Rapport du projet de M1

Combien de temps pour faire une espèce ?

Wiam Chaoui Sophie Manuel Stéphane Sadio

2021

Contents

1 Introduction **7**

1.1 Motivation 7

1.2 Problématique 8

2 Méthodes non-paramétriques **9**

2.1 Introduction 9

2.2 Estimation par histogramme 10

2.3 Estimateurs par projection : 10

2.4 Estimateurs à noyau de densité 11

3 Estimateur de densité à noyau **13**

3.1 Evaluer un estimateur 13

3.2 Méthodes adaptatives 19

4 Applications **31**

4.1 Fonction dens 31

4.2 Applications aux données du vivant 31

5 Conclusion **39**

6 References **41**

Table des matières

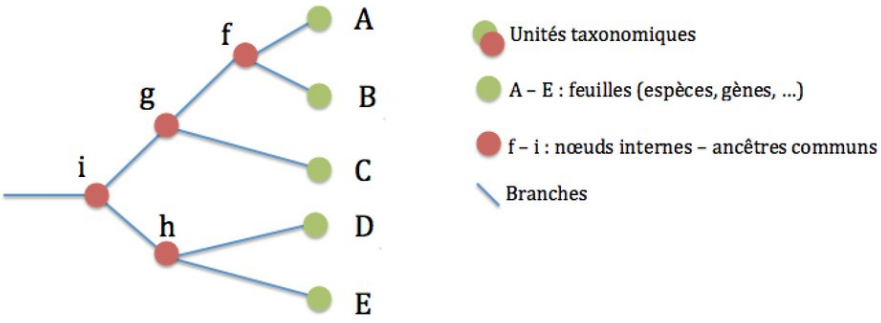
Chapter 1

Introduction

1.1 Motivation

La classification du vivant est depuis longtemps un vrai casse-tête pour les biologistes, surtout en ce qui concerne la notion d'*espèce*. De fait, il existe plusieurs définitions du mot espèce, ce qui rend encore plus compliqué un consensus. C'est pour cela que dans la suite nous ne nous étendrons pas sur cette notion et nous ne nous concentrerons que sur des espèces prédéfinies.

Les *arbres phylogénétiques* sont des outils permettant de représenter graphiquement certaines données de classification. En effet, ils présentent les relations de parenté entre *espèces*. On retrouve dessous différentes espèces actuelles, mais aussi leurs ancêtres communs (les *branchements évolutifs* qui correspondent à l'apparition d'une nouvelle homologie), ou encore la durée avant l'apparition d'une nouvelle espèce qui est donnée par la longueur des branches.



Dans la suite, nous nous intéresserons aux *branchements évolutifs*. Supposons qu'un branchement évolutif apparaît après une durée aléatoire d'une loi fixée μ indépendamment du passé et du futur évolutif des espèces.

Quelle est cette loi μ? Sa variance ? Sa moyenne ?

On observe des branchements successifs qui composent l'arbre phylogénétique et à partir de ces données quantitatives observées, on veut estimer la fonction de densité f qui donne la probabilité qu'un nouveau branchement évolutif apparaisse après un certain temps.

Formellement on a le modèle densité suivant, soient les vecteurs aléatoires $X_1, \dots, X_n$ tel que $n \in \mathbb{N}^*$ à valeur dans $\mathbb{R}$, indépendants et identiquement distribués de longueurs de branche observées qui ont pour une fonction de densité f par rapport à la mesure de Lebesgue sur $\mathbb{R}$ supposée inconnue. Notre objectif est d'estimer cette fonction densité f sur laquelle on fait le moins d'hypothèses possibles. On fera seulement les hypothèses d'existence, de continuité et de positivité de la fonction,en servant une observation $(X_1, ...X_n)$, ce qui nous mène en statistique non-paramétrique, où le paramètre cherché est une densité de probabilité qui appartient à un espace fonctionnel infini, d'où la problématique de notre sujet.

1.2 Problématique

Comment estimer la loi de densité de la création d'une nouvelle espèce avec une méthode d'estimation non-paramétrique ?

Pour commencer, en se basant sur quelques définitions nous présenterons les méthodes d'estimations non-paramétriques et en introduisant quelques types d'estimateurs.
Par la suite, nous approfondirons les estimateurs de densité à noyau en menant une discussion sur leurs critères d'évaluation. Ainsi nous consacrerons un chapitre pour présenter des méthodes adaptatives.
Enfin, pour répondre à la problématique, nous implémenterons un estimateur à noyau adaptatif, de type Goldenshluger-Lepski puis l'utiliser sur des données d'arbres phylogénétiques.

Chapter 2

Méthodes non-paramétriques

En statistique, on parle d'estimation quand on cherche à trouver certains paramètres inconnus caractérisant une distribution à partir d'un échantillon de données observées, en se basant sur différentes méthodes. On se tourne vers l'estimation non-paramétrique lorsque l'on traite des paramètres à dimension infinie. Dans notre cas, ce paramètre est la fonction densité appartenant à un espace fonctionnel.

Nous présenterons dans la suite une courte introduction à l'estimation non-paramétrique et nous introduirons les deux classes principales de l'estimation fonctionnelle (l'estimation par projection et l'estimation à noyau).

2.1 Introduction

L'estimation non-paramétrique vise à résoudre des problèmes d'estimation dans le cadre statistique où le modèle auquel on s'intéresse n'est pas décrit par un nombre fini de paramètres et dont chacun de ces paramètres ne permet pas de décrire la structure générale de la distribution des variables aléatoires.
Cela signifie qu'on utilise des modèles statistiques à dimension infinie.

Dans le cadre de notre problématique on s'intéresse à l'estimation de densité.
Un des principes de base de l'estimation de la densité selon une méthode d'estimation non-paramétrique est le suivant

Soit un échantillon d'observations $X = \{X_1, \dots, X_n\}$ de variables aléatoires i.i.d admettant une densité $f = F'$. Supposons que $f \in \mathcal{F}$ où $\mathcal{F}$ est un espace fonctionnel. On cherche à estimer la fonction de densité inconnue f à partir de

ces observations.
On notera $\widehat{f}$ l'estimateur de f.
On se trouve donc avec le modèle suivant $\{\mathbb{P} = \mathbb{P}_f,\ f \in \mathcal{F}\}$ où $\mathbb{P}_f$ est la mesure probabilité de la densité f.

L'estimation ici concerne donc la fonction elle même plutôt que les paramètres, ce qui explique le nom d'estimation non-paramétrique.

Remarque 1 - *On considère souvent les distances L^p avec $p = 1, 2$ ou ∞ .*

2.2 Estimation par histogramme

Une des premières approches possibles d'estimations non-paramétriques de la fonction de densité est l'estimation par l'histogramme. C'est une méthode qui consiste à obtenir un graphique de type histogramme pour la répartition des observations et à considèrer cet histogramme comme une approximation de la fonction densité f. (ajout formule graphique)

Nous traiterons dans la suite deux grandes familles de méthodes linéaires pour estimer une fonction densité :
- L'estimation par projection,
- L'estimation par noyau.

2.3 Estimateurs par projection :

Définition 1 *Soient $X = (X_1, ..., X_n)$ considérer dans le modèle de densité présenté précédemment et $T_j : \mathbb{R}^d x \mathbb{R}^d \to \mathbb{R}$ des fonctions mesurables. Un estimateur $x \to \widehat{f}(x) = \widehat{f}(x, X)$ est dit linéaire si il peut s'écrire sous la forme*

$$\widehat{f}(x) = \sum_{j=1}^{N} T_j(x, X_j), \forall x \in \mathbb{R}^d.$$

Définition 2 *(Estimateur par projection)*
Soit $f \in \mathcal{F} = (L^2, \| \ . \ \|, < \ ., \ . \ >_)$ où $\mathcal{F}$ un espace de Hilbert muni d'une base orthonormée $(\Phi_j)_{j \in \mathbb{N}^*}$ de L^2.Si $\mathbb{E}_N$ un sous-espace fini de $\mathcal{F}$ et $a_j =< f, \Phi_j >= \int_{\mathbb{R}} f(x)\Phi_j(x)dx$ avec $1 \leq |N| < \infty$.
On appelle un estimateur par projection de la fonction f sur $\mathbb{E}_N$, la fonction $\Pi_N\ f$ définit comme suit

$$\Pi_{vect(1,..N)}\ f = \sum_{j=1}^{N} a_j \Phi_j$$

Remarque 2 -*Cette méthode nous ramène au cas paramétrique.*

Dans la suite on procédera à la méthode la plus fréquemment utilisée pour l'estimation d'une densité : L'estimation à noyau.

(Argument pour le choix de la méthode à voir)

2.4 Estimateurs à noyau de densité

Notre but est d'estimer la densité f. Pour cela, on s'appuiera sur un échantillon *i.i.d.* $X = (X_1, ..., X_n)$ où chacune des variables X_i admet la densité f (par rapport à la mesure de Lebesgue).

Pour estimer une densité on peut utiliser une méthode à noyau. Les méthodes à noyau sont des méthodes non-paramétriques qui permettent de proposer une estimation de la densité plus lisse que celle obtenue par un histogramme.

2.4.1 Comment construit-on un estimateur à noyau ?

L'idée pour la construction de cet estimateur est d'utiliser l'approximation suivante , valable lorsque h est petit :

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

Pour estimer la densité f on peut passer par un estimateur $\hat{F}_n$ de la fonction de répartition F. $\hat{F}_n$ est la fonction de répartition empirique ($\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{2h} \mathbb{1}_{X_i \in]x-h, x+h[}$).

$$\hat{f}_n(x) = \frac{\hat{F}_n(x+h) - \hat{F}_n(x-h)}{2h} = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{2h} \mathbb{1}_{X_i \in]x-h; x+h]}$$

Notons $\hat{f}(x)$ l'estimateur à noyau de la densité f, alors celui-ci s'écrit :

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

où h est la fenêtre (ou paramètre de lissage), n le nombre d'observations, et K le noyau. Cette formule n'est valable que si h est petit et strictement positif.

ici $K(u) = \frac{1}{2}\mathbb{1}_{u\in]-1;1]}$, il s'agit du noyau de Rosenblatt, mais il existe d'autres noyaux.

2.4.2 Qu'est un noyau ?

Définition 3 *(Noyau)*

Un noyau (kernel en anglais) est une application $K : \mathbb{R} \to \mathbb{R}$ intégrable et centrée telle que :

$$\int_{\mathbb{R}} K(u)du = 1 \qquad et \qquad \int_{\mathbb{R}} uK(u) = 0$$

si le noyau est en plus positif alors il correspond à une fonction de densité.

Exemples de noyau :

- Noyau de Rosenblatt, ou rectangulaire : $K(u) = \frac{1}{2}\mathbb{1}_{u\in]-1;1]}$
- Noyau Gaussien : $K(u) = \frac{1}{\sqrt{2\pi}} exp(-\frac{u^2}{2})$
- Noyau d'Epanechnikov : $K(u) = \frac{3}{4}(1-u^2)\mathbb{1}_{[-1,1]}(u)$
- Noyau triangulaire : $K(u) = (1- \mid u \mid)\mathbb{1}_{[-1,1]}(u)$
- Noyau Biweight : $K(u) = \frac{15}{16}(1-u^2)^2\mathbb{1}_{[-1,1]}(u)$

Les propriétés du noyau (continuité, différentiabilité…) se transmettent à l'estimateur $\hat{f}_n$.

Chapter 3

Estimateur de densité à noyau

3.1 Evaluer un estimateur

Avant de commencer cette partie on va d'abord introduire quelques notions et définitions

Définition 4 *Noyau*
On note par le noyau la fonction intégrable $K : \mathbb{R} \to \mathbb{R}$ *tel que:*

$$\int_{\mathbb{R}} K(u)du = 1.$$

Lemme 1 *Soient :*
$h > 0$ *le paramètre de lissage et* $K_h : u \in \mathbb{R} \to K(\frac{u}{h})/h$*.On peut approximer la famille* $(K_h)_{h>0}$ *par l'identité du produit de convolution.*

Démonstration 1 *A Faire*

Corollaire 1 $K_h * f : x \to \int_{\mathbb{R}} K_h(y-x)f(x)dx$ *tend vers la fonction f quand h tend vers 0 pour la distance* L^2*.*

Pour évaluer un estimateur $\widehat{f}$ on définit son risque associé.

Définition 5 *La fonction de risque :*

$$\mathcal{R}(\widehat{f}, f) = \mathbb{E}_f[\| \widehat{f} - f \|^2]$$

Remarque 3 *La fonction de risque associé nous permet de comparer l'estimateur $\hat{f}$ et f.*
On cherche à minimiser ce risque associé (i.e tend vers 0 pour un nombre d'observation assez grand).

3.1.1 Risque quadratique des estimateurs à noyau sur les classe des espaces de Hölder

Nous nous intéressons au risque quadratique de $\hat{f}_n$, définit par :
étant donné $x_0 \in \mathbb{R}$

$$R(\hat{f}_n, f) = \mathbb{E}[|\hat{f}_n(x_0) - f(x_0)|^2]$$

Rappelons la décomposition "biais-variance" du risque quadratique :

$$\mathbb{E}[|\hat{f}_n(x_0) - f(x_0)|^2] = (\mathbb{E}[\hat{f}_n(x_0)] - f(x_0))^2 + \mathbb{V}[\hat{f}_n(x_0)]$$

3.1.1.1 Majoration du biais et de la variance

Dans cette section, nous allons nous intéresser au compromis biais-variance afin de minimiser le risque quadratique. Nous introduirons après quelques définitions deux propositions qui montrent que sous certaines hypothèses, on peut majorer le biais ainsi que la variance.

Définition 6 *Pour tout $\beta > 0$ et $L > 0$, on définit la classe de Hölder de régularité β et de rayon L par*

$$\Sigma(\beta, L) = \{f : \mathbb{R} \longrightarrow \mathbb{R} \text{ t.q. } f \text{ est } \lfloor\beta\rfloor \text{ fois dérivable et} \forall\ (x, y) \in \mathbb{R}_2 \quad | f^{\lfloor\beta\rfloor}(y) - f^{\lfloor\beta\rfloor}(x) | \leq L | x - y |^{\beta - \lfloor\beta\rfloor}$$

On notera $\Sigma_d(\beta, L)$ l'intersection de $\Sigma(\beta, L)$ et l'ensemble des densités.

Remarque 4 - *Si $\beta = 1$ on obtient l'ensemble des fonctions L-lipschitziennes.*
- *Si $\beta > 1$ alors $f' \in \Sigma(\beta - 1, L)$.*

Proposition 1 *(admise) Soit $\beta > 0$ et $L > 0$, il existe une constante $M(\beta, L)$ telle que*

$$\sup_{f \in \Sigma_d(\beta, L)} \| f \|_\infty = \sup_{x \in \mathbb{R}} \sup_{f \in \Sigma_d(\beta, L)} f(x) \leq M(\beta, L)$$

Définition 7 *Soit $\ell \in \mathbb{N}^*$. On dit que le noyau K est d'ordre ℓ si $u^j K(u)$ est intégrable et $\int u^j K(u)du = 0, \quad j = 1, ..., \ell$*

Proposition 2 *: Si $f \in \Sigma(\beta, L)$ avec $\beta > 0$ et $L > 0$ et si K est un noyau d'ordre $\ell = \lfloor \beta \rfloor$ tel que $\int |u^\beta| . |K(u)| \, du < \infty$ alors pour tout $x_0 \in \mathbb{R}$, et pour tout $h > 0$ le biais peut être borné comme suit :*

$$|\mathbb{E}[\widehat{f}_n(x_0)] - f(x_0)| \leqslant \frac{h^\beta L}{\ell!} \int |u|^\beta |K(u)| du$$

Démonstration 2 *On a*

$$\begin{aligned}
\mathbb{E}(\widehat{f}_n(x_0)) &= \mathbb{E}(\frac{1}{n}\sum_{i=1}^{n} \frac{1}{h} K(\frac{X_i - x_0}{h})) \\
&= \mathbb{E}(\frac{1}{h} K(\frac{X_1 - x_0}{h})) \\
&= \frac{1}{h} \int K(\frac{u - x_0}{h}) f(u) du \\
&= \int K(v) f(x_0 + hv) dv, \ (\textit{en posant } v = \frac{u - x_0}{h}).
\end{aligned}$$

De plus

$$f(x_0) = f(x_0) \times 1 = f(x_0) \int K(v) dv.$$

Comme $f \in \Sigma(\beta, L)$, f admet $\lfloor \beta \rfloor$ dérivées et par un développement de Taylor-Lagrange on a, pour tout $x \in \mathbb{R}$,

$$f(x) = \sum_{i=1}^{\ell-1} \frac{(x - x_0)^k}{k!} f^{(k)}(x_0) + \frac{(x - x_0)^\ell}{\ell!} f^{(\ell)}(x_0 + \zeta(x - x_0))$$

avec $\zeta \in]0, 1[$. Autrement dit on a, avec $x = x_0 + hv$,

$$f(x_0 + hv) - f(x_0) = \sum_{i=1}^{\ell-1} \frac{(hv)^k}{k!} f^{(k)}(x_0) + f^{(\ell)}(x_0 + hv\zeta) \frac{(hv)}{\ell!}$$

pour un certain $\zeta \in]0, 1[$. Donc

$$\begin{aligned}
\int K(v)(f(x_0 + hv) - f(x_0)) dv &= \int K(v)(\sum_{i=1}^{\ell-1} f^{(k)}(x_0) + f^{(\ell)}(x_0 + hv\zeta) \frac{(hv)^\ell}{\ell!}) dv \\
&= \frac{h^\ell}{\ell!} \int K(v) v^\ell f^{(\ell)}(x_0 + hv\zeta) dv
\end{aligned}$$

Comme K est d'ordre ℓ, on a aussi $\int K(v)v^\ell f'^{(\ell)}(x_0)dv = 0$. Donc on a

$$\int K(v)(f(x_0+hv)-f(x_0))dv = \frac{h^\ell}{\ell!}\int K(v)v^{(\ell)}(f^{(\ell)}(x_0+hv\zeta)-f^{(\ell)}(x_0))dv$$

Or, $f \in \Sigma(\beta, L)$, on a donc $\mid f^{(\ell)}(x_0+hv\zeta)-f^{(\ell)}(x_0) \mid \leq L|hv|^{\beta-\ell}$. Et finalement

$$\mid \int K(v)(f(x_0+hv\zeta)-f(x_0))dv \mid \leq \frac{\mid h \mid^\beta}{\ell!}\int \mid K(v) \mid |v|^\ell l|hv|^{\beta-\ell}dv$$

ce qui signifie que

$$|\mathbb{E}(\hat{f}_n(x_0)-f(x_0)| \leq \frac{L|h|^\beta}{\ell!}\int |K(v)||v|^\beta dv$$

Corollaire 2 *Le biais au carré tend vers zéro à la vitesse $h^{2\beta}$. En particulier, le biais tend vers zéro quand h tend vers zéro. Plus la fonction f est régulière, plus le biais tend vite vers zéro quand h tend vers zéro (à condition bien sûr que l'ordre du noyau soit suffisamment grand). Nous en déduisons la convergence de l'espérance de l'estimateur à noyau $\hat{f}_n$ vers la fonction f. Et donc, l'estimateur à noyau est asymptotiquement sans biais, $\hat{f}_n$ est donc consistant.*

Proposition 3 *: Si f est bornée et si K est de carré intégrable alors*

$$\mathbb{V}(\hat{f}_n(x_0)) \leqslant \frac{\|f\|_\infty \|K\|_2^2}{nh}$$

En particulier, si $f \in \Sigma(\beta, L)$ alors

$$\mathbb{V}(\hat{f}_n(x_0)) \leqslant \frac{M(\beta,L)\|K\|_2^2}{nh}$$

Démonstration 3 *:*

$$\begin{aligned}
\mathbb{V}(\hat{f}_n(x_0)) &= \mathbb{V}(\frac{1}{nh}\sum_{i=1}^n K(\frac{X_i-x_0}{h})) = \sum_{i=1}^n \mathbb{V}(\frac{1}{nh}K(\frac{X_i-x_0}{h})) \\
&= \sum_{i=1}^n \frac{1}{n^2h^2}\mathbb{V}(K(\frac{X_i-x_0}{h})) = \frac{1}{nh^2}\mathbb{V}(K(\frac{X_1-x_0}{h}) \\
&\leqslant \frac{1}{nh^2}\mathbb{E}(K^2(\frac{X_1-x_0}{h})) = \frac{1}{nh^2}\int K^2(\frac{u-x_0}{h})f(u)du \\
&\leqslant \frac{1}{nh}\int K^2(v)f(x_0+vh)dv
\end{aligned}$$

Et enfin,on utilise la proposition ? : il existe une constante positive $M(\beta, L)$ tel que $\|f\|_\infty \leqslant M(\beta, L)$. Ceci implique que :

$$\mathbb{V}(\hat{f}_n(x_0)) \leqslant \frac{1}{nh} M(\beta, L) \int K^2(v)dv$$

Corollaire 3 *Pour que la variance tende vers zéro, il faut que nh tende vers l'infini. En particulier, à n fixé, la variance est une fonction décroissante de h. Il y a donc une valeur optimale de h qui doit réaliser l'équilibre entre le biais au carré et la variance. On peut à présent donner un contrôle du risque quadratique par le théorème suivant.*

Théorème 1 *Soit $\beta > 0$ et $L > 0$ et K un noyau de carré intégrable et d'ordre $\lfloor \beta \rfloor$ tel que $\int |u^\beta| . |K(u)|du < \infty$. Alors, en choisissant une fenêtre de la forme $h = cn^{-\frac{1}{2\beta+1}}$ avec une constante $c > 0$, on obtient pour tout $x_0 \in \mathbb{R}$,*

$$R(\hat{f}_n(x_0)), \Sigma_d(\beta, L)) := \sup_{f \in \Sigma_d(\beta, L)} \mathbb{E}[|\hat{f}_n(x_0) - f(x_0)|^2] \leqslant Cn^{-\frac{2\beta}{2\beta+1}}$$

où C est une constante dépendant de L, β, c et K.

Démonstration 4 *On a :*

$$R(\hat{f}_n(x_0), f(x_0)) = Biais^2 + Variance$$

Si nous nous référons aux deux propositions précédentes, nous pouvons écrire :

$$R(\hat{f}_n(x_0), f(x_0)) \leqslant (\frac{h^\beta L}{l!} \int |u|^\beta |K(u)|du)^2 + \frac{M(\beta, L) \|K\|_2^2}{nh}$$

On cherche ensuite la fenêtre h qui minimise cette quantité.Comme on cherche la vitesse de convergence en h, on utilisera la notation $c_1 = (\frac{L}{l!} \int |u|^\beta |K(u)|du)^2$ et $c_2 = M(\beta, L) \|K\|_2^2$ qui ne dépendent pas de h. On doit alors minimiser en h la quantité :

$$c_1 h^{2\beta} + \frac{c_2}{nh}$$

On a une somme d'une quantité croissante et une quantité décroissante en h. On cherche la fenêtre h qui nous donne l'ordre minimal du risque. Quand h est trop grand, le biais est trop grand, et quand h est trop petit, c'est la variance qui est trop grande (voir exemple ci-dessous). On cherche donc la fenêtre h qui réalise un équilibre entre le biais au carré et la variance:

$$h^{2\beta} \approx \frac{1}{nh}$$

où le signe $\approx$ signifie ici "de l'ordre de". Cela donne :

$$h \approx n^{-\frac{1}{2\beta+1}}$$

Autrement dit, pour une fenêtre h de l'ordre de $n^{-\frac{1}{2\beta+1}}$, le biais au carré et la variance sont de même ordre.Plus exactement, on choisit la fenêtre $h_* = cn^{-\frac{1}{2\beta+1}}$, avec c une constante strictement positive, on a :

$$\textit{Biais au carré} \approx h_*^{2\beta} \approx \textit{Variance} \approx \frac{1}{nh_*}$$

De plus, on a alors :

$$h_* \approx n^{-\frac{2\beta}{2\beta+1}}$$

Autrement dit, il existe une certaine constante C telle que, pour cette fenêtre h_*, on a :

$$R(\hat{f}_n(x_0), \sum_d(\beta, L)) \leqslant Cn^{\frac{-2\beta}{2\beta+1}}$$

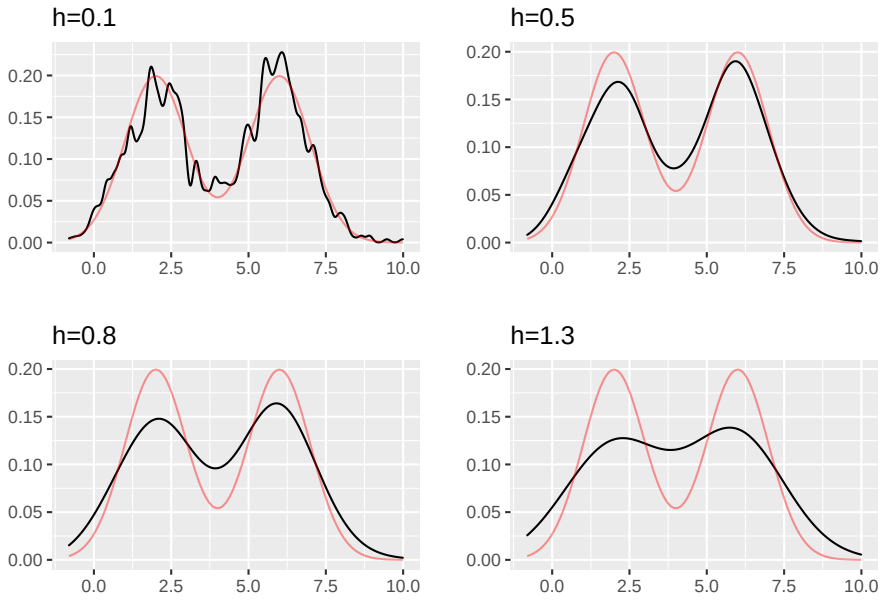
Cette fenêtre est donc optimale à une constante près (si on change c, on change C ça ne change pas le taux qui est $n^{\frac{-2\beta}{2\beta+1}}$).

Commentaire 1 *: L'estimateur dépend de β à travers la fenêtre h. Or, sans connaissance a priori sur les propriétés de la fonction f, on ne peut donc pas utiliser cet estimateur. On essaie alors de trouver un choix de fenêtre ne dépendant que des données et qui soit aussi performant (ou presque) que l'estimateur utilisant cette fenêtre optimale. A ce sujet, on introduira plus loin un choix de fenêtre ne dépendant que des données et qui est basé sur ce qu'on appelle la validation croisée (ou "cross validation" en Anglais).*

\begin{exem} (Simulation numérique) Nous estimons la fonction densité d'une somme de deux variables gaussiennes ci-contre avec la méthode à noyau avec différentes fenêtres.

$$f(x) = \frac{1}{2}\frac{1}{\sqrt{2\pi}}(exp(-\frac{(x-2)^2}{2}) + exp(-\frac{(x-6)^2}{2}))$$

On va en fait utiliser `ggplot` pour représenter l'estimateur à noyau. La fonction qui permet de dessiner l'estimateur à noyau est `geom_density`. Le paramètre représentant le fenêtre h rappelle `bw` (comme bandwidth en Anglais). \end{exem}



3.2 Méthodes adaptatives

On introduit précédemment la notion de l'estimation de la densité qui dépend d'un paramètre de lissage h. Soit $(\hat{f}_h)_{h\in\mathcal{H}}$ une famille des estimateurs de la fonction densité f que que l'on cherchons cherche à estimer.
Une question s'impose : comment peut-on construire un estimateur à risque optimal à partir de cette famille en tenant en considération les observations ?
Dans la théorie adaptative f est toujours supposée appartenir à une classe fonctionnelle. Cette classe n'est pas connu à priori mais supposé appartenir à une famille de classes fonctionnelles$\{\mathcal{F}_\alpha, \alpha \in \mathcal{A}\}$ où $\mathcal{A}$ est un ensemble des paramètres de nuisance. (ref : Sur l'estimation adaptative d'une densité multivariée sous l'hypothèse de la structure d'indépendance-Approche minimax adaptative).

Dans cette partie, afin de répondre à la question posée auparavant, nous allons discuter du choix du noyau en premier lieu. Ensuite, nous introduirons deux méthodes pour le choix du paramètre de lissage h.

3.2.1 Choix du noyau

Avant de présenter le critère de choix du noyau nous allons introduire quelques outils mathématiques qui simplifient l'écriture du critère.

Tout d'abord, nous avons besoin du risque quadratique $\mathcal{R}$ aussi appelé l'erreur quadratique moyenne (**M**ean **S**quared **E**rror en anglais). Dans cette partie nous allons noter MSE le risque quadratique pour des questions pratiques.

$$\begin{aligned} MSE = \mathcal{R} &= \mathbb{E}\left[\{\hat{f}_n(x) - f(x)\}\right] \\ &= \mathbb{V}\left[\hat{f}_n(x)\right] + Biais^2\left[\hat{f}_n(x)\right] \\ &= MSE(x; n, h, K, f). \end{aligned}$$

comme nous l'avons montré précédemment. @ref{#holder}

Propriété 1 *(Biais ponctuel) Pour tout x fixé dans $\mathbb{R}$, le biais de l'estimateur $\hat{f}_n$ est donnée par l'équation suivante :*

$$Biais\left[\hat{f}_n(x)\right] \approx \frac{1}{2}h^2 f''(x) \int_{\mathbb{R}} t^2 K(t)dt.$$

Démonstration 5 *Démontrons la propriété du biais. Rappelons que les variables aléatoires $\{X_1, \ \ldots, \ X_n\}$ sont i.i.d., alors*

$$\begin{aligned} Biais\left[\hat{f}_n(x)\right] &= \mathbb{E}\left[\hat{f}_n(x)\right] - f(x) \\ &= \mathbb{E}\left[\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) - f(x)\right] \\ &= \frac{1}{nh}\sum_{i=1}^{n} \mathbb{E}\left[K\left(\frac{X_i - x}{h}\right)\right] - f(x) \text{ par linéarité de l'espérance} \\ &= \frac{1}{h}\mathbb{E}\left[K\left(\frac{X_1 - x}{h}\right)\right] - f(x) \qquad \text{car les } X_i \text{ sont i.i.d.} \\ &= \frac{1}{h}\int_{\mathbb{R}} K\left(\frac{x_1 - x}{h}\right) f(x_1)dx_1 - f(x). \end{aligned}$$

Procédons à un changement de variables.

$$t = \frac{x_1 - x}{h} \qquad et \qquad dt = \frac{dx_1}{h}.$$

On obtient alors,

$$Biais\left[\hat{f}_n(x)\right] = \int_{\mathbb{R}} K(t)f(x + ht)dt - f(x).$$

Après avoir transformé l'écriture du biais, nous pouvons donner une approximation de celui-ci en utilisant la formule de Taylor-Young à l'ordre 2.

$$f(x + ht) = f(x) + htf'(x) + \frac{h^2t^2}{2}f''(x) + o(h^2t^2).$$

Ce qui nous donne en remplaçant dans la formule

$$Biais\left[\hat{f}_n(x)\right] \approx \frac{1}{2}h^2 f''(x)\int_{\mathbb{R}} t^2 K(t)dt - f(x) + o(h^2t^2).$$

et puisque le noyau est centré et son intégrale sur $\mathbb{R}$ *est égale à* 1*, on a bien que*

$$Biais\left[\hat{f}_n(x)\right] \approx \frac{1}{2}h^2 f''(x)\int_{\mathbb{R}} t^2 K(t)dt.$$

Propriété 2 *(Variance ponctuelle) Pour tout* x *fixé dans* $\mathbb{R}$*, la variance de l'estimateur* $\hat{f}_n$ *est donnée par l'équation suivante :*

$$\mathbb{V}\left[\hat{f}_n(x)\right] \approx \frac{1}{nh}f(x)\int_{\mathbb{R}} K(t)^2 dt.$$

Démonstration 6 *Démontrons la propriété de la variance. Rappelons que les variables aléatoires* $\{X_1,\ \ldots,\ X_n\}$ *sont i.i.d., alors*

$$\begin{aligned}
\mathbb{V}\left[\hat{f}_n(x)\right] &= \mathbb{V}\left[\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)\right] \\
&= \frac{1}{nh^2}\mathbb{V}\left[K\left(\frac{X_1 - x}{h}\right)\right] \qquad \textit{car les } X_i \textit{ sont i.i.d.} \\
&= \frac{1}{nh^2}\mathbb{E}\left[K\left(\frac{X_1 - x}{h}\right)^2\right] - \frac{1}{nh^2}\mathbb{E}\left[K\left(\frac{X_1 - x}{h}\right)\right]^2 \\
&= \frac{1}{nh^2}\int_{\mathbb{R}} K^2\left(\frac{x_1 - x}{h}\right) f(x_1)dx_1 - \frac{1}{nh^2}\left[\int_{\mathbb{R}} K\left(\frac{x_1 - x}{h}\right) f(x_1)dx_1\right]^2.
\end{aligned}$$

On effectue encore une fois le même changement de variables.

$$t = \frac{x_1 - x}{h} \quad \textit{et} \quad dt = \frac{dx_1}{h}.$$

Nous trouvons désormais,

$$\begin{aligned}
\mathbb{V}\left[\hat{f}_n(x)\right] &= \frac{1}{nh}\int_{\mathbb{R}} K^2(t)\, f(x+ht)dt - \frac{1}{n}\left[\int_{\mathbb{R}} K(t)\, f(x+ht)dt\right]^2 \\
&= \frac{1}{nh}\int_{\mathbb{R}} K^2(t)\, f(x+ht)dt - \frac{1}{n}\left[Biais\left(\hat{f}_n(x)\right) + f(x)\right]^2 \\
&= \frac{1}{nh}\int_{\mathbb{R}} K^2(t)\, f(x+ht)dt - \frac{1}{n}\left[O\left(h^2\right) + f(x)\right]^2.
\end{aligned}$$

Donc si la condition $\int_{mathbbR} K(u)^2 du < +\infty$ *est vérifiée et que la taille de l'échantillon est importante, l'équation suivante est avérée :*

$$\mathbb{V}\left[\hat{f}_n(x)\right] \approx \frac{1}{nh}f(x)\int_{\mathbb{R}} K(t)^2 dt.$$

L'approximation de l'erreur quadratique (**A**verage of **M**ean **S**quared **E**rror en anglais) est donnée par l'équation suivante :

$$AMSE(x) = \frac{1}{nh} f(x) \int_{\mathbb{R}} K(t)^2 dt + \left(\frac{1}{2} h^2 f''(x) \int_{\mathbb{R}} t^2 K(t) dt \right)^2 .$$

Elle a été calculée à partir de la variance approchée et du biais approché.

L'erreur quadratique moyenne intégrée (**M**ean **I**ntregrated **S**quared **E**rror en anglais) est une mesure théorique communément utilisée pour évaluer la différence entre f et $\hat{f}_n$. Pour l'évaluer on utilise l'erreur quadratique moyenne qu'on intègre sur le support $\mathbb{R}$ de l'estimateur.

$$\begin{aligned} MISE(n, h, K, f) &= \int_{\mathbb{R}} MSE(x; n, h, K, f) dx \\ &= \int_{\mathbb{R}} \mathbb{V}\left[\hat{f}_n(x)\right] dx + \int_{\mathbb{R}} Biais^2\left[\hat{f}_n(x)\right] dx \end{aligned}$$

De la même façon que nous l'avons fait avec l'erreur quadratique moyenne, nous allons calculer l'expression approchée de l'erreur quadratique moyenne (**A**verage of **M**ean **I**ntegrated **S**quared **E**rror en anglais).

$$\begin{aligned} AMISE(x) &= \frac{1}{nh} \int_{\mathbb{R}} f(x) dx \int_{\mathbb{R}} K(t)^2 dt + \frac{1}{4} h^4 \int_{\mathbb{R}} f''(x) dx \quad \left(\int_{\mathbb{R}} t^2 K(t) dt \right)^2 \\ &= \frac{1}{nh} \int_{\mathbb{R}} K(t)^2 dt + \frac{1}{4} h^4 \int_{\mathbb{R}} f''(x) dx \quad \left(\int_{\mathbb{R}} t^2 K(t) dt \right)^2 \\ &= \frac{1}{nh} \int_{\mathbb{R}} K(t)^2 dt + \frac{1}{4} h^4 \ \mathbb{V}(K)^2 \int_{\mathbb{R}} f''(x) dx \end{aligned}$$

à présent que nous avons défini les outils nécessaires au choix du noyau, nous allons pouvoir présenter un critère de choix pour les noyaux continus symétriques. Afin de mesurer l'efficacité des noyaux, nous utilisons une mesure qui calcule le rapport du critère $AMISE$ de deux noyaux.

$$eff(K_1, K_2) = \frac{AMISE(K_1)}{AMISE(K_2)}.$$

Supposons que K_1 est le noyau d'Epanechnikov, il est souvent utilisé comme référence par rapport aux autres noyaux continus. Après quelques calculs, on obtient que l'efficacité d'un noyau K par rapport à celui d'Epanechnikov est donnée par

$$eff(K) = \frac{3}{5 \times \int_{\mathbb{R}} K(t)^2 dt \sqrt{5 \times \int_{\mathbb{R}} t^2 K(t) dt}} \leq 1.$$

Voici un tableau récapitulatif de l'efficacité de plusieurs noyaux continus symétriques.

\begin{table}

\caption{(#tab:ker_tab)Efficacité des noyaux continus symétriques}

Noyau	Efficacité
Epanechnikov	1.000
Bigweight	0.994
Triangular	0.986
Gaussien	0.951
Rectangulaire	0.930

\end{table}

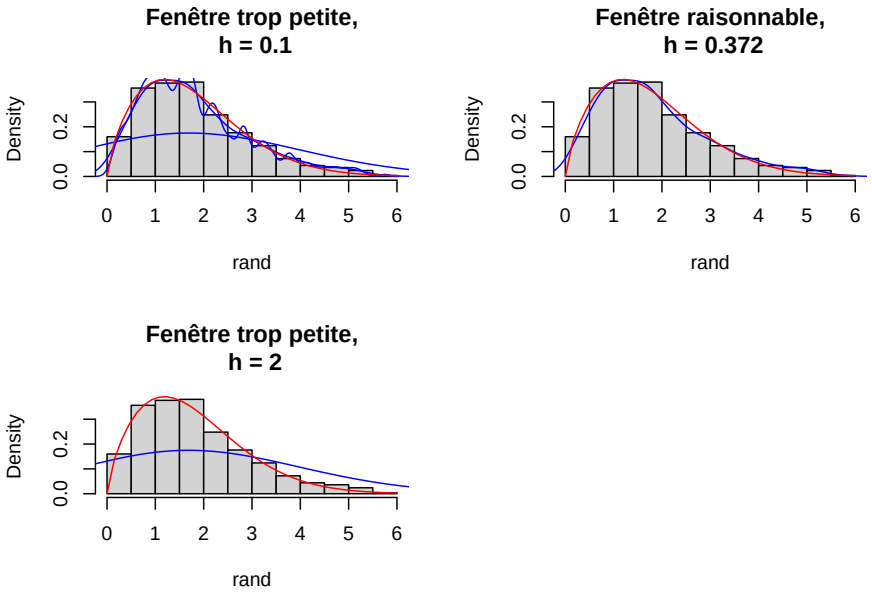
Remarque 5 *Les valeurs d'efficacité des noyaux sont très proches les unes des autres dans le cas étudié, surtout pour les trois premiers noyaux du tableau. c'est pour cela que le choix du noyau n'a au final que peu d'importance dans l'estimation de la densité.*

3.2.1.1 Comment choisir les paramètres de la méthode ?

Dans la méthode d'estimation à noyau le choix du noyau n'est pas le plus important, le vrai enjeu de cette méthode est le choix de la fenêtre h (*bandwidth*). En effet, la fenêtre détermine l'influence des données dans l'estimation. Si h est petit, l'effet local est important donc on aura beaucoup de bruit. Si h est grand, on aura une estimation plus douce, plus lisse.

Nous pouvons constater l'influence du paramètre h sur l'exemple suivant :

Nous avons simulé 500 variables suivant une loi de Weibull de paramètres ($\alpha = 1.7$, $\lambda = 2$) représentées dans l'histogramme. La courbe en rouge est la fonction de densité de la loi de Weibull et la bleue est l'estimation avec la méthode des noyaux sur les variables simulées.



La fenêtre h du second graphique est calculée automatiquement par la fonction `density`de R.

3.2.2 Choix de la fenêtre

L'estimation de densité nécessite le choix de la fenêtre qu'on note h.
En statistique non-paramétrique, ils existent plusieurs méthodes et critères de qualité pour le choix de la fenêtre.

On présente dans la suite deux méthodes:
Méthode de validation croisée.
Méthode de Goldenshluger-Lepski.

3.2.2.1 Choix de la fenêtre h par validation croisée

Le choix de la fenêtre dans la section précédente est critiquable: comme on l'a mentionné, il dépend de la régularité la fonction f qui est inconnue dans notre cas. On peut donc essayer d'estimer cette fenêtre idéale par un estimateur $\hat{h}$. De façon à souligner la dépendance à la fonction, on va noter $\hat{f}_{n,h}$ l'estimateur associé à un choix de fenêtre h. L'estimateur final sera $\hat{f}_{n,\hat{h}}$, une fois le choix de $\hat{h}$ fait.

On cherche à minimiser en h le risque quadratique pour la distance L_2 :

$$\begin{aligned} R(\hat{f}_{n,h}) &= \mathbb{E}[\left\|\hat{f}_{n,h} - f\right\|_2^2] \\ &= \mathbb{E}[\left\|\hat{f}_{n,h}\right\|_2^2] - 2\ \mathbb{E}[\int \hat{f}_{n,h}(x) f(x) dx] + \|f\|_2^2 \end{aligned}$$

Or la fonction f étant inconnue, ce risque n'est pas calculable à partir des données. On cherche donc à estimer ce risque en utilisant uniquement les données. Remarquons que minimiser en h la quantité $R(\hat{f}_{n,h}, f)$ est équivalent à minimiser en h la quantité $R(\hat{f}_{n,h}, f) - \|f\|_2^2$. On va en fait remplacer la minimisation de la quantité inconnue $R(\hat{f}_{n,h}, f) - \|f\|_2^2$ par la minimisation d'un estimateur $\hat{R}(h)$ de cette quantité. Plus précisément on va chercher un estimateur sans biais de cette expression:

$$\mathbb{E}[\left\|\hat{f}_{n,h}\right\|_2^2] - 2\ \mathbb{E}[\int \hat{f}_{n,h}(x) f(x) dx]$$

Le premier terme admet $\left\|\hat{f}_{n,h}\right\|_2^2$ comme estimateur trivial (d'après la propriété des estimateurs sans biais : $\mathbb{E}[\hat{\beta}] = \beta$).
Il reste à trouver un estimateur sans biais du second terme.

Lemme 2 *Soit $\hat{G}$ définit en tout points sauf en X_i (c'est le principe du Leave-one-out):*

$$\hat{G} = \frac{1}{n} \sum_{i=1}^{n} \hat{f}_{n,h}^{(-i)}(X_i)$$

avec :

$$\hat{f}_{n,h}^{(-i)}(x) = \frac{1}{n-1} \frac{1}{h} \sum_{j=1, j \neq i}^{n} K(\frac{x - X_j}{h})$$

Cette estimateur $\hat{G}$, par construction est l'estimateur sans biais de $\int \hat{f}_{n,h}(x) f(x) dx$.

Démonstration 7 *Montrons que $\mathbb{E}(\hat{G}) = \mathbb{E}[\int \hat{f}_{n,h}(x) f(x) dx]$.*
Comme les X_i sont i.i.d., d'une part nous avons :

$$\begin{aligned} \mathbb{E}[\int \hat{f}_{n,h}(x) f(x) dx] &= \mathbb{E}[\int \frac{1}{nh} \sum_{i=1}^{n} K(\frac{x - X_i}{h}) f(x) dx] \\ &= \frac{1}{h} \mathbb{E}[\int K(\frac{x - x_1}{h}) f(x) dx] \\ &= \frac{1}{h} \int f(x) \int K(\frac{x - X_1}{h}) f(x_1) dx_1 dx \end{aligned}$$

D'autre part, nous avons :

$$\begin{aligned}
\mathbb{E}[\widehat{G}] &= \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n}\hat{f}_{n,h}^{(-i)}(X_i)] = \mathbb{E}[\hat{f}_{n,h}^{(-1)}(X_1)] \\
&= \mathbb{E}[\frac{1}{(n-1)h}\sum_{j\neq 1}K(\frac{X_j - X_1}{h})] \\
&= \mathbb{E}[\frac{1}{h}K(\frac{X - X_1}{h})] \\
&= \frac{1}{h}\int f(x)\int K(\frac{x - x_1}{h})f(x_1)dx_1 dx \\
&= \mathbb{E}[\int \hat{f}_{n,h}(x)f(x)dx]
\end{aligned}$$

Donc, $\hat{G}$ est un estimateur sans biais de $\int \hat{f}_{n,h}(x)f(x)dx$.

Finalement, l'estimateur sans biais de $R(\hat{f}_{n,h}, f) - \|f\|_2^2$ est donné par:

$$\hat{R}(h) = \|\hat{f}_{n,h}\|_2^2 - \frac{2}{n(n-1)}\sum_{i=1}\sum_{j=1,j\neq i}\frac{1}{h}K(\frac{X_i - X_j}{h})$$

On définit alors

$$\hat{h} = arg\ \underset{h\in H}{min}\hat{R}(h)$$

Si ce minimum est atteint. On cherche une fenêtre parmi une grille finie de valeurs, grille qu'on a notée H dans la formule ci-dessus.
L'estimateur $\hat{f}_{n,\hat{h}}$ a de bonnes propriétés pratiques et de consistance.

La validation croisée est une méthode très générale mais nous l'utilisons ici pour le choix la fenêtre h optimale.

3.2.2.2 Méthode de Goldenshluger-Lepski

La méthode de Goldenshluger-Lepski donne principalement des critères de sélection dans une famille d'estimateurs linéaires à noyau, afin d'obtenir un estimateur vérifiant une inégalité d'oracle.
Avant de présenter ces critères de sélection, commençons d'abord par une introduction aux inégalités d'oracle dans l'estimation adaptative.

3.2.2.2.1 Inégalités d'oracle References pour cette partie. et
Supposons que la fonction estimée appartient à une classe fonctionnelle $\mathcal{F}$ et qu'on a un nombre d'observations n fixé.
On a pour objectif de choisir, dans une famille d'estimateurs $\mathbb{F} = \{\hat{f}_h; h \in \mathcal{H}\}$ indexée par le paramètre $h \in \mathcal{H}$, un estimateur $\hat{f}_{h^*}$ qui soit le meilleur possible. Cela revient à résoudre le problème de minimisation

$$h^* = arg\ inf_{h\in\mathcal{H}} R(\hat{f}_h, f)$$

L'estimateur $\hat{f}_{h^*}$ n'est pas calculable en pratique puisqu'il dépend de la fonction inconnue f. C'est aussi pourquoi il est souvent appelé oracle. Le but est donc de se servir de son risque pour trouver un estimateur qui fonctionne presque comme cet oracle. Pour cela nous utilisons l'échantillon des observations pour sélectionner un paramètre $\hat{h} \in \mathcal{H}$ tel que $\hat{f}_{\hat{h}}$ vérifie une égalité d'oracle

$$\mathcal{R}(\hat{f}_{\hat{h}}, f) \leq C\ inf_{h\in\mathcal{H}}\ \mathcal{R}(\hat{f}_h, f) + \delta, \quad \forall\ \ f \in \mathcal{F}.$$

où $C \geq 1$ est une constante indépendante de n et de f, $inf_{h\in\mathcal{H}} R(\hat{f}_h, f)$ est le risque d'oracle et δ un terme résiduel indépendant de f, souvent négligeable devant le risque d'oracle.

Remarque 6 *Lorsque C vaut 1, l'inégalité est dite exacte et $\hat{f}_{\hat{h}}$ imite l'oracle sur $\mathcal{H}$.*
Lorsque $C > 1$, l'estimateur imite seulement la vitesse de convergence de l'oracle. Parfois quand il est difficile de comparer le risque de l'estimateur sélectionné avec celui de l'oracle. on cherche à obtenir une inégalité d'oracle

$$\mathcal{R}(\hat{f}_{\hat{h}}, f) \leq inf_{h\in\mathcal{H}} \mathcal{R}(h, f) + \delta, \quad \forall\ \ f \in \mathcal{F}.$$

Avec $\mathcal{R}(h, f)$ une approximation du risque $\mathcal{R}(\hat{f}_h, f)$

La question qui se pose dans la suite est
Soit $\mathcal{F}$ une collection d'estimateurs construits à partir des données et $\hat{f}_h \to \mathcal{R}(\hat{f}_h, f)$ un risque pour l'estimation de f, comment construire un estimateur $\hat{f}_{\hat{h}}$ tel que $\mathcal{R}(\hat{f}_{\hat{h}}, f) \approx \quad inf_{h\in\mathcal{H}} \mathcal{R}(\hat{f}_h, f)$ où $\mathbb{E}[\mathcal{R}(\hat{f}_{\hat{h}}, f)] \approx inf_{h\in\mathcal{H}} \mathbb{E}[\mathcal{R}(\hat{f}_h, f)]$? C'est là que s'applique la méthode de Goldenshluger et Lepski. Il s'agit une des méthodes usuelles qui imite la décomposition biais-variance du risque de l'estimateur.
Cette méthode se base sur l'observation. Elle consiste à choisir un estimateur dans une famille d'estimateurs linéaires $\mathbb{F} = \{\hat{f}_h, h \in \mathcal{H}\}$. Pour cela on doit imposer d'abord quelques suppositions.

Suppositions

-1) Le noyau K est lipschitzienne

$$|K(x) - K(y)| \leq c|x-y|, \quad \forall (x,y) \in \mathbb{R}.$$

OÙ $|.|$ est la distance euclidienne.
-2)Il existe un réel $k_\infty < \infty$ tel que $||K||_\infty \leq k_\infty$

Passons ensuite au critère de sélection.

Critère de sélection

Ce critère comme abordé au dessus se base sur la comparaisons des estimateurs deux à deux en faisant intervenir des estimateurs auxiliaires $\{\hat{f}_{h,\mu}, h \ \ et \ \ \mu \in \mathcal{H}\}$ définies comme suit

$$\hat{f}_{h,\mu}(x) = \frac{1}{n}\sum_{i=1}^{n}[K_h * K_\mu](x - X_i),$$

où * est le produit de convolution sur $\mathbb{R}$.
On définit aussi

$$\forall h \in \mathcal{H} \quad \hat{\mathcal{R}}_h = sup_{\mu \in \mathcal{H}}[||\hat{f}_{h,\mu} - \hat{f}_\mu||_s - m_s(h,\mu)]_+ + m_s^*(h),$$

où la fonction m_s est appelé le majorant et $m_s^*(h) = sup_{\mu \in \mathcal{H}} m_s(h,\mu) \ \ \forall h \in \mathcal{H}$

Proposition 4 *Soient $\xi_{h,\mu}$ et ξ_μ les erreurs stochastiques relatives aux estimateurs $\hat{f}_{h,\mu}$ et $\hat{f}_\mu$.*
La fonction m_s est une majorante uniforme de la perte aléatoire$||\xi_{h,\mu} - \xi_\mu||_s$

Démonstration 8 *BANDWIDTH SELECTION IN KERNEL DENSITY ESTIMATION: ORACLE INEQUALITIES AND ADAPTIVE MINIMAX OPTIMALITY By ALEXANDER GOLDENSHLUGER1 AND OLEG LEPSKI*

Remarque 7 *La fonction majorante m_s ne dépend pas de la fonction densité f.*

De tout ce qui précède $\hat{h}$ est définie par

$$\hat{h} = arg \ inf_{h \in \mathcal{H}} \hat{\mathcal{R}}_h.$$

Et le critère de comparaison

$$\hat{\Delta}(h) = sup_{\mu \in \mathcal{H}}[||\hat{f}_{h,\mu} - \hat{f}_\mu||_s - m_s(h,\mu)]_+, \quad \forall h, \mu \in \mathcal{H}, \ \ \forall \hat{f} \in \mathbb{F}.$$

Remarque 8 *Soient h et$\mu \in \mathcal{H}$, m_s la fonction majorante définie auparavant et $\widehat{f} \in \mathbb{F}$ m_s est de l'ordre de l'écart type de $|\widehat{f}_{h,\mu}(x) - \widehat{f}_{\mu}(x)|$ pour tout x dans $\mathbb{R}$.*

Donc si l'inégalité suivante est vérifiée

$$sup_{h\in\mathcal{H}}(\mathbb{E}_f \; sup_{\mu\in\mathcal{H}}[||\xi_{h,\mu} - \xi_{\mu}|| - m_s(h,\mu)]_+^2)^{\frac{1}{2}} \leq \delta, \quad \forall f \in \mathbb{F}, \quad \forall h, \mu \in \mathcal{H}.$$

Et si il existe $\widehat{h} \in \mathcal{H}$ mesurable par rapport à l'observation et vérifiant

$$\widehat{\Delta}(\widehat{h}) + sup_{\mu\in\mathcal{H}} m_s(\widehat{h},\mu) \leq inf_{h\in\mathcal{H}}(\widehat{\Delta}(h) + sup_{\mu\in\mathcal{H}} m_s(h,\mu)).$$

Alors l'estimation sectionnée est $\widehat{f}_{\widehat{h}}$.

Chapter 4

Applications

Rappelons que nous cherchons à estimer la fonction de densité f de la durée avant la création d'une nouvelle espèce. Dans ce cadre nous allons faire nos estimations à noyau de la densité sur R en appliquant la théorie que nous avons vue jusque là.

4.1 Fonction dens

Dans cette partie nous présenterons la fonction `dens` que nous avons créée en voulant reproduire ce que fait la fonction `density` de R. Voici son code :

Insérer le script de fonc_dens.R

4.2 Applications aux données du vivant

Maintenant que nous avons notre fonction `dens` nous allons pouvoir la comparer à la fonction `density` de R. Pour les comparer nous allons en profiter pour en même temps les appliquer aux données qu'on veut étudier.

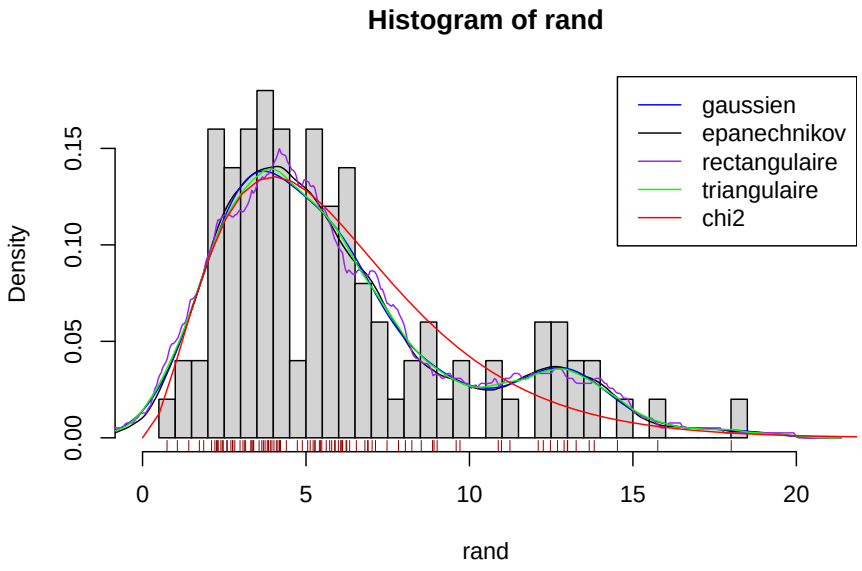
Pour commencer nous allons tester les fonctions avec un premier exemple fictif.

```
# Estimation de la densité par la méthode des noyaux (Test)

## Chi 2 à 6 ddl

seq <- seq(0,30, length.out = 60)
ychi2 <- dchisq(seq,6)
#plot(seq, ychi2, type = "l", col = "red")
```

```
rand <-rchisq(100,6)
hist(rand, breaks = 60, freq = F, xlim=c(0, max(rand)+3))
lines(density(rand, kernel =  "epanechnikov", bw = 1), col ="black")
lines(density(rand),col="blue")
lines(density(rand, kernel =  "rectangular"), col ="purple")
lines(density(rand, kernel =  "triangular"), col ="green")
lines(seq, ychi2, col = "red")
legend("topright",c("gaussien", "epanechnikov", "rectangulaire", "triangulaire", "chi2"
       lty=1, col = c("blue", "black","purple","green","red"))
rug(rand,col="darkred") # Visualisation 1d
```



Comme nous l'avons vu théoriquement, on voit relativement bien ici que la différence au niveau du choix du noyau n'est pas très impactante sur la qualité de l'estimation.

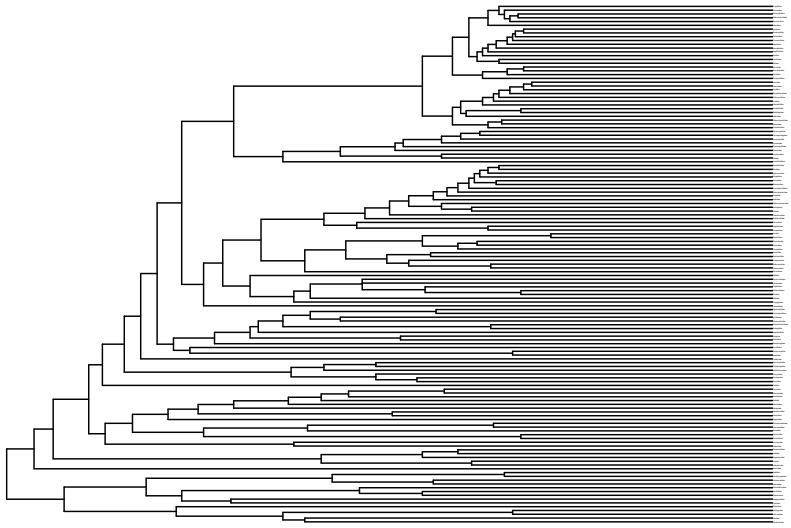
Nous allons maintenant tester les fonctions d'estimation de la densité sur deux exemples. Commençons par les familles d'oiseaux (`bird.families`.

```
# Voici des librairies contenant des arbres phylogénétiques.
library(ape)
library(geiger)

## test avec bird families

### arbre
```

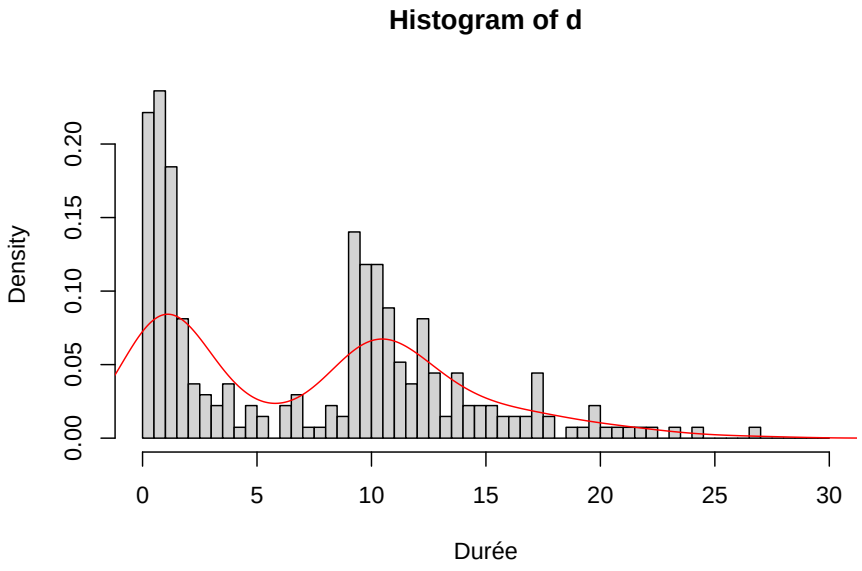
```r
data("bird.families")
op <- par()
par(cex = 0.3)
plot(bird.families)
```



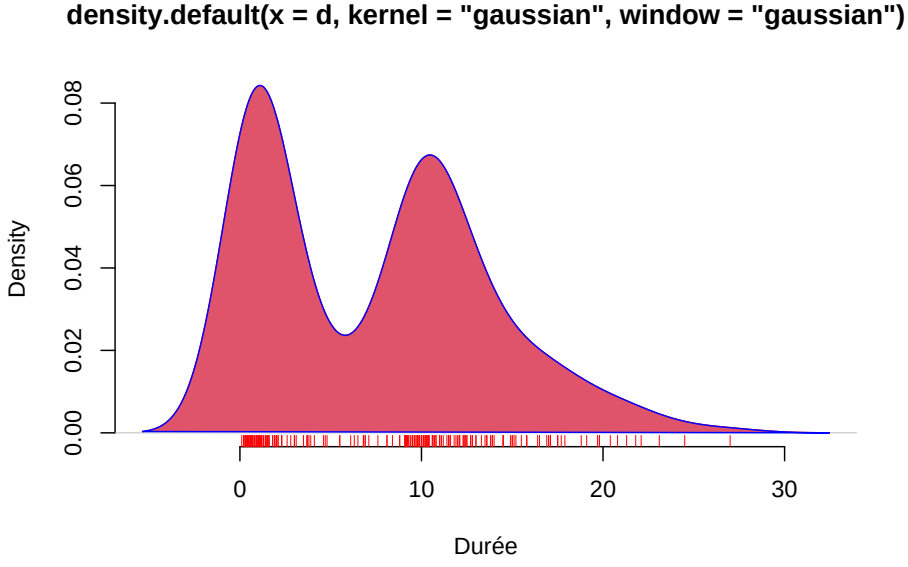
```r
par(cex = op$cex)

#### noyau gaussien

d<-bird.families$edge.length
vec <- pretty(0:(max(d)+3),((max(d)+3)*2))
hist(d, breaks = vec, freq = F, xlab = "Durée")
lines(density(d), col = "red") # bw = 1.82
```




```
plot(density(d, kernel= "gaussian", window = "gaussian"),
     col = "red", bty = "n", xlab = "Durée")
polygon(density(d, kernel= "gaussian"), col=2, border = "blue")
rug(d, col= "red")
```



```r
cat("La moyenne de cet échantillon est de : ", mean(d), "\n")
```

```
## La moyenne de cet échantillon est de :  7.413653
```

```r
cat("La variance de cet échantillon est de : ", var(d)* (length(d)-1)/length(d))
```

```
## La variance de cet échantillon est de :  38.35033
```

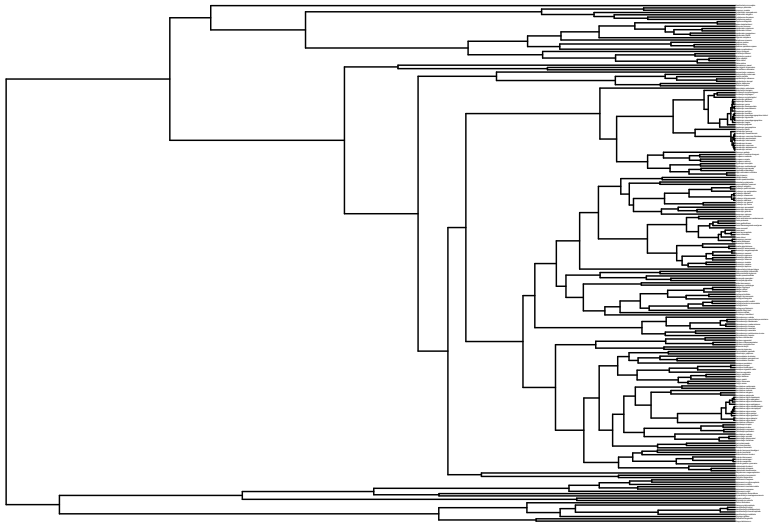
```r
# estimateur biaisé
# max(d) = 27
```

Sur cette échantillon on remarque une répartition bimodale de l'effectif. Effectivement, on observe un mode autour de 1 et un autre autour de 10. On peut supposer en regardant ce tableau que cet échantillon contient des familles d'oiseaux distinctes, ayant suffisamment de différences pour ne pas avoir le même temps d'évolution. C'est pour cela que la moyenne de l'échantillon se trouve au final, là où il y a peu d'individus (moins de 5%). On remarque aussi que les durée sont relativement concentrées avec une variance de 38, elles ne dépassent pas 27.

Maintenant, faisons une autre estimation de la densité sur un échantillon d'espèces de tortues (`chelonia`).

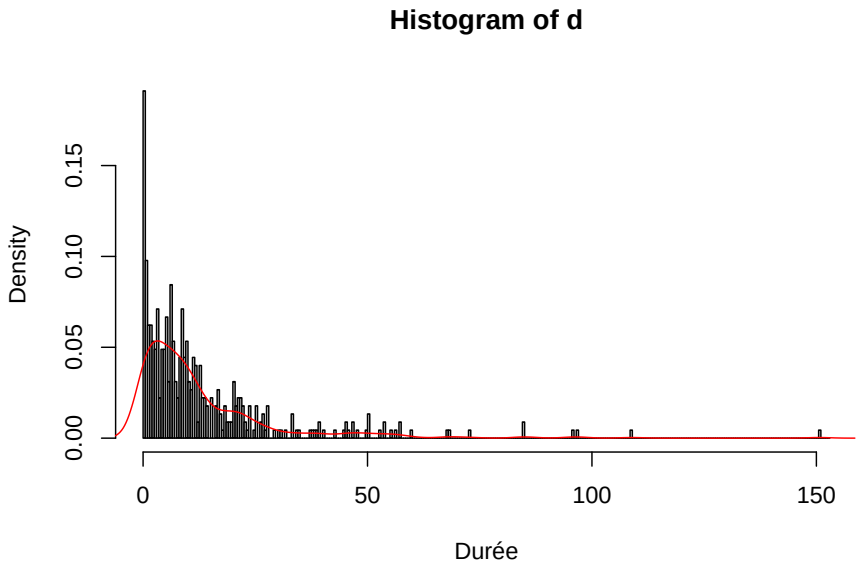
```r
### test avec chelonia (turtules)

#### arbre
data(chelonia)
op <- par()
par(cex = 0.3)
plot(chelonia$phy)
```

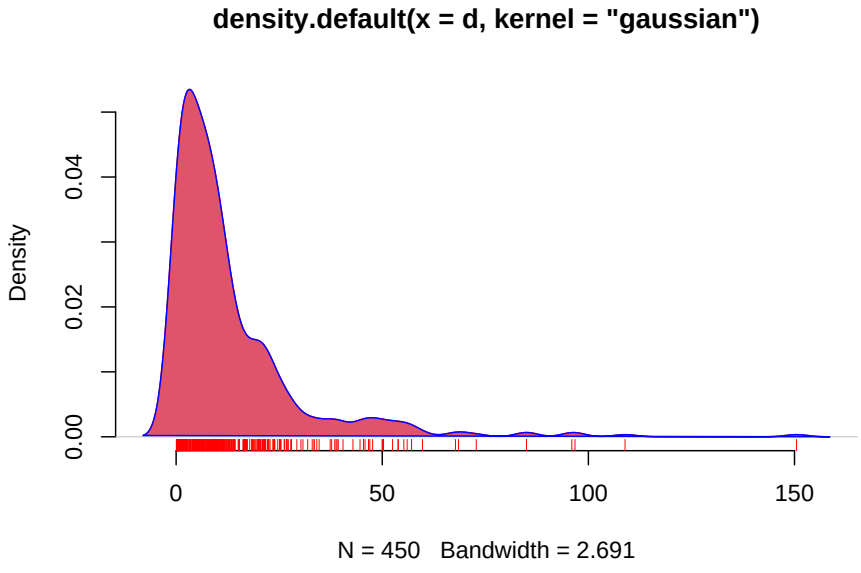


```
par(cex = op$cex)

#### noyau gaussien
d<-chelonia$phy$edge.length
#d1<-chelonia$dat
vec <- pretty(0:(max(d)+3), ((max(d)+3)*2))
hist(d, breaks= vec, freq = F, xlab = "Durée")
lines(density(d), col = "red") # bw = 0.201
```




```
plot(density(d, kernel= "gaussian"), col = "red", bty = "n")
polygon(density(d, kernel= "gaussian"),col=2,border = "blue")
rug(d, col= "red")
```



```r
cat("La moyenne de cet échantillon est de : ", mean(d), "\n")
```

```
## La moyenne de cet échantillon est de :  12.93598
```

```r
cat("La variance de cet échantillon est de : ", var(d)* (length(d)-1)/length(d)) # est
```

```
## La variance de cet échantillon est de :  285.3433
```

```r
# max(d) = 150
```

Si on compare cet échantillon avec le précédent, on remarque cette fois la distribution est plus étalée (avec une variance de 258.3) et unimodale. Cependant la moyenne n'est pas énormément supérieure à celle des familles d'oiseaux. Cela s'explique par le fait que beaucoup d'espèces de tortue sont apparues en moins de 25 ans et qu'on observe le pic autour de 1. On en déduit qu'en moyenne on met plus de temps à obtenir une nouvelle espèce de tortue plutôt qu'une nouvelle espèce d'oiseau.

Chapter 5

Conclusion

Chapter 6

References

Lucie Montuelle. Inégalités d'oracle et mélanges. Statistiques [math.ST]. Université Paris-Sud, 2014. Français.

Gilles Rebelles. Sur l'estimation adaptative d'une densité multivariée sous l'hypothèse de la structure d'indépendance.

Cours-Estimation de densité.https://cedric.cnam.fr/vertigo/Cours/ml/coursEstimationDensite.html