposera que $f \in \mathcal{F}$ espace fonctionnel et on notera $\hat{f}_n$ un estimateur de f.

Objectifs

Obtenir des informations de nature géométrique sur la distribution des variables. Ex :

- Combien de modes ?
- Zones peu denses ? très denses ?

Notation

- On notera $\mathbb{E}_f$ l'espérance quand la densité des X_i vaut f.

Mesure de la qualité d'un estimateur : risque I

Ingrédients de la définition du risque

1) Distance sur $\mathcal{F}$ pour mesurer l'écart entre $\hat{f}_n$ et f. Ex :
   - $d(f,g) = \|f-g\|_p = [\int |f-g|^p]^{1/p}$, pour $p \geq 1$. Par exemple $p = 1$ ou 2.
   - $d(f,g) = \|f-g\|_\infty = \text{supess}_x |f(x)-g(x)|$.
   - $d(f,g) = |f(x_0)-g(x_0)|$ où x_0 fixé.
2) Définition d'une fonction de perte $\omega : \mathbb{R} \mapsto \mathbb{R}^+$ convexe, telle que $\omega(0) = 0$. Ex : $\omega : u \mapsto u^2$ fonction de perte quadratique.
3) L'erreur $w(d(\hat{f}_n, f))$ (par ex $d(\hat{f}_n, f)^2$) dépend de l'échantillon observé. On définit donc une fonction de risque

$$R(\hat{f}_n, f) = \mathbb{E}_f(\omega(d(\hat{f}_n, f))).$$

C'est en moyenne, l'erreur que l'on commet en estimant f par $\hat{f}_n$, pour la distance d et la perte ω.

Mesure de la qualité d'un estimateur : risque II

Exemples de fonctions de risque

- En prenant la distance $\mathbb{L}_2$ et la perte quadratique, on obtient le risque quadratique intégré : MISE = mean integrated squared error

$$R(\hat{f}_n, f) = \mathbb{E}_f \|\hat{f}_n - f\|_2^2 = \mathbb{E}_f \int_x (\hat{f}_n(x) - f(x))^2 dx.$$

- En prenant la distance ponctuelle en x_0 et la perte quadratique, on obtient le risque quadratique ponctuel en x_0 : MSE = mean squared error

$$R_{x_0}(\hat{f}_n, f) = \mathbb{E}_f |\hat{f}_n(x_0) - f(x_0)|^2.$$

Décomposition biais-variance du risque quadratique I

Décomposition "biais-variance" du MSE

$$R_{x_0}(\hat{f}_n, f) = \mathbb{E}_f[\hat{f}_n(x_0) - f(x_0)]^2$$
$$= \mathbb{E}_f[\mathbb{E}_f(\hat{f}_n(x_0)) - f(x_0)]^2 + \mathbb{E}_f[\hat{f}_n(x_0) - \mathbb{E}_f\hat{f}_n(x_0)]^2 + \text{dp}.$$

Or $[\mathbb{E}_f(\hat{f}_n(x_0)) - f(x_0)]^2$ est déterministe donc $\mathbb{E}_f$ disparaît, et le double produit vérifie

$$\text{dp} = 2\mathbb{E}_f\Big([\mathbb{E}_f(\hat{f}_n(x_0)) - f(x_0)][\hat{f}_n(x_0) - \mathbb{E}_f\hat{f}_n(x_0)]\Big) = 0.$$

Donc

$$R_{x_0}(\hat{f}_n, f) = |\mathbb{E}_f(\hat{f}_n(x_0)) - f(x_0)|^2 + \mathbb{E}_f|\hat{f}_n(x_0) - \mathbb{E}_f\hat{f}_n(x_0)|^2$$
$$= \text{Biais}^2 + \mathbb{V}\text{ar}(\hat{f}_n(x_0)).$$

Décomposition biais-variance du risque quadratique II

Décomposition "biais-variance" du MISE

$$R(\hat{f}_n, f) = \mathbb{E}_f \|\hat{f}_n - f\|_2^2 = \mathbb{E}_f \int_x |\hat{f}_n(x) - f(x)|^2 f(x)\, dx$$
$$= \mathbb{E}_f \|\mathbb{E}_f(\hat{f}_n) - f\|_2^2 + \mathbb{E}_f \|\hat{f}_n - \mathbb{E}_f(\hat{f}_n)\|_2^2 + \text{dp}.$$

Or $\|\mathbb{E}_f(\hat{f}_n) - f\|_2^2$ est déterministe donc $\mathbb{E}_f$ disparaît,
et le double produit vérifie

$$\text{dp} = 2\mathbb{E}_f\Big(< \hat{f}_n - \mathbb{E}_f(\hat{f}_n), \mathbb{E}_f(\hat{f}_n) - f >_{\mathbb{L}_2(\mathbb{R})} \Big)$$
$$= 2 < 0, \mathbb{E}_f(\hat{f}_n) - f >_{\mathbb{L}_2(\mathbb{R})} = 0.$$

Donc

$$R(\hat{f}_n, f) = \|\mathbb{E}_f(\hat{f}_n) - f\|_2^2 + \mathbb{E}_f \|\hat{f}_n - \mathbb{E}_f(\hat{f}_n)\|_2^2 = \text{Biais}^2 + "\mathbb{V}\text{ar}(\hat{f}_n)".$$

Décomposition biais-variance du risque quadratique III

Compromis biais/variance

- L'étude du risque quadratique de l'estimateur se ramène donc à l'étude de son biais et de sa variance.
- On pourra accepter des estimateurs biaisés mais de variance petite, tels que le risque quadratique soit contrôlé.

Oracle

Idéalement, le meilleur estimateur, au sens du risque, est

$$f_n^\star = \underset{\hat{f}_n}{\mathrm{Argmin}}\, R(\hat{f}_n, f).$$

- Pbm : $R(\hat{f}_n, f)$ dépend de la densité f inconnue et n'est donc pas calculable. L'argmin $f_n^\star$ n'est pas un estimateur, c'est un oracle.
- Souvent, on dispose d'une famille d'estimateurs $\hat{f}_{\lambda,n}$ dépendants d'un paramètre λ (partition, fenêtre, etc . . .). L'oracle est donné par

$$\lambda^\star = \underset{\lambda}{\mathrm{Argmin}}\, R(\hat{f}_{\lambda,n}, f).$$

mais $\hat{f}_{\lambda^\star,n}$ n'est pas un estimateur. On veut sélectionner le meilleur λ à partir de l'étude du risque de $\hat{f}_{\lambda,n}$ et de sa dépendance en λ.

Contrôles du risque

Puisque f est inconnue, le risque $R(\hat{f}_n, f)$ au point f n'est pas calculable. Alternatives :

- Avoir recours à une méthode de validation croisée pour estimer ce risque au point f.
- S'intéresser au risque maximal sur une classe de fonctions $\mathcal{F}$. On introduit alors

$$R(\hat{f}_n, \mathcal{F}) = \sup_{f \in \mathcal{F}} R(\hat{f}_n, f).$$

  C'est un point de vue pessimiste, puisqu'en général les observations n'ont pas été générées sous le "pire des cas".
- En général, dans le second cas, on prend un point de vue asymptotique.

Contrôle asymptotiques du risque maximal I

Vitesses de convergence

- On veut construire un estimateur $\hat{f}_n$ tel que

$$R(\hat{f}_n, \mathcal{F}) = \sup_{f \in \mathcal{F}} R(\hat{f}_n, f) \to_{n \to \infty} 0,$$

- et exhiber la vitesse de convergence de $\hat{f}_n$ pour le risque R : la plus petite suite $(\phi_n)_{n \geq 0} \to 0$ telle que $\{\phi_n^{-1} R(\hat{f}_n, \mathcal{F})\}_n$ bornée :

$$\exists C > 0, \forall n \in \mathbb{N}, \forall f \in \mathcal{F}, \quad R(\hat{f}_n, f) \leq C \phi_n.$$

- On dit alors que $(\hat{f}_n)_n$ atteint la vitesse de convergence $(\phi_n)_n$ sur la classe $\mathcal{F}$ pour la distance d et la perte ω.

Contrôle asymptotiques du risque maximal II

Point de vue minimax

- Minimax : la recherche du meilleur estimateur pour le risque maximal, à classe $\mathcal{F}$ fixée. Le risque minimax est défini par

$$\inf_{\hat{f}_n} \sup_{f \in \mathcal{F}} R(\hat{f}_n, f).$$

S'il existe une suite $(\phi_n)_n$ telle que $\exists c, C > 0, \forall n \in \mathbb{N}$,

$$c\phi_n \leq \inf_{\hat{f}_n} \sup_{f \in \mathcal{F}} R(\hat{f}_n, f) \leq C\phi_n,$$

alors $(\phi_n)_n$ est la vitesse minimax.

- Typiquement, les classes de fonctions $\mathcal{F}$ pour lesquelles ont sait contrôler le risque minimax sont des classes de fonctions régulières. Comme par exemple : Lipschitz, classe $\mathcal{C}^k$, etc.

Contrôle asymptotiques du risque maximal III

Point de vue maxiset

- Maxiset : la recherche de la plus grande classe de fonctions $\mathcal{F}$ telle que le risque maximal sur $\mathcal{F}$ d'un estimateur $\hat{f}_n$ fixé décroît à une certaine vitesse

$$\sup_{\mathcal{F}}\{\mathcal{F}; \sup_{f\in\mathcal{F}} R(\hat{f}_n, f) \leq c\phi_n\}.$$

Voir [Cohen *et al.* 01, Kerkyacharian & Picard 02].

Remarque préliminaire à la construction d'estimateurs

Approche plug-in naïve

- On a vu que pour estimer $T(F)$, on pouvait utiliser $\hat{T}_n = T(\hat{F}_n)$ où $\hat{F}_n$ fdr empirique,
- Ici, la densité est la dérivée de la fdr : $f = F'$ d'où l'idée naïve de prendre $\hat{f}_n = \hat{F}'_n$.
- Pbm : $\hat{F}_n$ n'est pas dérivable. Cet estimateur plug-in n'est pas défini !

Plan partie 3

Plan partie 3

Histogramme I

On suppose que la densité f est définie sur un intervalle borné $[a, b] \subset \mathbb{R}$.

Définition

- Soit $I = (I_k)_{1 \leq k \leq D}$ une partition de $[a, b]$ (*i.e.* intervalles disjoints dont l'union est $[a, b]$),
- On note $n_k = \mathrm{Card}\{i; X_i \in I_k\}$ le nombre d'observations dans I_k, et $|I_k|$ la longueur de l'intervalle I_k.
- L'estimateur par histogramme de f est défini par

$$\hat{f}_{I,n}(x) = \sum_{k=1}^{D} \frac{n_k}{n|I_k|} 1_{I_k}(x).$$

- Il affecte à chaque intervalle une valeur égale à la fréquence des observations dans cet intervalle, renormalisée par la longueur de l'intervalle.

Histogramme II

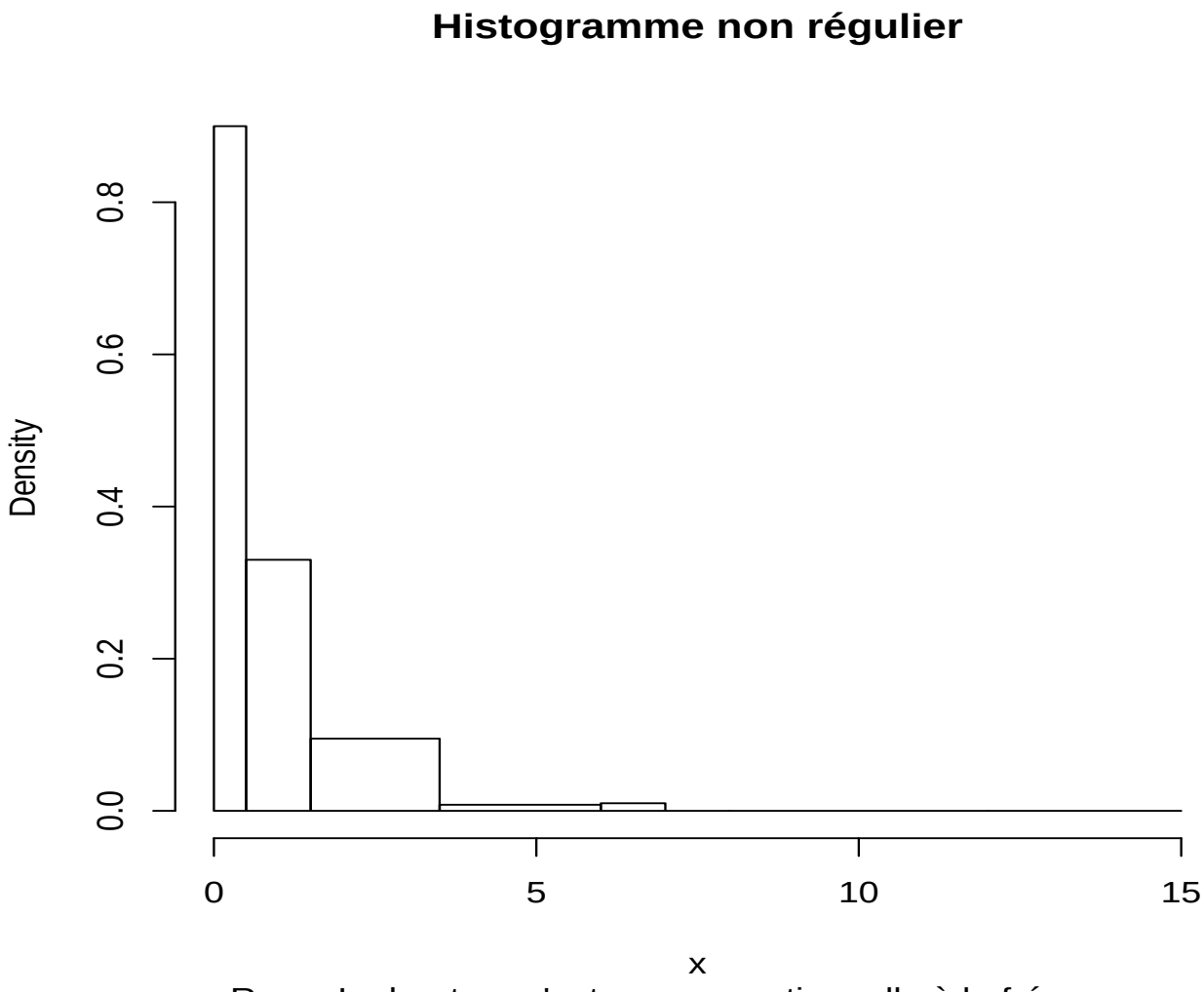
Histogrammes réguliers

- Un histogramme est dit régulier lorsque tous les intervalles I_k de la partition ont la même longueur.
- Dans ce cas, $|I|$ est aussi appelé fenêtre.
- Un histogramme régulier prend des valeurs proportionnelles à la fréquence des observations dans chaque intervalle.

Remarque

- L'histogramme est une fonction constante par morceaux. C'est donc une fonction très irrégulière. Cette notion de régularité n'a rien à voir avec la précédente ...

Histogramme III



Rem : La hauteur n'est pas proportionnelle à la fréquence

Risque quadratique des histogrammes I

Risque quadratique [Freedman and Diaconis 81]

- On suppose que f est deux fois dérivable, dans $\mathbb{L}_2([a,b])$, avec $f' \in \mathbb{L}_2([a,b])$ et $f'' \in \mathbb{L}_p([a,b])$ pour un certain $p \in [1,2]$,
- Alors on montre que pour des histogrammes réguliers, la fenêtre $|I|$ qui minimise le risque quadratique est de l'ordre de $O(n^{-1/3})$.
- De plus, pour ce choix de fenêtre, le risque quadratique de l'estimateur par histogramme décroît en $O(n^{-2/3})$.
- En particulier, asymptotiquement, si la fenêtre $|I|$ décroît comme $O(n^{-1/3})$, on obtient un estimateur consistant.

Risque quadratique des histogrammes II

Remarque

- La densité f est supposée à support borné $[a, b]$. En pratique, on observe des valeurs dans l'intervalle $[X_{(1)}, X_{(n)}]$ et on ne peut pas estimer f en dehors de ces bornes sans hypothèses de régularité supplémentaires.

Risque quadratique des histogrammes III

Considérations pratiques

- Ce résultat ne permet pas de choisir en pratique la fenêtre $|I|$.
- Règle empirique de Sturges :
  - Choisir le nombre de segments de la partition $D = 1 + \log_2 n$.
  - Règle empirique, fondée sur la loi normale, le TCL et le triangle de Pascal.
  - C'est la règle par défaut de la fonction *hist* dans R.
- La librairie *MASS* contient la fonction *truehist* qui est plus évoluée que *hist*. En particulier, on peut
  - contrôler la taille des intervalles ou bien leur nombre,
  - sélectionner automatiquement la partition avec des considérations de type [Freedman and Diaconis 81].
- Cependant, peut-on construire une règle moins empirique que Sturges et utilisable en pratique ?

Plan partie 3

Choix de la partition optimale

Minimisation du risque MISE et estimateur oracle

- On veut choisir la partition I qui minimise le risque quadratique intégré (MISE) $R(I, n, f) := \mathbb{E}_f \|\hat{f}_{I,n} - f\|_2^2$. Ansi

$$I^\star = \underset{I \in \mathcal{I}}{\text{Argmin}}\, R(I, n, f),$$

où $\mathcal{I}$ est l'ensemble des partitions de $[a, b]$.

- Pbm : Le MISE dépend de la densité inconnue f.

$$\underset{I \in \mathcal{I}}{\text{Argmin}}\, R(I, n, f) = \underset{I \in \mathcal{I}}{\text{Argmin}}\, \mathbb{E}_f \left[\|\hat{f}_{I,n}\|_2^2 - 2 \int_x \hat{f}_{I,n}(x) f(x)\, dx \right].$$

Donc $I^\star$ n'est pas un estimateur. On dit que c'est un oracle.

- On va donc estimer ce risque pour sélectionner une partition I.

Méthodes d'estimation du risque I

Validation croisée (V-fold cross validation)

- On découpe l'échantillon de départ $X_1, \ldots, X_n$ en V paquets $C_1, \ldots, C_V$ de même taille n/V,
- Pour $1 \leq v \leq V$,
  - Lorsque c'est possible : on construit l'estimateur $\hat{f}_I^v$ à partir de toutes les observations sauf le paquet C_v
  - On construit $\hat{R}^v(I)$ estimateur du risque $R(\hat{f}_I^v, f)$ de $\hat{f}_I^v$, à partir des observations du paquet C_v qui sont indépendantes des précédentes.
- On construit un estimateur du risque global $R(I, n, f)$ via

$$\hat{R}^{CV}(I) = \frac{1}{V} \sum_{v=1}^{V} \hat{R}^v(I).$$

Méthodes d'estimation du risque II

Variantes : Leave-one-out et leave-p-out

- Leave-one-out : validation croisée avec $V = n$, *i.e.* à chaque étape, on utilise $n - 1$ observations pour construire l'estimateur et l'observation restante permet d'estimer son risque.
- Leave-p-out : même principe que CV appliqué à tous les paquets possibles de p variables parmi n, *i.e.* à chaque étape, on utilise $n - p$ observations pour construire l'estimateur et les p observations restantes permettent d'estimer son risque.

Différence entre V-fold CV et leave-p-out

- Dans la validation croisée, chaque variable appartient à un paquet et un seul. En particulier, chaque variable n'est utilisée qu'une seule fois pour estimer le risque.

Estimation du MISE de l'estimateur par histogramme I

Mise en œuvre pour l'estimateur par histogramme

- Il reste donc à construire les estimateurs $\hat{R}^v(I)$ des risques $R(\hat{f}_I^v, f)$ de chaque estimateur $\hat{f}^v$. Or (on note $p = n/V$)

$$R(\hat{f}_I^v, f) = \mathbb{E}_f\left[\|\hat{f}_I^v\|_2^2 - 2\int_x \hat{f}_I^v(x) f(x)\,dx\right] + \text{cte}$$

$$= \int_x \sum_{k=1}^{D} \mathbb{E}_f\left(\frac{n_k^v}{(n-p)|I_k|}\right)^2 1_{I_k}(x)\,dx$$

$$- 2\int_x \sum_{k=1}^{D} \mathbb{E}_f\left(\frac{n_k^v}{(n-p)|I_k|}\right) 1_{I_k}(x) f(x)\,dx + \text{cte},$$

où $n_k^v = \text{Card}\{i \notin C_v; X_i \in I_k\}$.

- Il faut donc estimer $\mathbb{E}_f(n_k^v)$ et $\mathbb{E}_f[(n_k^v)^2]$.

Estimation du MISE de l'estimateur par histogramme II

- Or $\mathbb{E}_f(n_k^v) = (n-p)\mathbb{P}(X \in I_k)$ s'estime sur les observations dans C_v par : $(n-p)(n_k - n_k^v)/p$,
- Formule plus compliquée mais analogue pour $\mathbb{E}_f[(n_k^v)^2]$.
- Au final, on peut montrer par exemple pour l'estimateur leave-p-out [Celisse and Robin 08]

$$\hat{R}^{\mathrm{lpo}}(I) = \frac{2n-p}{(n-1)(n-p)} \sum_k \frac{n_k}{n|I_k|} - \frac{n(n-p+1)}{(n-1)(n-p)} \sum_k \frac{1}{|I_k|} \left(\frac{n_k}{n}\right)^2.$$

Estimation par histogramme avec partition LpO optimale

Procédure

- On se donne un ensemble de partitions $\mathcal{I}$ de $[a, b]$
  - Ex : $\mathcal{I} = \{I^{2^m}, m_{\min} \leq m \leq m_{\max}\}$ où I^N est la partition régulière de $[a, b]$ en intervalles de longueur $(b-a)/N$.
  - En pratique, $\mathrm{Card}(\mathcal{I})$ doit rester raisonnable pour pouvoir explorer toutes les partitions.
- Pour chaque $I \in \mathcal{I}$, on calcule l'estimateur LpO $\hat{R}^{\mathrm{lpo}}(I)$, on sélectionne

$$\hat{I} = \operatorname*{Argmin}_{I \in \mathcal{I}} \hat{R}^{\mathrm{lpo}}(I).$$

- On estime f par l'histogramme $\hat{f}_{\hat{I}}$.

Remarques

- Cet estimateur dépend encore du choix de $p \in \{1, \ldots, n-1\}$ utilisé pour la procédure LpO.
- Pourquoi s'arrêter là et ne pas sélectionner le meilleur p ?

Sélection automatique de p pour procédure LpO

Risque MISE et estimateur LpO

- L'estimateur $\hat{R}^{\text{lpo}}(I)$ dépend de p. On le note $\hat{R}_p(I)$.
- Le risque quadratique de l'estimateur $\hat{R}_p(I)$ est donné par

$$MSE(p, I) = \mathbb{E}_f \left[(\hat{R}_p(I) - R(I, n, f))^2 \right].$$

- Cette quantité dépend de f. [Celisse and Robin 08] ont montré que c'est une fonction $\Phi(p, I, \alpha)$ où $\alpha = (\alpha_1, \ldots, \alpha_D)$ et $\alpha_k = \mathbb{P}(X \in I_k)$. On peut donc l'estimer par $\Phi(p, I, (n_k/n)_{1 \leq k \leq D})$ et sélectionner

$$\hat{p}(I) = \underset{1 \leq p \leq n-1}{\text{Argmin}}\ \phi(p, I, (n_k/n)_{1 \leq k \leq D}).$$

Estimation adaptative par histogramme

(On parle d'estimation adaptative lorsque les paramètres optimaux -ex fenêtre- sont sélectionnés automatiquement à partir des observations).

Procédure

- On se donne un ensemble de partitions $\mathcal{I}$ de $[a, b]$
- Pour chaque $I \in \mathcal{I}$, on calcule
  - la valeur optimale de p

$$\hat{p}(I) = \underset{1 \leq p \leq n-1}{\text{Argmin}}\ \phi(p, I, (n_k/n)_{1 \leq k \leq D}),$$

  - puis l'estimateur leave-$\hat{p}(I)$-out $\hat{R}_{\hat{p}(I)}(I)$,
- et on sélectionne

$$\hat{I} = \underset{I \in \mathcal{I}}{\text{Argmin}}\ \hat{R}_{\hat{p}(I)}(I).$$

- On estime f par l'histogramme $\hat{f}_{\hat{I}}$.

Résultats supplémentaires

[Celisse and Robin 08] ont montré que

- La procédure leave-p-out est meilleure que V-fold CV pour estimer le risque quadratique des estimateurs par histogramme,
- Ils ont fourni des expressions du biais et de la variance de l'estimateur du risque.

Plan partie 3

Illustration : Estimation de la proportion d'hypothèses nulles pour les tests multiples [Celisse & Robin 10] I

Contexte

- Tests multiples : on veut tester simultanément un très grand nombre d'hypothèses. (ex : gènes différentiellement exprimés entre individus malades et sains)
- Les règles de type Bonferroni sont peu adaptées à la détection d'hypothèses non nulles.
- On modélise la distribution des statistiques de test par un mélange semi-paramétrique : $X_1, \ldots, X_m$ i.i.d de loi

$$g = \pi_0 g_0 + (1 - \pi_0) g_1,$$

où g_0= distribution de la stat de test sous H_0 est supposée exactement connue, π_0 est la proportion de statistiques qui suivent la loi H_0, et g_1 loi sous H_1 des X_i est inconnue.

Illustration : Estimation de la proportion d'hypothèses nulles pour les tests multiples [Celisse & Robin 10] II

- Quitte à transformer les données, on peut se ramener aux p-values $p_1, \ldots, p_m$ i.i.d sur $[0,1]$ de loi

$$\forall x \in [0,1], \quad f(x) = \pi_0 + (1-\pi_0) f_1(x),$$

où π_0 est la proportion de statistiques qui suivent la loi H_0, et f_1 loi sous H_1 des p_i est inconnue.

- Il est naturel de supposer que f_1 s'annule au voisinage de 1, mais sur les données réelles, on observe que la distribution des p-values croît au voisinage de 1 (modèle mal spécifié).
- On cherche à estimer π_0, sous l'hypothèse que f_1 s'annule sur un intervalle $[\lambda^\star, \mu^\star] \subset [0,1]$, où $\lambda^\star, \mu^\star$ sont inconnus.
- Sous cette hypothèse, $\pi_0 = f(x)$ pour tout $x \in [\lambda^\star, \mu^\star]$.

Illustration : Estimation de la proportion d'hypothèses nulles pour les tests multiples [Celisse & Robin 10] III

Approche [Celisse & Robin 10]

- Construire un estimateur $\hat{f}_m$ par histogramme de la densité des observations $p_1, \ldots, p_m$ en utilisant la stratégie développée dans [Celisse and Robin 08] (LpO) sur l'ensemble de partitions $\mathcal{I}_N = \left\{ I = (I_i)_i : \forall i \neq k+1, |I_i| = \frac{1}{N},\ |I_{k+1}| = \frac{l-k}{N},\ 2 \leq k+2 \leq l \leq N \right\}$.
- Estimer π_0 par la valeur de $\hat{f}_m$ sur l'intervalle $[\hat{\lambda}, \hat{\mu}]$ de l'histogramme.

Illustration : Estimation de la proportion d'hypothèses nulles pour les tests multiples [Celisse & Robin 10] IV

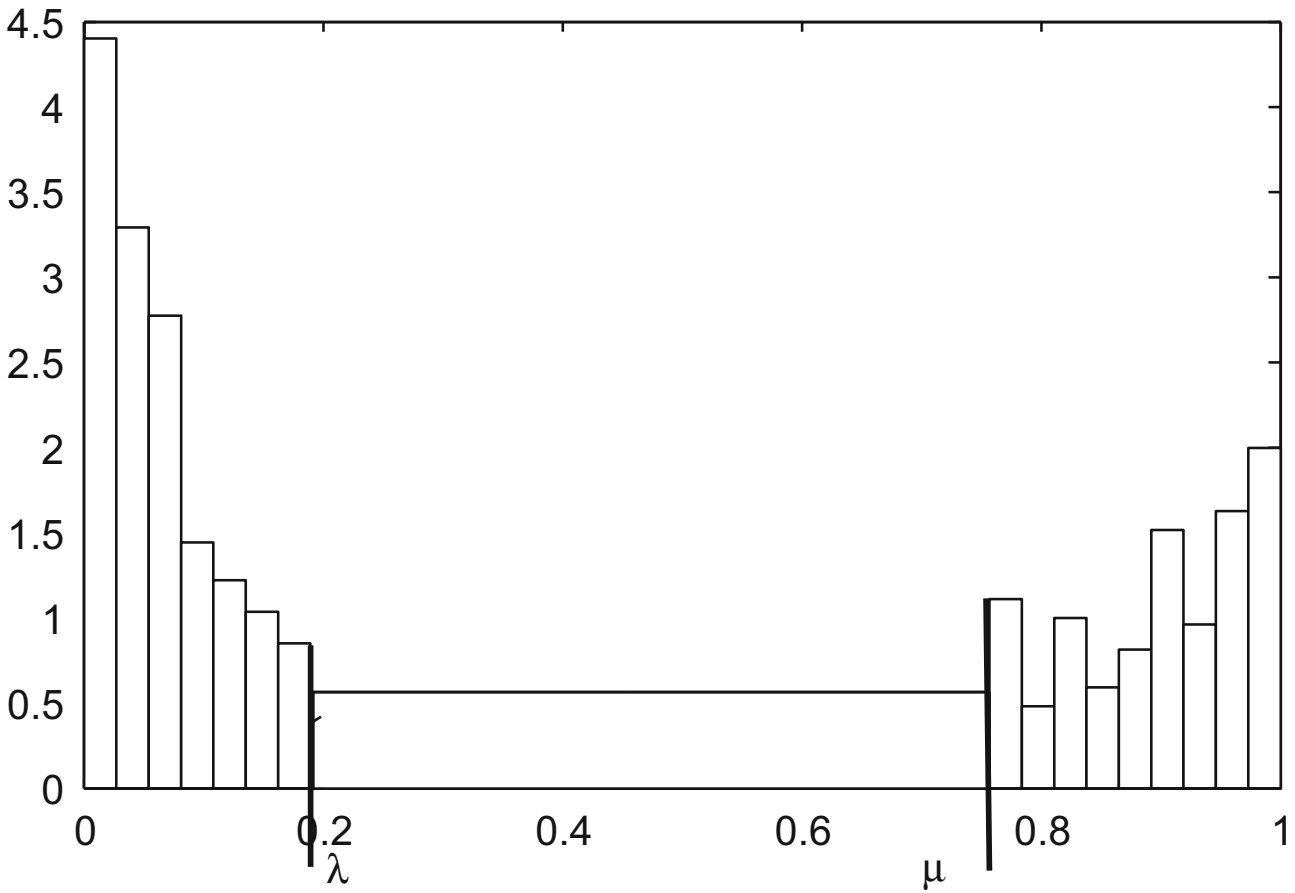


Illustration : Estimation de la proportion d'hypothèses nulles pour les tests multiples [Celisse & Robin 10] V

Propriétés théoriques de l'estimateur

Estimateur consistant (quand $m \to +\infty$) et asymptotiquement normal [Nguyen and Matias 12].

Performances sur simulations

Comparaison des performances de plusieurs procédures : histogrammes +LOO, histogrammes +LPO, estimateurs de Storey (2 choix différents de paramètres).

Illustration : Estimation de la proportion d'hypothèses nulles pour les tests multiples [Celisse & Robin 10] VI

$\pi_0 = 0.9$	$\lambda^* = 0.2$, s=4		
	Bias	Std	MSE
LPO	**0.39**	2.5	$\mathbf{6.41} \times 10^{-2}$
LOO	**0.46**	2.3	5.52×10^{-2}
$\widehat{\pi}_0^{St}$	-0.15	3.2	$\mathbf{9.94} \times 10^{-2}$
$\widehat{\pi}_0^{St}$ Th.	0	3.0	$\mathbf{9.00} \times 10^{-2}$

(toutes les quantités sont multipliées par 100).

Plan partie 3

Plan partie 3

Construction des estimateurs à noyaux I

Retour aux estimateurs plug-in

Pour h assez petit, on a

$$f(x) = F'(x) \simeq \frac{F(x+h) - F(x-h)}{2h}.$$

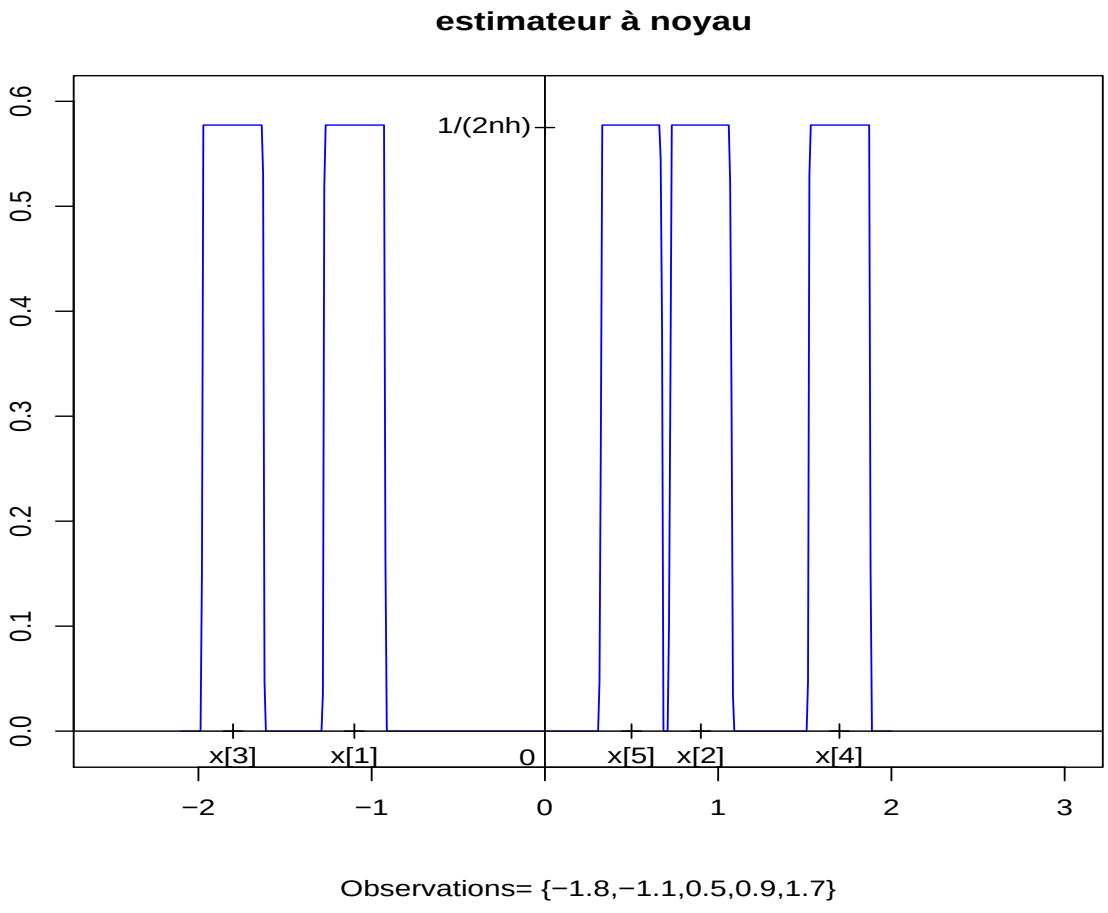
D'où l'estimateur plug-in (non naïf)

$$\begin{aligned}\hat{f}_{n,h}(x) &= \frac{\hat{F}_n(x+h) - \hat{F}_n(x-h)}{2h} \\ &= \frac{1}{n}\sum_{i=1}^{n} \frac{1}{2h} 1\{X_i \in]x-h; x+h]\} = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h} K_0\left(\frac{X_i - x}{h}\right),\end{aligned}$$

où $K_0(u) = 1_{]-1;1]}(u)/2$ est le noyau de Rosenblatt (1956).

Construction des estimateurs à noyaux II

Estimateur à noyau (rectangulaire)



Parzen (1962), propose de remplacer K_0 par un noyau plus général.

Définition des estimateurs à noyau I

Définition

- Soit $K : \mathbb{R} \to \mathbb{R}$ intégrable telle que $\int K(u)du = 1$. Alors K est appelé noyau.
- Pour tout $h > 0$ petit (en fait $h = h_n \to_{n\to\infty} 0$), on peut définir

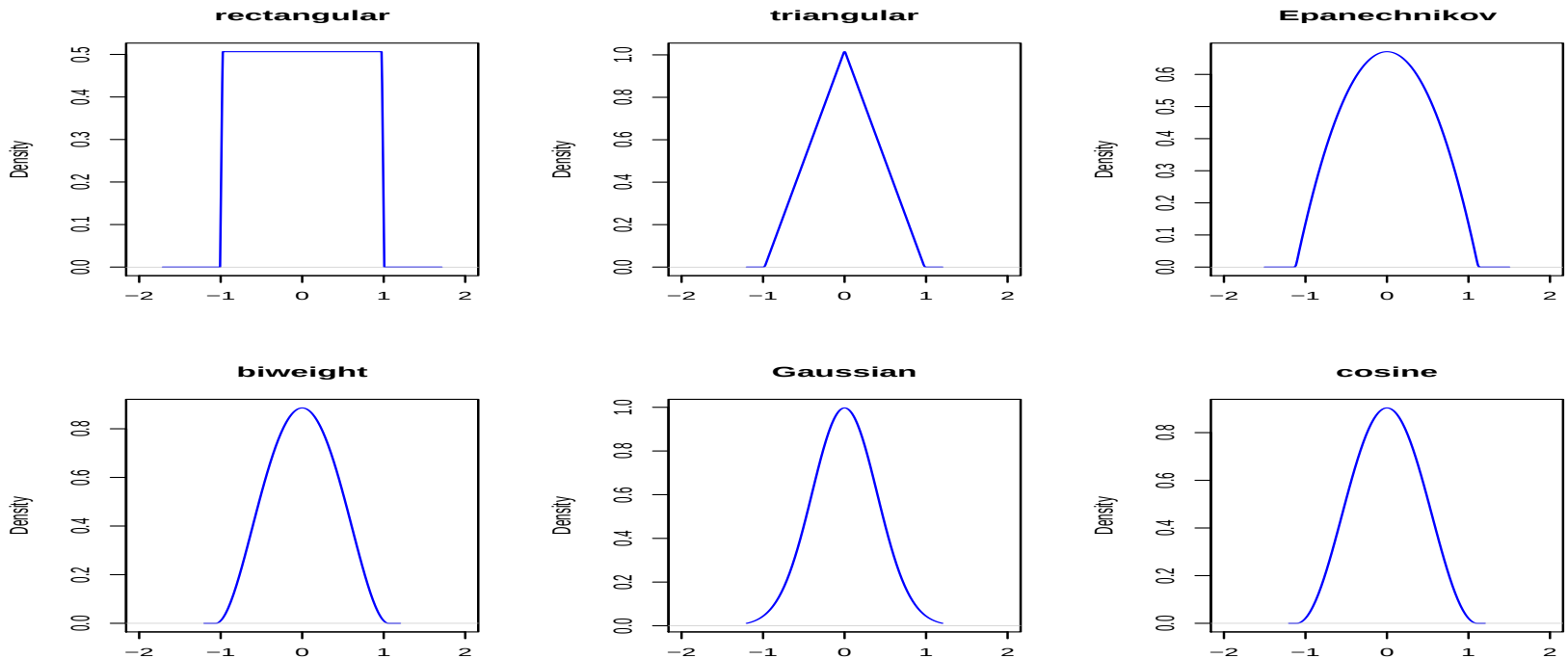
$$\hat{f}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h} K\left(\frac{X_i - x}{h}\right),$$

  estimateur à noyau de f. On a $\int \hat{f}_n(x)dx = 1$ et si $K > 0$ alors $\hat{f}_n$ est une densité.
- Le paramètre $h > 0$ est appelé fenêtre. C'est un paramètre de lissage : plus h est grand, plus l'estimateur est régulier.

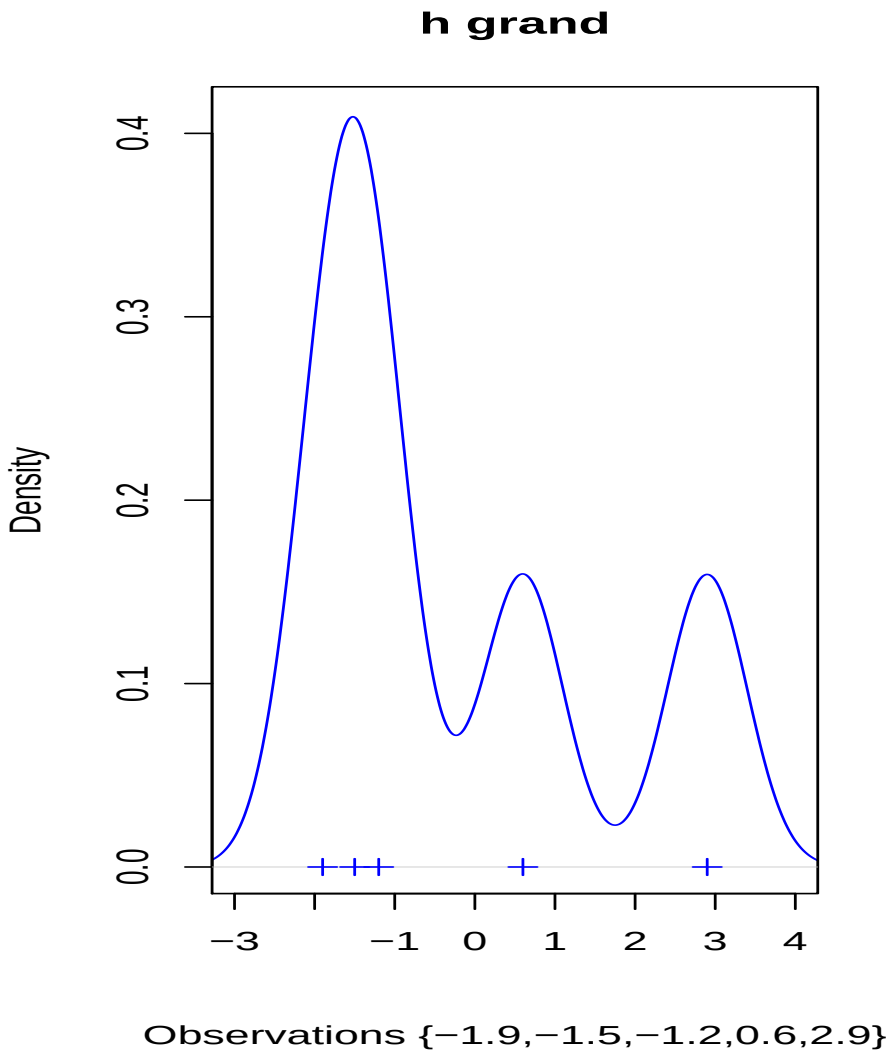
Rem : On considérera souvent des noyaux positifs et pairs, mais ce n'est pas obligatoire.

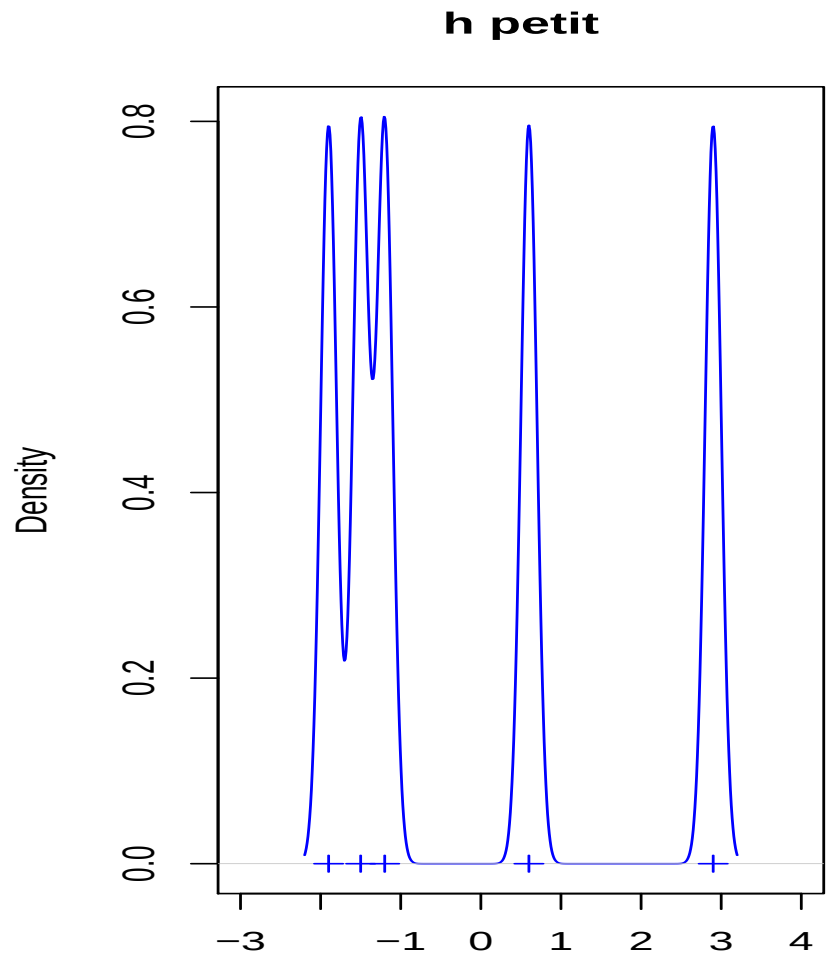
Exemples de noyaux

- Rosenblatt, ou noyau rectangulaire $K(u) = 1_{[-1;1]}(u)/2$.
- Noyau triangle $K(u) = (1 - |u|)1_{[-1;1]}(u)$.
- Epanechnikov $K(u) = \frac{3}{4}(1 - u^2)1_{[-1;1]}(u)$.
- Biweight $K(u) = \frac{15}{16}(1 - u^2)^2 1_{[-1;1]}(u)$.
- Gaussien $K(u) = \frac{1}{\sqrt{2\pi}} \exp(-u^2/2)$.
- Cosine $K(u) = \frac{\pi}{4} \cos(u\pi/2)1_{[-1;1]}(u)$.



Effet de la variation de h sur l'estimateur à noyau





Mise en perspective histogrammes/noyaux

- Dans l'estimateur par histogramme, on calcule la fréquence des observations dans des intervalles fixés à l'avance,
- Dans l'estimateur à noyau rectangulaire, on calcule la fréquence des observations dans une fenêtre glissante.
- Dans l'estimateur à noyau gaussien, toutes les observations sont prises en compte : celles qui sont proches du point x où on estime la densité ont un poids plus important que les autres.

Biais des estimateurs à noyau I

Rappel sur le risque quadratique ponctuel

$$R_x(\hat{f}_n, f) = \mathbb{E}_f(\hat{f}_n(x) - f(x))^2 = \text{Biais}_f^2(\hat{f}_n(x)) + \mathbb{V}\text{ar}_f(\hat{f}_n(x)).$$

Étude du biais : principe

$$\mathbb{E}_f(\hat{f}_n(x)) = \mathbb{E}_f\left(\frac{1}{h}K\left(\frac{X-x}{h}\right)\right) = \int \frac{1}{h}K\left(\frac{u-x}{h}\right) f(u)\,du$$
$$= \int K(v) f(x+hv)\,dv.$$

Si f est une fonction dérivable au voisinage de x, alors on peut écrire $f(x+hv) = f(x) + hvf'(x+\xi hv)$, où $\xi \in]0;1[$. D'où

$$\mathbb{E}_f(\hat{f}_n(x)) = \int K(v)[f(x) + hvf'(x+\xi hv)]\,dv$$
$$= f(x) + h\int vK(v)f'(x+\xi hv)\,dv.$$

Biais des estimateurs à noyau II

Si de plus $\|f'\|_\infty < +\infty$ et $\int |vK(v)|dv < \infty$, alors on obtient que

$$\mathbb{E}_f(\hat{f}_n(x)) = f(x) + O(h), \text{ lorsque } h \to 0.$$

Contrôle du biais

- Dans ce cas, on a montré que le biais $|\mathbb{E}_f(\hat{f}_n(x)) - f(x)|$ converge vers 0 lorsque $h \to 0$.
- Plus généralement, si on suppose que f appartient à une classe de fonctions suffisamment régulières, on va pouvoir montrer une décroissance du terme de biais vers 0.

Classe de Hölder (régularité locale) I

Définitions

- Pour tout $\beta \in \mathbb{R}$, on note $\lfloor \beta \rfloor$ le plus petit entier strictement inférieur à β.
- Pour tous $\beta > 0, L > 0$, on définit la classe des fonctions de Hölder sur l'ensemble T par

$$\Sigma(\beta, L) = \{f : T \to \mathbb{R}; f \text{ est } \ell = \lfloor \beta \rfloor \text{ fois dérivable et } \forall x, y \in T, |f^{\ell}(x) - f^{(\ell)}(y)| \leq L|x - y|^{\beta - \ell}\}.$$

- On note également $\Sigma_d(\beta, L)$ l'intersection entre $\Sigma(\beta, L)$ (pour $T = \mathbb{R}$) et l'ensemble des densités sur $\mathbb{R}$.

Classe de Hölder (régularité locale) II

Remarques

- Si $\beta \in]0;1]$ alors $\ell = 0$ et $\Sigma(\beta, L)$ est la classe des fonctions contractantes (ou Hölderiennes). De plus lorsque $\beta = 1$, on obtient les fonctions Lipschitziennes.
- Si $\beta \in]1;2]$ alors $\ell = 1$ et f' est contractante.

Noyaux d'ordre ℓ

Définition

Soit $\ell \in \mathbb{N}^\star$. Le noyau $K : \mathbb{R} \to \mathbb{R}$ est dit d'ordre ℓ si

- $\forall j \in \{1, \ldots, \ell\}$, on a $u \to u^j K(u)$ est intégrable,
- et $\forall j \in \{1, \ldots, \ell\}$, $\int u^j K(u) du = 0$.

Remarques

- Si K est un noyau pair alors K est d'ordre au moins 1.
- Pour $j = 0$, on a $\int u^j K(u) du = \int K(u) du = 1$.
- On sait construire des noyaux d'ordre ℓ pour tout entier $\ell \geq 1$.

Biais des estimateurs à noyaux sur la classe $\Sigma_d(\beta, L)$

Proposition

Si $f \in \Sigma_d(\beta, L)$ avec $\beta, L > 0$ et si K noyau d'ordre $\ell = \lfloor \beta \rfloor$ tel que $\int |u|^\beta |K(u)| du < +\infty$, alors pour tout $x \in \mathbb{R}$, tout $h > 0$ et tout entier $n \geq 1$ on a

$$\text{Biais}_f(\hat{f}_n(x)) = |\mathbb{E}_f(\hat{f}_n(x)) - f(x)| \leq \frac{L}{\ell!}\left(\int |u|^\beta |K(u)| du\right) h^\beta.$$

En particulier, le biais tend vers 0 lorsque $h \to 0$.

Variance des estimateurs à noyaux

On montre que

Proposition

Si f est une densité bornée sur $\mathbb{R}$ (i.e. $\|f\|_\infty < \infty$) et si K est un noyau tel que $\int K^2(u)du < +\infty$, alors pour tout $x \in \mathbb{R}$, pour tout $h > 0$ et tout $n \geq 1$, on a

$$\mathbb{V}\mathrm{ar}_f(\hat{f}_n(x)) \leq \frac{\|f\|_\infty(\int K^2(u)du)}{nh}.$$

Si de plus, $f(x) > 0$ et f continue au voisinage de x et $\int |K(u)|du < +\infty$, alors

$$\mathbb{V}\mathrm{ar}_f(\hat{f}_n(x)) = \frac{f(x)}{nh}(\int K^2(u)du)(1 + o(1)), \text{ lorsque } h \to 0.$$

Compromis biais/variance

Commentaires

- Si $nh \to \infty$ alors on aura $\mathbb{V}\mathrm{ar}_f(\hat{f}_n(x)) \to 0$. Donc on veut $h \to 0$ (à cause du biais), mais pas trop vite (*i.e.* $nh \to \infty$) : il ne faut pas sous-lisser.
- Sur la classe de Hölder $\Sigma_d(\beta, L)$, le biais de $\hat{f}_n(x)$ est en $O(h^\beta)$ et si la densité f est bornée, on sait contrôler sa variance. La question naturelle qui se pose alors est : les fonctions de $\Sigma_d(\beta, L)$ sont elles bornées ?

Lemme

Soit $\beta, L > 0$. Il existe une constante $M(\beta, L) > 0$ telle que $\forall f \in \Sigma_d(\beta, L)$, on a

$$\sup_{x \in \mathbb{R}} \sup_{f \in \Sigma_d(\beta, L)} f(x) \leq M(\beta, L).$$

Contrôle du risque quadratique ponctuel I

Théorème

Soit $\beta > 0, L > 0$ *et* K *un noyau d'ordre* $\ell = \lfloor \beta \rfloor$ *tel que* $\int K^2(u)du < +\infty$ *et* $\int |u|^\beta |K(u)|du < +\infty$. *Alors, en choisissant une fenêtre* $h = cn^{-1/(2\beta+1)}$, *avec* $c > 0$, *on obtient*

$$\forall x \in \mathbb{R}, \quad R_x(\hat{f}_n, \Sigma_d(\beta, L)) = \sup_{f \in \Sigma_d(\beta,L)} \mathbb{E}_f[|\hat{f}_n(x) - f(x)|^2]$$
$$\leq Cn^{-2\beta/(2\beta+1)},$$

où $C = C(c, \beta, L, K)$.

Contrôle du risque quadratique ponctuel II

Remarques

- L'estimateur $\hat{f}_n$ atteint la vitesse de convergence $\phi_{n,\beta} = n^{-2\beta/(2\beta+1)}$ sur la classe $\Sigma_d(\beta, L)$ pour le risque quadratique ponctuel maximal.
- En particulier, pour $\beta = 2$ (densité dérivable avec f' Lipschitz), on obtient la vitesse $n^{-4/5}$ pour le risque quadratique (ou $n^{-2/5}$ pour sa racine carrée).
- Le choix de la fenêtre optimale h dépend de β= régularité maximale de la densité f inconnue. Il peut paraître artificiel de supposer qu'on connaît β quand on ne connaît pas f. Il existe des méthodes d'estimation dites adaptatives qui n'utilisent pas la connaissance a priori de β. Voir [Lepski 92].

Risque quadratique intégré (MISE) I

Rappel

$$MISE(\hat{f}_n, f) = \mathbb{E}_f \|\hat{f}_n - f\|_2^2 = \mathbb{E}_f [\int (\hat{f}_n(x) - f(x))^2 dx]$$
$$= \mathbb{V}\mathrm{ar}_f(\hat{f}_n) + \mathrm{Biais}_f^2(\hat{f}_n),$$

où $\mathbb{V}\mathrm{ar}_f(\hat{f}_n) = \mathbb{E}_f \|\hat{f}_n - \mathbb{E}_f \hat{f}_n\|_2^2$ et $\mathrm{Biais}_f^2 = \|\mathbb{E}_f \hat{f}_n - f\|_2^2$.

Contrôle de la variance

Si $f \in \mathbb{L}_2(\mathbb{R})$, et si K noyau tel que $\int K^2(u)du < \infty$, alors pour toute fenêtre $h > 0$ et tout entier $n \geq 1$, on a

$$\mathbb{V}\mathrm{ar}_f(\hat{f}_n) = \mathbb{E}_f \|\hat{f}_n - \mathbb{E}_f \hat{f}_n\|_2^2 = \frac{1}{nh}(\int K^2(u)du)(1 + o(1)).$$

Risque quadratique intégré (MISE) II

Contrôle du biais

Pour contrôler le biais de cet estimateur, il faut introduire une classe de fonctions régulières. Ici, le contrôle souhaité étant global, on introduit une classe qui contrôle la régularité globale de la fonction f.

Classe de Nikol'ski (régularité globale)

Soient $\beta, L > 0$, on définit la classe de fonctions

$$\mathcal{N}(\beta, L) = \Big\{ f : \mathbb{R} \to \mathbb{R}, f \text{ est } \ell = \lfloor \beta \rfloor \text{ fois dérivable et } \forall t \in \mathbb{R},$$

$$\|f^{(\ell)}(\cdot + t) - f^{(\ell)}\|_2 = \Big(\int \Big(f^{(\ell)}(x+t) - f^{(\ell)}(x) \Big)^2 dx \Big)^{1/2} \leq L|t|^{\beta - \ell} \Big\}.$$

De plus, on note $\mathcal{N}_d(\beta, L)$ l'ensemble des densités qui sont dans la classe $\mathcal{N}(\beta, L)$.

Risque quadratique intégré (MISE) III

Proposition

Si $f \in \mathcal{N}_d(\beta, L)$ et si K est un noyau d'ordre $\ell = \lfloor \beta \rfloor$ tel que $\int |u|^\beta |K(u)| du < +\infty$, alors pour tout $h > 0$ et tout $n \geq 1$, on a

$$\text{Biais}_f^2 = \|\mathbb{E}_f \hat{f}_n - f\|_2^2 \leq \left(\frac{L}{(\ell)!} \int |u|^\beta |K(u)| du \right)^2 h^{2\beta}.$$

Contrôle du risque

La fenêtre optimale qui minimise le risque quadratique intégré est alors $h = cn^{-1/(2\beta+1)}$, et pour cette fenêtre, l'estimateur $\hat{f}_{n,h}$ vérifie

$$MISE(\hat{f}_n, \mathcal{N}_d(\beta, L)) = O(n^{-2\beta/(2\beta+1)}).$$

Risque quadratique des histogrammes versus noyaux I

Rappels sur les hypothèses

- Pour l'histogramme, on suppose le support de la densité borné, ce qui n'est pas le cas avec les noyaux.
- Les noyaux estiment des fonctions régulières (au sens local ou global).
- Dans la définition des classes de Hölder ou Nikol'ski, on suppose l'existence d'une constante $L > 0$ fixée qui majore les "normes" des densités.
- Les estimateurs à noyau utilisent la connaissance a priori de la régularité $\beta > 0$, mais il existe des méthodes adaptatives pour faire sans.

Risque quadratique des histogrammes versus noyaux II

Comparaison des résultats

- Le risque quadratique des histogrammes décroît en $n^{-2/3}$,
- Si la densité f est régulière (en fait dès que $\beta > 1$), le risque quadratique des estimateurs à noyau, qui décroît en $n^{-2\beta/(2\beta+1)}$ est plus rapide.

Vitesses minimax

En fait la vitesse $n^{-2\beta/(2\beta+1)}$ est une vitesse minimax

- pour le MSE sur la classe $\Sigma_d(\beta, L)$,
- pour le MISE sur la classe $\mathcal{N}_d(\beta, L)$.

Remarque

Les résultats sur les contrôles du risque ne donnent pas de règle pratique pour choisir la fenêtre h de l'estimateur.

Plan partie 3

Choix de la fenêtre optimale pour le risque MISE I

Rappel sur le MISE

$$
\begin{aligned}
MISE(h) &= \mathbb{E}_f \int [\hat{f}_{n,h}(x) - f(x)]^2 dx \\
&= \mathbb{E}_f \int \hat{f}^2_{n,h}(x)\, dx - 2\mathbb{E}_f \int f(x)\hat{f}_{n,h}(x)\, dx + cte, \\
\operatorname*{Argmin}_{h>0} MISE(h) &= \operatorname*{Argmin}_{h>0} \mathbb{E}_f \int \hat{f}^2_{n,h}(x)\, dx - 2\mathbb{E}_f \int f(x)\hat{f}_{n,h}(x)\, dx \\
&= \operatorname*{Argmin}_{h>0} J(h).
\end{aligned}
$$

- Comme J est inconnue (puisque dépend de f inconnue), on propose de l'estimer et de choisir la fenêtre $h > 0$ qui minimise son estimateur.
- Par validation croisée, on calcule $\hat{f}_{n,h}$ sur toutes les variables sauf un paquet, sur lequel on estime le risque.

Choix de la fenêtre optimale pour le risque MISE II

Stratégie d'estimation de J

- $\mathbb{E}_f \int \hat{f}_{n,h}^2(x)\,dx$ est estimé sans biais par $\int \hat{f}_{n,h}^2(x)\,dx$,
- Pour estimer sans biais $\mathbb{E}_f \int f(x)\hat{f}_{n,h}(x)\,dx$ on pourrait penser prendre $\int \hat{f}_{n,h}^2(x)\,dx$ mais ça ne marche pas (puisque que cette quantité est un estimateur sans biais de $\mathbb{E}_f \int \hat{f}_{n,h}^2(x)\,dx$).
- On remarque plutôt que

$$\mathbb{E}_f \int f(x)\hat{f}_{n,h}(x)\,dx \overset{\text{Fubini}}{=} \int f(x)\mathbb{E}_f[\hat{f}_{n,h}(x)]\,dx$$
$$= \int f(x)\frac{1}{h}\int K\left(\frac{u-x}{h}\right) f(u)\,du\,dx$$

et on introduit $\hat{T}_n = \frac{1}{n(n-1)h}\sum_i \sum_{j\neq i} K\left(\frac{X_i - X_j}{h}\right)$, estimateur sans biais de $\frac{1}{h}\int\int f(x)K\left(\frac{u-x}{h}\right) f(u)\,du\,dx$.

Choix de la fenêtre optimale pour le risque MISE III

Remarque

De façon très générale, il est important quand on estime par une somme double de la forme $\sum_{i,j} \phi(X_i - X_j)$, de la priver de sa diagonale $i \neq j$, sinon on augmente le biais. En effet, considérons par exemple

$$\tilde{T}_n = \frac{1}{n^2 h} \sum_{i=1}^{n} \sum_{j=1}^{n} K\left(\frac{X_i - X_j}{h}\right).$$

Alors la moyenne de $\tilde{T}_n$ fait apparaître un terme parasite :

$$\begin{aligned}
\mathbb{E}_f \tilde{T}_n &= \frac{1}{n^2 h} \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbb{E}_f K\left(\frac{X_i - X_j}{h}\right) \\
&= \frac{1}{nh} K(0) + \frac{n-1}{nh} \mathbb{E}_f K\left(\frac{X_1 - X_2}{h}\right).
\end{aligned}$$

Choix de la fenêtre optimale pour le risque MISE IV

Ainsi, on définit

$$\hat{J}_{n,h} = \frac{1}{V}\sum_{v=1}^{V}\left\{\int (\hat{f}^{v}_{n,h})^2(x)\,dx - \frac{2}{n(n-1)h}\sum_{\substack{i,j\in C_v\\ j\neq i}} K\left(\frac{X_i - X_j}{h}\right)\right\},$$

où $\hat{f}^{v}_{n,h}$ est calculé sur toutes les variables sauf celles du paquet C_v.
Puis

$$h^{CV} = \underset{h>0}{\text{Argmin}}\, \hat{J}_{n,h} \quad \text{et} \quad \hat{f}_n^{CV} \equiv \hat{f}_{n,h^{CV}}.$$

On obtient un estimateur à noyau qui est construit avec la fenêtre aléatoire h^{CV} (qui dépend des observations).

- On peut mq asymptotiquement, $\hat{f}_n^{CV}$ minimise en $h > 0$ le risque $MISE(h) = R(\hat{f}_{n,h}, f)$ pour la densité observée.
- Rem : on ne sait pas estimer le MSE par CV.

Plan partie 3

Illustration I

Données

On utilise la base de données *geyser* de la librairie *MASS* qui contient des données d'éruption (temps d'attente et durée) de l'"Old Faithful" geyser du Yellowstone National Park.

On commence par montrer la variabilité des histogrammes même lorsque la taille des intervalles est fixée mais les centres des intervalles changent.

```
> library(MASS)
> attach(geyser)
> par(mfrow=c(2,3))
> # histograms are sensitive to placing of bin centres
> truehist(duration,h=0.5,x0=0.0,xlim=c(0,6),ymax=0.7)
> truehist(duration,h=0.5,x0=0.1,xlim=c(0,6),ymax=0.7)
> truehist(duration,h=0.5,x0=0.2,xlim=c(0,6),ymax=0.7)
> truehist(duration,h=0.5,x0=0.3,xlim=c(0,6),ymax=0.7)
> truehist(duration,h=0.5,x0=0.4,xlim=c(0,6),ymax=0.7)
```

Illustration II

```
> # Improve by averaging these histograms
> breaks <- seq(0, 5.9, 0.1)
> counts <- numeric(length(breaks))
> for(i in (0:4)) {
+ counts[i+(1:55)] <- counts[i+(1:55)]+
+   rep(hist(duration,breaks=0.1*i+seq(0, 5.5, 0.5),
+   plot=FALSE)$intensities, rep(5,11))
+ }
> plot(breaks+0.05, counts/5, type="l", xlab="duration",
+ ylab="averaged", bty="n", xlim=c(0, 6), ylim=c(0,0.7))
> detach(geyser)
```

Illustration III

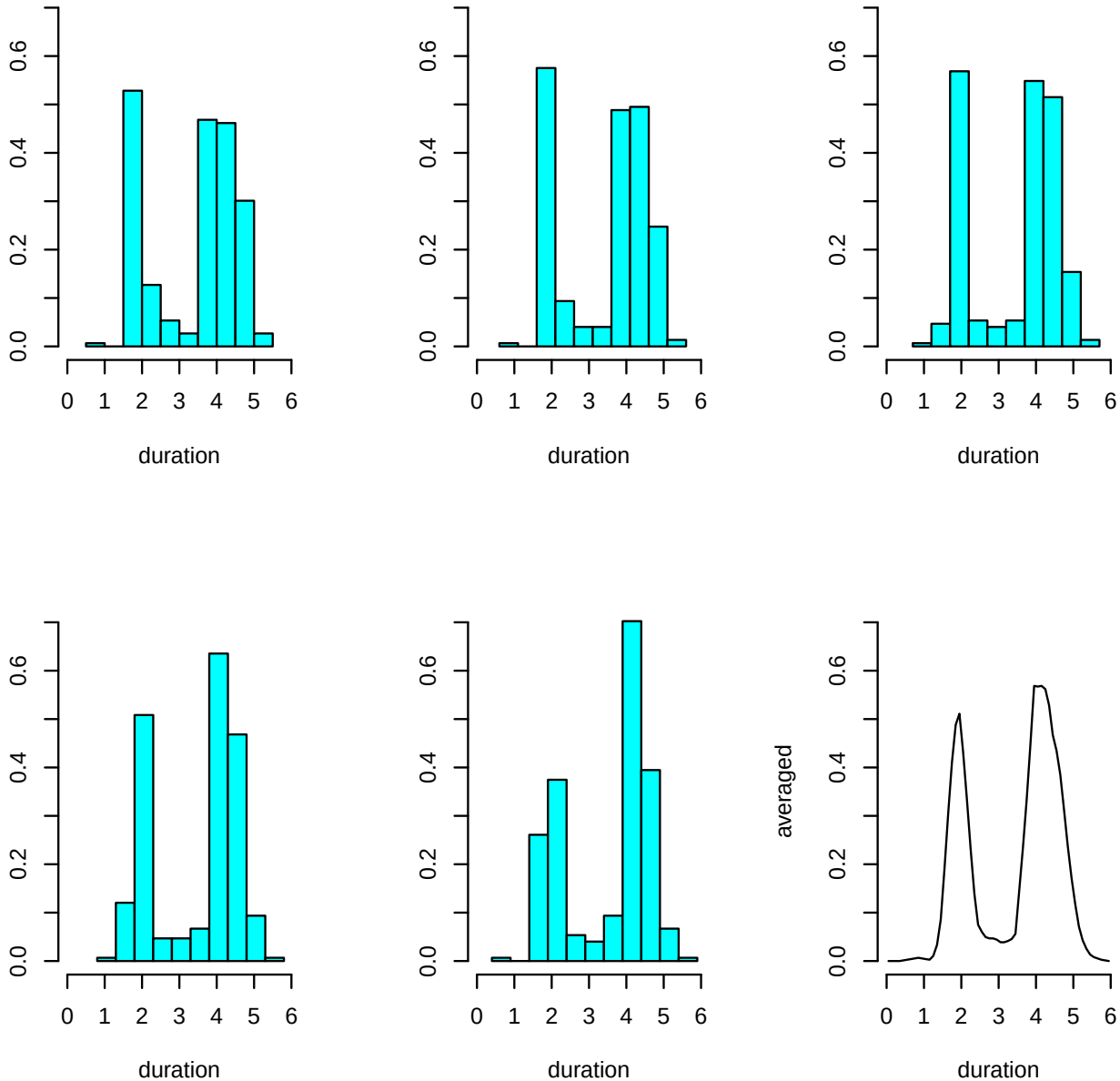


Illustration (suite) I

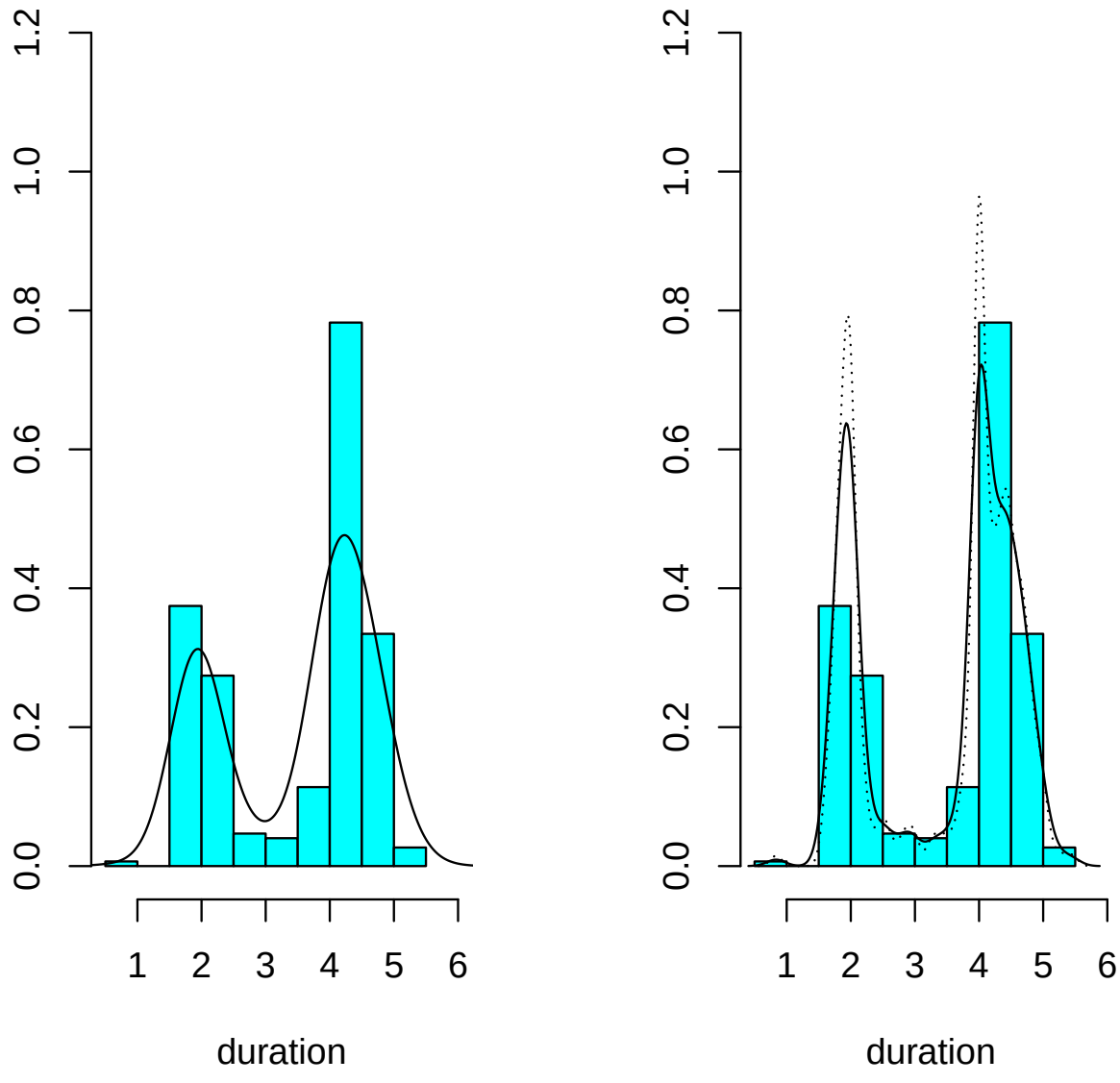
On utilise à présent un estimateur à noyau gaussien avec différentes fenêtres.

```
> # nrd, SJ and SJ-dpi are different methods for choosing
> # the fixed bandwidth
> # do ?bw.nrd for more details. To get values used:
> # bw.nrd(geyser$duration) = 0.389
> # bw.SJ(geyser$duration) = 0.090
> # bw.SJ(geyser$duration, method='dpi') = 0.144
> par(mfrow=c(1,2))
> attach(geyser)
> truehist(duration,nbins = 15,xlim=c(0.5,6),ymax=1.2)
> lines(density(duration,width="nrd"))
> truehist(duration,nbins=15,xlim=c(0.5,6),ymax=1.2)
> lines(density(duration,width="SJ",n=256),lty = 3)
> # dotted line
```

Illustration (suite) II

```
> lines(density(duration,n=256,width ="SJ-dpi"),lty=1)
> detach(geyser)
```

Illustration (suite) III



Plan partie 3

Estimateur à noyau des k plus proches voisins

Principe

- On se donne une distance d sur l'ensemble des observations ($\mathbb{R}$ ou $\mathbb{R}^m$) et un noyau $K : \mathbb{R} \to \mathbb{R}^+$ (ex : noyau gaussien),
- Estimateur à noyau de fenêtre h variable : en tout point x, on considère les k observations les plus proches de x

$$\hat{f}_n^{knn}(x) = \frac{1}{n[V_k(x)]^m} \sum_{i=1}^{n} K\left(\frac{d(X_i, x)}{V_k(x)}\right),$$

où $V_k(x)$ est le rayon de la plus petite boule de centre x qui contient k observations (m dimension de l'espace).

- La "fenêtre" s'adapte. Pbm : choisir k.

Propriétés des k plus proches voisins [Mack & Rosenblatt 79]

- Sous l'hyp. f bornée et 2 fois dérivable,
- En choisissant K tel que $\int |x| K(x)dx = 0$ et $\int x^2 K(x)dx < +\infty$,

On obtient (unidimensionnel)

$$\mathbb{V}\mathrm{ar}(\hat{f}_n^{knn}(x)) = O(k^{-1}), \quad \mathrm{Biais}(\hat{f}_n^{knn}(x)) = O((k/n)^2).$$

Contrôles similaires à ceux des estimateurs à noyau ($k = nh$).

Remarques

- Les méthodes de k plus proches voisins sont peu utilisées en estimation de densité,
- Par contre, elles sont très populaires en classification supervisée.

Plan partie 3

Plan partie 3

Construction I

Dans cette section, on suppose que $f \in \mathbb{L}_2(I)$ où $I = \mathbb{R}$ ou $[a, b]$.

Base orthonormée (b.o.n.)

- Soit $(\phi_j)_{j \geq 1}$ une b.o.n. de $\mathbb{L}_2(I)$,
- On a $f = \sum_{j \geq 1} \theta_j \phi_j$, au sens d'une série convergente dans $\mathbb{L}_2(I)$, et où $\theta_j = \int_I f(x)\phi_j(x)dx$ est la projection de f sur la $j$ème coordonnée de la base.

Estimateur des coordonnées θ_j

$$\hat{\theta}_j = \frac{1}{n} \sum_{i=1}^{n} \phi_j(X_i)$$

estimateur sans biais de θ_j.

Construction II

Définition

- Soit $\{\hat{\theta}_j\}_{j\geq 1}$ une suite d'estimateurs des coordonnées $\{\theta_j\}_{j\geq 1}$, on définit l'estimateur par projection de f via

$$\hat{f}_{n,N} = \sum_{j=1}^{N} \hat{\theta}_j \phi_j ,$$

  i.e. on prend la projection de f sur les N premières coordonnées de la base et on estime ses coordonnées.
- N joue le rôle d'un paramètre de lissage, comme h auparavant.
- Compromis sur la taille de la base ($=$ la valeur de N). Plus N est grand, plus le biais est petit, plus la variance est grande.

Exemples de bases I

Bases (régulières dyadiques) d'histogrammes (sur $[0,1]$)

On fixe $p \geq 1$ et

$$\phi_k(x) = c_k 1\Big\{\Big[\frac{k-1}{2^p}, \frac{k}{2^p}\Big[\Big\}, \quad 1 \leq k \leq 2^p,$$

où $c_k = 2^{p/2}$ constante de normalisation.

Base trigonométrique (de Fourier) de $\mathbb{L}_2([0,1])$

Définie par : $\phi_1 \equiv 1$, et

$$\begin{aligned} \forall k \geq 1, \quad & \phi_{2k} : x \to \sqrt{2}\cos(2\pi k x) \\ \text{et} \quad & \phi_{2k+1} : x \to \sqrt{2}\sin(2\pi k x). \end{aligned}$$

Exemples de bases II

Bases d'ondelettes de $\mathbb{L}_2(\mathbb{R})$

Soit $\psi : \mathbb{R} \to \mathbb{R}$ une fonction suffisamment régulière. On définit les fonctions translatées en échelle et en temps

$$\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k) \quad \forall k, j \in \mathbb{Z}.$$

Alors, sous certaines hypothèses sur ψ, les fonctions $\{\psi_{j,k}\}_{j,k\in\mathbb{Z}}$ forment une b.o.n. de $\mathbb{L}_2(\mathbb{R})$ et pour tout $h \in \mathbb{L}_2(\mathbb{R})$, on a

$$h = \sum_{j\in\mathbb{Z}} \sum_{k\in\mathbb{Z}} \theta_{j,k}\psi_{jk},$$

où $\theta_{jk} = \int h\psi_{jk}$ et la série ci-dessus converge dans $\mathbb{L}_2(\mathbb{R})$.

Exemples de bases III

Remarques

- Une base d'ondelettes est constituée de deux indices : j pour l'échelle (=fréquence) et k pour la translation (=temps),
- La base trigonométrique localise les fonctions en fréquence tandis que les bases d'ondelettes localisent les fonctions en fréquence et en temps.

Plan partie 3

Propriétés

Risque quadratique

- Les estimateurs par projection ont le même type de performances que les estimateurs à noyaux (sur des classes de fonctions régulières, pour un bon choix de N)
- On peut mettre en œuvre des techniques de sélection de modèle pour choisir N.

Version multivariée

On abordera dans la prochaine partie les projection pursuit density estimator dans le cadre de l'estimation de densité multivariée.

Références I


[Celisse & Robin 2008] Celisse, A. and S. Robin (2008).
Nonparametric density estimation by exact leave-p-out cross-validation.
Comput. Statist. Data Anal., 52(5) :2350–2368.


[Celisse & Robin 2010] Celisse, A. and S. Robin (2010).
A cross-validation based estimation of the proportion of true null hypotheses.
Journal of Statistical Planning and Inference, 140 :3132–3147, 2010.


[Cohen *et al.* 01] A. Cohen and R. DeVore and G. Kerkyacharian and D. Picard
Maximal spaces with given rate of convergence for thresholding algorithms.
Bernoulli, 8(2) :219–253, 2002.

Références II


[Freedman & Diaconis 81] D. Freedman and P. Diaconis
On the Histogram as a Density Estimator : $\mathbb{L}_2$ Theory.
Z. Wahrscheinlichkeitstheorie verw. Gebiete, 57(4) :453–476, 1981.


[Kerkyacharian & Picard 02] G. Kerkyacharian and D. Picard
Minimax or maxisets ?
ACHA, 11 :167–191, 2001.


[Lepski 92] O.V. Lepski
Asymptotically minimax adaptive estimation. II. Schemes without optimal adaptation. Adaptive estimates.
Theory Probab. Appl., 37(3) :433–448, 1992.


[Mack & Rosenblatt 79] Y.P. Mack and M. Rosenblatt
Multivariate k-nearest neighbour density estimates.
J. Multivariate Anal., 9 :1–15, 1979.

Références III

[Nguyen & Matias 12] V. H. Nguyen and C. Matias
Is it possible to construct an asymptotically efficient estimator of the proportion of true null hypotheses in a multiple testing setup ?
Preprint hal-00647082, 2012.