Contents

1 Introduction génerale

1.1 Motivation

On a *une arbre phylogénétique* présentant les relations de parenté entre *espèces* et on s'interesse au *branchement évolutif* apparaît après une durée aleatoire d'une loi fixée μ indépendamment du passé et du future évolutifs des espèces .
Quelle est cette loi μ ? Sa variance ? Sa moyenne ?.

1.2 Problématique

On obsérve les successions de branchements qui composent un arbre phylogénétique et a parir des ces données quantitatives obsérvées on veut a éstimer la fonction f qui donne le temps qu'il faut pour qu'un branchement évolutif apparaisse .
Formellement ; On a un échantillon χ=$\{X_1, ..., X_n\}$ de variables obsérvés qui ont pour fonction de densité f dans $\mathcal{F}$ où $\mathcal{F}$ est un éspace fonctionnel .On cherche a éstimer cette fontion densité f sur la quelle on fait le moins d'hypothéses possibles .D'où le modéle suivant $\{P = P_f, f \in \mathcal{F}\}$ qui revient a faire une éstimation non-paramétrique .
Pour répondre a la problématque on cherchera a comprendre et finalement implémenter un éstimateur à noyau adaptif , de type Goldenshluger-Lepski sur des données d'arbres phylogénétiques .

2 Méthodes non-paramétriques

En statistique ,on parle de l'éstimation quand cherche à trouvé certains paramétres inconus caractérisant une distribution a partir d'un échantillon de données observées en se basant sur differetes méthodes. On se tourne vers l'estimataion non-paramétrique lorsqu'on traite des paramétres à dimension infini . c'e qui est'est bien notre cas, comme on cherche à éstimer une fonction densité qui apartien au espace fonctionnel .

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

2.1 Définitions

Définition 1. *Èstimation non-paramétrique:*
L'éstimation non-paramétrique vise à résoudre des problémes d'éstmation dans le cadre statistique où le modéle dont on s'intéresse n'est pas décrit par un nombre fini de paramètres et dont chacun de ces paramétres ne permet pas de décrire la structure générale de la distribution des variables aléatoires.
Cela signifit qu'on utilise des modéles statistiques à dimension infinis .

Dans le cadre de notre problématique on s'interesse a l'estimation de densité .

Définition 2. *Èstimation à densité:*
Un des proncipes de base de l'éstimation à densité selon une méthode d'éstimation non-paramétrique est le suivant :
Pour un échantillon des obsérvations quantitatives $\mathbb{X} = X_1 ....., X_n$ *des variables aléatoires i.i.d admettant une densité* $f = F'$ *. Supposons que* $f \in \mathcal{F}$ *où* $\mathcal{F}$*un espace fonctionnel on cherche a éstimer la fonction inconnu* f *à partir de ces observations.*
On notera $\hat{f}n$ *l'estimateur de f.*
On se trouve donc avec le modéle suivant $\mathbb{P} \;\; \mathbb{P}_f, f \in \mathcal{F}$ *, tel que* $\mathbb{P}_f$ *est la mesure probabilté de la densité* f *.*

L'éstimation ici concerne donc la fonction elle même plutôt que les paramétres ,ce qui éxplique le nom de l'éstimation non-paramétrique .

Remarque 1. *-On notera dans la suite* $\hat{f}$ *l'éstimateur de la vrai fonction* f*.*
-On prend souvent les distance L^p *avec p =1,2 ou* ∞ *.*

Ils éxiste deux grandes familles de méthodes pour éstimer une fonction densité :
Èstimation par projection .
Èstimation par le noyau .

2.2 Èstimateurs par projéction :

Définition 3. *Èstimation par projection:*
Supposant que la fonction f *à éstimer est dans l'éspace 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 éstime la fonction f *par son projeté*

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

Remarque 2. *-Cette méthode nous raméne au cas paramétrique .*
-Plus la valeur de N est grande plus le biais est petit et la variance est grande .

Dans la suite on procedra avec la méthode la plus fréquente utilisée pour l'éstimation d'une densité : L'éstimation par le noyau.

3 Èstimateur de densité à noyau :

3.1 Èvaluer un éstimateur

Pour évaluer un éstimateur on définit le risque associé d'un éstimateur $\hat{f}$ pour l'éstimateur f.

Définition 4. *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'éstimateur $\hat{f}$ et l'éstimation f .*
On cherche a ce que ce risque associé soit minimal (i.e tend vers 0 pour un nombre d'obsérvation assez grand.)

Proposition. *Dans le cas d'un éstimateur à noyau , on a :*

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

Démonstration. *On que :*

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

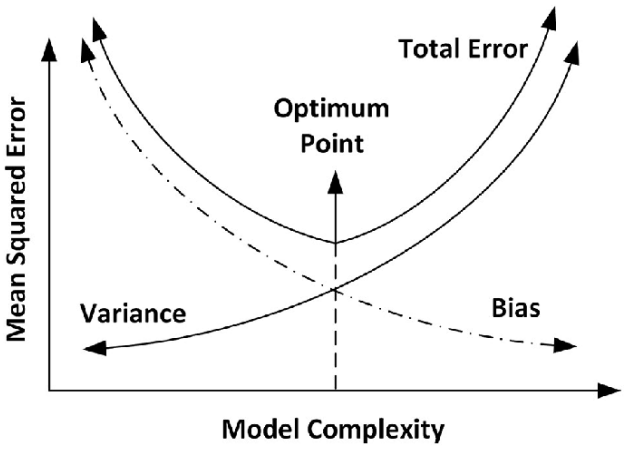
$E_f||\hat{f} - f||]^2 = \mathbb{E}_f||\hat{f} + E_f(\hat{f})||^2 + \mathbb{E}_f||\mathbb{E}_f(\hat{f}) - f||^2 - 2\mathbb{E}_f(< \hat{f} - \mathbb{E}_f(\hat{f}); \mathbb{E}_f(\hat{f}) - f >)$.

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

On à bien trouver que
$R(\hat{f}, f) = biais^2 + \mathbb{V}ar$

Remarque 4. *Plus la valeur de h est grande,plus le biais devient grand et la variance petite et vis-vers-ca ; plus la valeur de h est petite plus le biais devient petit et la variance s'explose.*

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



3.2 Méthodes adabtatives :

On a introduit précedement la notion de l'éstimation de densité qui dépend d'un paramétre de lissage h. Soit $(\hat{f}_h)_{h\in\mathcal{H}}$ une famille des éstimateurs de la vrai fonction densité f .
La question qui se pose est donc la suiva,te : comment peut on construire un éstimateur à risque optimal à partir de cette famille (en prenant en considération les obsérvations) ?

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 introduir deux méthodes pour le choix du paramétre de lissage h .

3.2.1 Choix du noyau

On note par le noyau la fonction intégrable K : $\mathbb{R} \to \mathbb{R}$ tel que , $\int_{\mathbb{R}} K(u)du = 1$.
Soient :
h>0 le paramétre 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. *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)*

3.2.2 Choix de la fenêtre

L'éstimation 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 .

3.2.2.1 Méthode par validation croisée

3.2.2.2 Méthode de Goldenshluger-Lepski La méthode de Lepski donne principalement des critéres pour le choix entre éstimateurs à noyau $(\hat{f}_h)_{h\in\mathcal{H}}$ avec différentes fenêtres qu'on fixe en prennat on considération l'échantillon des obsérvations.
Cette méthode propose de choisir le $\hat{h}$ qui minimise l'éxpression 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 paramétre 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 5. *Le terme de pénalisation choisi est proportionne a la variance de l'éstimateur .*

On s'interesse 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 fenétre h suivant cette méthode dans ce cas est $n^{-\frac{1}{2\alpha+1}}$