Rapport du projet de M1

Combien de temps pour faire une espèce ?

Wiam Chaoui Sophie Manuel Stéphane Sadio

2021

Contents

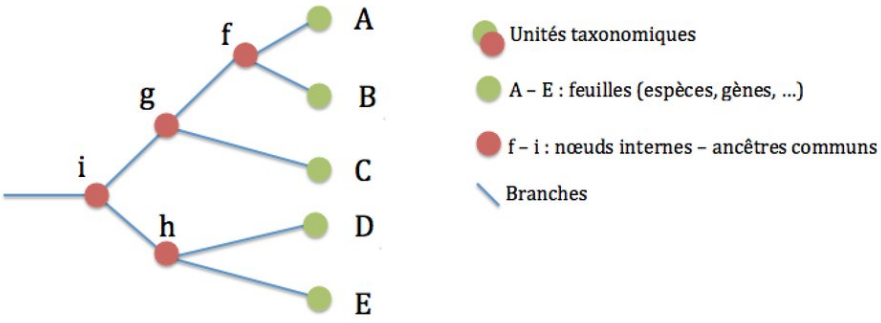
Chapter 1

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 concensus. C'est pour cela que dans la suite nous ne nous étendrons pas sur cette notion et on se concentrera 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 longeur des branches.



Dans la suite, on s'intéresse aux *branchements évolutifs*. On suppose 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 un échantillon $X=\{X_1, \dots, X_n\}$ de longueurs de branche observées qui ont pour une fonction de densité $f \in \mathcal{F}$ où $\mathcal{F}$ est un espace fonctionnel. On cherche à 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 f.

D'où le choix du modèle suivant $\{P = P_f,\ f \in \mathcal{F}\}$ qui revient à faire une estimation non-paramétrique de la densité. Ce qui nous mène à 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, nous introduirons les méthodes d'estimations non-paramétriques, en donnant quelques définitions et en présentant quelques types d'estimateurs. Ensuite, nous allons approfondir sur les estimateurs de densité à noyau en parlant de leur évaluation et des méthodes adaptatives. Enfin, pour répondre à la problématique, on cherchera à implémenter 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 lorsqu'on traite des paramètres à dimension infini. Ce qui est bien notre cas, comme on cherche à estimer une fonction densité qui appartient à un espace fonctionnel.

On présente dans la suite une courte introduction à l'estimation non paramétrique. On introduira ensuite les deux classes principales de l'estimation fonctionnelle (l'estimation par projection et l'estimation à noyau) afin de discuter de ces deux classes et expliquer pourquoi on fait le choix de l'estimation à noyau.

2.1 Définitions

Définition 1 *Estimation non-paramétrique :*
L'estimation non-paramétrique vise à résoudre des problèmes d'estmation 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 infini.

Dans le cadre de notre problématique on s'intéresse a 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

Définition 2 *Pour un échantillon d'observations quantitatives* $X = \{X_1, ..., X_n\}$ *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 $\hat{f}_n$ *l'estimateur de f.*
On se trouve donc avec le modèle suivant $\{\mathbb{P} = \mathbb{P}_f,\ f \in \mathcal{F}\}$, *tel que* $\mathbb{P}_f$ *est la mesure probabilté 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 notera dans la suite* $\hat{f}$ *l'estimateur de la vraie fonction* f.
- *On considerera souvent les distances* L^p *avec* $p = 1, 2$ *ou* ∞ .

Une des premières estimations non-paramétriques de la fonction de densité qui est possible est l'histogramme.

(ajout formule graphique)

L'histogramme fait partie de la famille des estimations à noyau qu'on détaillera plus tard.

Nous traiterons deux grandes familles de méthodes pour estimer une fonction densité :
- L'estimation par projection et - L'estimation par noyau.

2.2 Estimateurs par projection :

Définition 3 *Estimation par projection :*
Supposant que la fonction f *à estimer est dans l'espace de Hilbert* $\mathcal{F} = (L^2, \| . \|, < ., . >)$ *avec* $(\Phi_j)_{j>0}$ *une base orthonormée de* L^2, $\mathbb{E}_N$ *un sous-espace fini de* $\mathcal{F}$ *et* $1 \leq |N| < \infty$.
De plus $a_\lambda =< f, \Phi_\lambda >= \int_{\mathbb{R}} f(x)\Phi_\lambda(x)dx$.
Alors, on estime la fonction f *par son projeté*

$$\Pi_N\ f = \sum_{\lambda \in N} a_\lambda \Phi_\lambda$$

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

Dans la suite on procèdera 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)

2.3 Estimateurs à noyau de densité

Notre but est d'estimer la densité f. Pour cela, on s'appuiera sur un échantillon iid $X = (X_1, ..., X_n)$ 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.

2.3.1 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 répartion 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 :

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

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 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.

2.3.2 Explication ce qu'est un noyau ?

Définition 4 *(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 $\widehat{f}_n$.

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

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

et soient $h < 0$ *le paramètre de lissage.*
$K_h : u \in \mathbb{R} \to K(\frac{u}{h})/h$

Lemme 1 *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 on définit le risque associé d'un estimateur $\hat{f}$ pour l'estimateur f.
Pas compris

Définition 6 *La fonction de risque :*

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

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

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

Nous nous intéressons au risque quadratique ponctuel de $\hat{f}_n$, i.e 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 au carré-variance" du risque quadratique:

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

3.2.1 Majoration du biais et de la variance

Dans cette section, nous allons nous intéresser au compromis biais-variance afin de minimiser le risque quadratique. Les deux propositions suivantes montrent que sous certaines hypothèses, on peut majorer le biais ainsi que la variance.

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

Proposition 1 *: Si $f \in \sum(\beta, L)$ avec $\beta > 0$ et $L > 0$ et si K est un noyau d'ordre $l = \lfloor \beta \rfloor$ 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:*

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

Démonstration 2 *: (voir Esti-non para.pdf page 97, prop 4.10).*

Le biais au carré tend vers zéro à la vitesse $h^{2\beta}$. 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).

Proposition 2 *: 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 $f \in \sum(\beta, L)$ alors

$$\mathbb{V}(\hat{f}_n(x_0)) \leqslant \frac{M(\beta, L)}{nh}$$

Démonstration 3 *:*

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

Et enfin,on admet le résultat suivant :
il existe une constante positive $M(\beta, L)$ tel que $\|f\|_\infty \leqslant M(\beta, L)$. Ceci implique que :

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

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 carrré et la variance. On peut à présent donner un contrôle du risque quadratique par le théorême 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 choissant une fenêtre de la forme $h = cn^{-\frac{1}{2\beta+1}}$ avec une constante $c > 0$, on obtient pour tout $x_0 \in \mathbb{R}$,*

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

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

Démonstration 4 *: On a :*

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

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

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

On cherche ensuite la fenêtre h qui minimise cette quantité. Comme on ne se soucie pas vraiment des constantes exactes quand on cherche la vitesse de convergence d'un estimateur, on utilisera la notation $c_1 = (\frac{L}{l!} \int |u|^\beta |K(u)| du)^2$ et $c_2 = \frac{M(\beta,L) \left\|K\right\|_2^2}{nh}$. On doit alors minimiser en h la quantité :

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

On a une quantité croissante et une quantité décroissante en h. Encore une fois, comme on ne se soucie pas pas des constantes, donc 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. 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'odre 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 positive, on a :

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

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 $n^{\frac{-2\beta}{2\beta+1}}$).

Remarque 4 *: * L'estimatimateur 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 fenêtr ne dépendant que des données et qui est basé sur ce qu'on appelle la validation croisée (ou "cross validation" en Anglais). * Nous avons vu plus haut que le biais au carré tend vers zéro quand h tend vers zéro (si β est 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 consistante.*

Proposition 3 *Dans le cas d'un estimateur à noyau, on a :*

$$R(\hat{f}, f) = \mathbb{E}_f[\| \hat{f} - f \|^2] = \| f - K_h * f \|^2 + \mathbb{E}_f[\| \hat{f} - K_h * f \|^2]$$

Démonstration 5 *On que :*

$\mathbb{E}_f \| \hat{f} - f \|^2 = \mathbb{E}_f[\| \hat{f} + \mathbb{E}_f(\hat{f}) - (\mathbb{E}_f(\hat{f}) - f) \|]^2$.

$\mathbb{E}_f[\| \hat{f} - f \|]^2 = \mathbb{E}_f[\| \hat{f} + \mathbb{E}_f(\hat{f}) \|]^2 + \mathbb{E}_f[\| \mathbb{E}_f(\hat{f}) - f \|]^2 - 2\ \mathbb{E}_f[< \hat{f} - \mathbb{E}_f(\hat{f}); \mathbb{E}_f(\hat{f}) - f >]$.

Comme $\hat{f}$ *est déterministe*
$2\ \mathbb{E}_f(< \hat{f} - \mathbb{E}_f(\hat{f})\ ;\ \mathbb{E}_f(\hat{f}) - f >) = 2 < 0,\ \mathbb{E}_f(\hat{f}) - f >= 0$
Ainsi $\parallel \mathbb{E}_f(\hat{f}) - f \parallel$ *est déterministe*
On obtient :
$R(\hat{f}, f) = \mathbb{E}_f[\parallel \hat{f} - f \parallel^2] = \parallel f - K_h * f \parallel^2 + \mathbb{E}_f[\parallel \hat{f} - K_h * f \parallel^2]$

On a bien trouvé que :

$$R(\hat{f}, f) = biais^2 + \mathbb{V}ar$$

Remarque 5 *Plus la valeur de h est grande, plus le biais devient grand et la variance petite et de même ; plus la valeur de h est petite plus le biais devient petit et la variance explose.*

Donc, afin de minimiser l'expression du risque le choix de h est très influent et même plus crucial pour la qualité de l'estimateur que celui de la noyau K.
On doit chercher le meilleur compromis biais-variance pour avoir un risque minimal.
Un paramètre trop faible provoque l'apparition de détails artificiels sur le graph de l'estimateur (La variance devient trop grande), par contre si on prend une valeur de h très grande on aura la majorité des caractéristiques effacées.

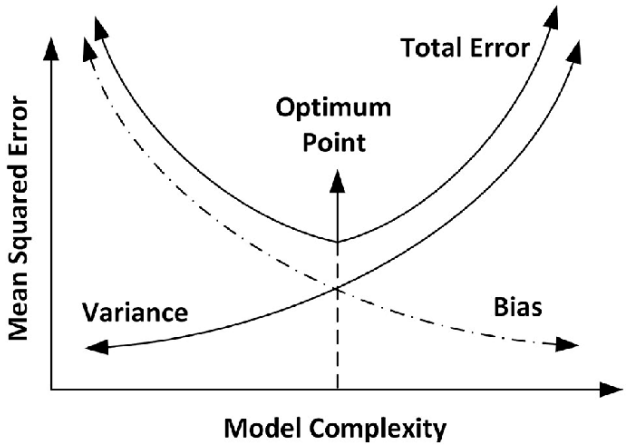


Figure 3.1: Bias-variance-trade-off

(http://chimix.com/an16/pol16/image/aspts35.jpg)

Chapter 4

Méthodes adaptatives :

On a 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 vrai fonction densité f .
La question qui se pose est donc la suivante : comment peut on construire un estimateur à risque optimal à partir de cette famille (en prenant en considération les observations) ?

Dans cette partie et afin de repondre a la question qu'on a poser on va discuter au premier temps du choix du noyau. Ensuite, on va introduire deux méthodes pour le choix du paramètre de lissage h .

4.1 Choix du noyau

4.1.1 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 vraie fonction de densité et la bleue est l'estimation avec la méthode des noyaux sur les variables simulées.

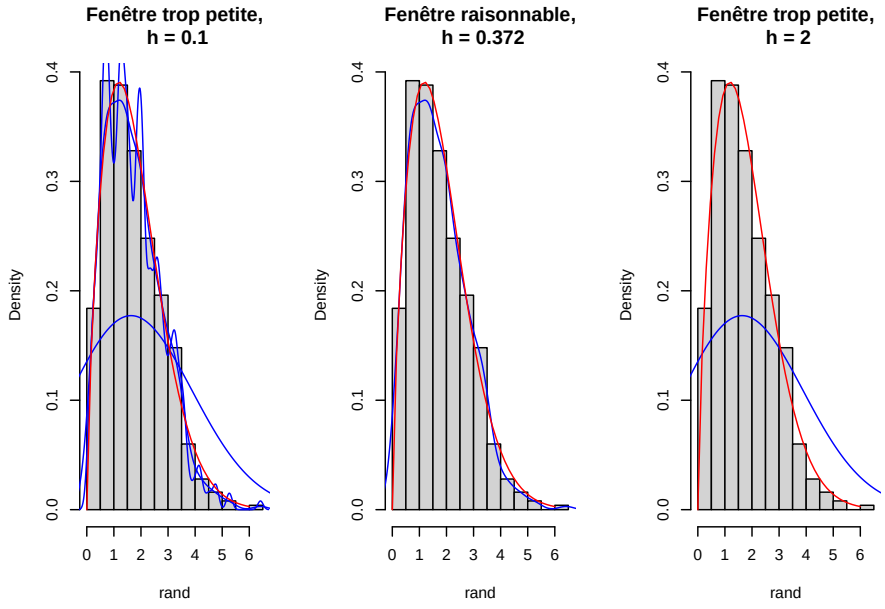
```
par(mfrow=c(1,3))
```

```
seq <- seq(0,6, length.out = 40)
yweib <- dweibull(seq,1.7,2)
rand <-rweibull(500,1.7,2)

hist(rand, breaks = 12, freq = F, main = "")
lines(density(rand, bw = 0.1), col ="blue")
lines(density(rand), col ="blue")
lines(density(rand, bw = 2), col ="blue")
lines(seq, yweib, col="red")
title("Fenêtre trop petite,\n h = 0.1")

hist(rand, breaks = 12, freq = F, main = "")
lines(density(rand), col ="blue")
lines(seq, yweib, col="red")
title("Fenêtre raisonnable,\n h = 0.372")

hist(rand, breaks = 12, freq = F, main = "")
lines(density(rand, bw = 2), col ="blue")
lines(seq, yweib, col="red")
title("Fenêtre trop petite,\n h = 2")
```



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

4.2 Choix de la fenêtre

L'estimation de densité nécessite le choix de la fentêtre qu'on note h.
En statistique non-paramétrique, ils éxistent plusieurs méthodes et critéres de qualité pour le choix de la fêntere.

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

4.2.1 Choix de la fenêtre h par validation croisée

Le choix de la fenêtre dans la section précédnte est criticable: 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 $\widehat{h}$. De façon à souligner la dépendance à la fonction, on va noter $\widehat{f}_{n,h}$ l'estimateur associé à un choix de fenêtre h. L'estimateur final sera $\widehat{f}_{n,\widehat{h}}$, une fois le choix de $\widehat{h}$ fait.
on fait un choix sur h ? On cherche à minimiser en h le risque quadratique pour la distance L_2 :

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

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 tout de suite que minimiser en h la quantité $R(\widehat{f}_{n,h}, f)$ est équivalent à minimiser en h la quantité $R(\widehat{f}_{n,h}, f) - \|f\|_2^2$. On va en fait remplacer la minimisation de la quantité inconnue $R(\widehat{f}_{n,h}, f) - \|f\|_2^2$ par la minimisation d'un estimateur $\widehat{R}(h)$ de cette quantité. Plus précisément on va chercher un estimateur sans biais de cette expression:

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

Le premier terme admet $\left\|\widehat{f}_{n,h}\right\|_2^2$ 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. Pour cela, nous

admettons par construction l'estimateur sans biais $\hat{G}$ défini 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 :

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

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 :

$$\begin{aligned}
\mathbb{E}[\int \hat{f}_{n,h}(x)f(x)dx] &= \mathbb{E}[\int \frac{1}{nh}\sum_{i=1}^{n} K(\frac{x-X_i}{h})f(x)dx] \\
&= \frac{1}{h}\mathbb{E}[\int K(\frac{x-x_1}{h})f(x)dx] \\
&= \frac{1}{h}\int f(x)\int K(\frac{x-X_1}{h})f(x_1)dx_1dx
\end{aligned}$$

D'autre part, nous avons :

$$\begin{aligned}
\mathbb{E}[\hat{G}] &= \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n}\hat{f}_{n,h}^{(-i)}(X_i)] \\
&= \mathbb{E}[\hat{f}_{n,h}^{(-1)}(X_1)] \\
&= \mathbb{E}[\frac{1}{n(n-1)h}\sum_{j\neq 1} K(\frac{X_j-X_1}{h})] \\
&= \mathbb{E}[\frac{1}{h}K(\frac{X-X_1}{h})] \\
&= \frac{1}{h}\int f(x)\int K(\frac{x-x_1}{h})f(x_1)dx_1dx \\
&= \mathbb{E}[\int \hat{f}_{n,h}(x)f(x)dx]
\end{aligned}$$

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

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

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

4.2.2 Méthode de Goldenshluger-Lepski

La méthode de Goldenshluger-Lepski donne principalement des critères pour le choix entre estimateurs à noyau $(\hat{f}_h)_{h \in \mathcal{H}}$ avec différentes fenêtres qu'on fixe en prennant n considération l'échantillon des observations.
Cette méthode propose de choisir le $\hat{h}$ qui minimise l'expression suivante:

$$B(h) + V(h)$$

Avec :

$$B(h) = sup_{h' \in \mathcal{H}}[\| \hat{f}_{h'} - \hat{f}_h \| - V(h')]$$

Et

$$V(h) = a \frac{\| K_{h'} \|^2}{n}$$

Tel que K est le noyau, a un paramètre et $V(h)$ est le terme de pénalisation.

On a donc le $\hat{h}$ est égale à:

$$\hat{h} = argmin_{h \in \mathcal{H}}(B(h) + V(h))$$

Remarque 6 *Le terme de pénalisation choisi est proportionnel à la variance de l'estimateur.*

On s'intéresse dans cette méthode à déterminer le terme de pénalisation minimal $V(h)$ tel que si on le dépasse on n'obtient plus l'équilibre biais-variance.

Dans ce cas, la valeur de $\hat{h}$ est d'ordre $\frac{1}{n}$,

Le choix optimal de la fenêtre h suivant cette méthode dans ce cas est $n^{-\frac{1}{2\alpha+1}}$

Chapter 5

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 applicant la théorie que nous avons vue jusque là.

5.1 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

5.2 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 veux étudier.

Insérer test noyau.R

Estimer loi espérance variance.

Chapter 6

Conclusion