

Contents

1	Introduction générale	2
1.1	Motivation	2
1.2	Problématique	2
2	Méthodes non-paramétriques	3
2.1	Définitions	3
2.2	Estimateurs par projection :	3
3	Estimateur de densité à noyau :	4
3.1	Évaluer un estimateur	4
3.1.1	Risque quadratique ponctuel des estimateurs à noyau sur les classe des espaces de Hölder	4
3.2	Méthodes adaptatives :	7
3.2.1	Choix du noyau	8
3.2.2	Choix de la fenêtre	8

1 Introduction générale

1.1 Motivation

On a un *arbre phylogénétique* présentant les relations de parenté entre *espèces* et on s'intéresse au *branchement évolutif* apparaît après une durée aléatoire d'une loi fixée μ indépendamment du passé et du futur évolutifs des espèces .

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

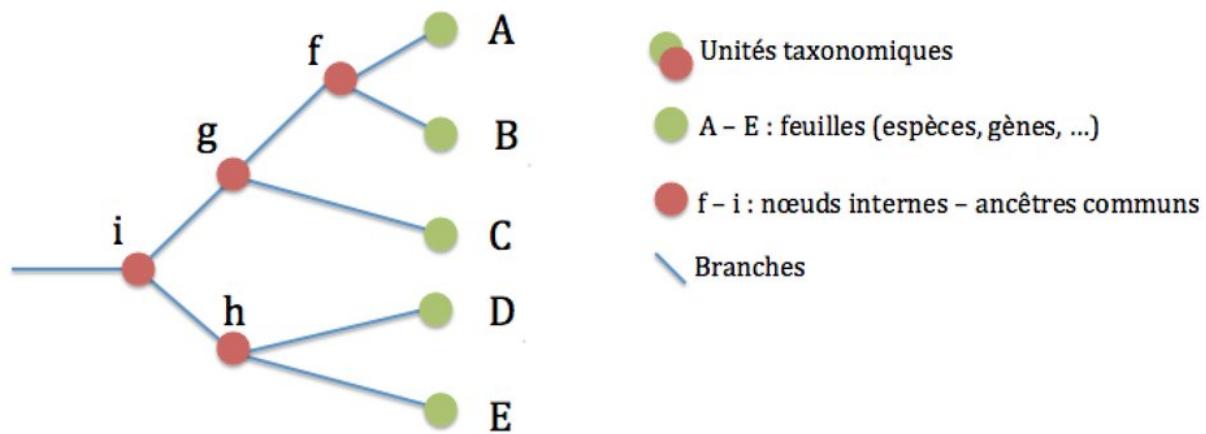


Figure 1: Arbre phylogénétique

1.2 Problématique

On observe des branchements successives qui composent l'arbre phylogénétique et à partir de ces données quantitatives observées, on veut estimer la fonction f qui donne le temps qu'il faut pour qu'un nouveau branchement évolutif apparaisse.

Formellement, On a un échantillon $\chi = \{X_1, \dots, X_n\}$ de variables observées qui ont pour fonction de densité f dans \mathcal{F} où \mathcal{F} est un espace fonctionnel. On cherche à estimer cette fonction de densité f sur la quelle on fait le moins d'hypothèses possibles. D'où le modèle suivant $\{P = P_f, f \in \mathcal{F}\}$ qui revient à faire une estimation non-paramétrique.

Pour répondre à la problématique on cherchera à comprendre et finalement implémenter un estimateur à noyau adaptatif, de type Goldenshluger-Lepski sur des données d'arbres phylogénétiques.

2 Méthodes non-paramétriques

En statistique, on parle de l'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 infinie. Ce qui est bien notre cas, comme on cherche à estimer une fonction densité qui appartient à l'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'estimation par projection et l'estimation à noyau) afin de discuter de ces deux classes et de 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'estimation dans le cadre statistique où le modèle dont 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.

Définition 2. *Un des principes de base de l'estimation à densité selon une méthode d'estimation non-paramétrique est le suivant*

Pour un échantillon des observations quantitatives $\mathbb{X} = X_1, \dots, X_n$ 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 à estimer la fonction inconnue f à partir de ces observations.

On notera \hat{f}_n l'estimateur de f .

On se trouve donc avec le modèle suivant $\mathbb{P} \quad \mathbb{P}_f, f \in \mathcal{F}$, tel que \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 de l'estimation non-paramétrique.

Remarque 1. *-On notera dans la suite \hat{f} l'estimateur de la vraie fonction f .*

-On prend souvent la distance L^p avec $p = 1, 2$ ou ∞ .

Ils existent deux grandes familles de méthodes pour estimer une fonction densité:

- Estimation par projection,
- Estimation par le noyau.

2.2 Estimateurs par projection :

Définition 3. *Estimation par projection:*

Supposant que la fonction f à estimer est dans l'espace de Hilbert $\mathcal{F} = (L^2, \|\cdot\|, \langle \cdot, \cdot \rangle)$ avec $(\Phi_j)_{j \geq 0}$ une base orthonormée de L^2 , \mathbb{E}_N un sous-espace fini de \mathcal{F} et $1 \leq |N| < \infty$.

De plus $a_\lambda = \langle f, \Phi_\lambda \rangle = \int_{\mathbb{R}} f(x) \Phi_\lambda(x) dx$.

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 procédera à la méthode la plus fréquente utilisée pour l'estimation d'une densité: L'estimation par le noyau.

(Argument pour le choix de la méthode à voir)

3 Estimateur de densité à noyau :

3.1 Évaluer 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 $K: \mathbb{R} \rightarrow \mathbb{R}$ tel que:

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

et soient $h > 0$ le paramètre de lissage.

$$K_h : u \in \mathbb{R} \rightarrow K\left(\frac{u}{h}\right)/h$$

Lemme 1. On peut approximer la famille $(K_h)_{h>0}$ par l'identité du produit de convolution.

Démonstration. A Faire

Corollaire 1. $K_h * f : x \rightarrow \int_{\mathbb{R}} K_h(y-x)f(x)dx$ 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 .

Définition 5. 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.)

3.1.1 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:

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

3.1.1.1 Majoration du biais et de la variance. Dans cette section, nous allons nous intéresser au compromis biais-variance afin de minimiser le risque quadatique. Les duex propositions suivantes montrent que sous certaines hypothèses, on peut majorer le biais ainsi que la variance.

Définition 6. : Soit $l \in \mathbb{N}^*$. On dit que le noyau K est d'ordre l si $u^j K(u)$ est intégrable et $\int u^j K(u) du = 0$, $j = 1, \dots, l$.

Proposition. : 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:

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

Démonstration. : (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. : Si f est bornée et si K est de carré intégrable alors

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

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

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

Démonstration. :

$$\mathbb{V}(\hat{f}_n(x_0)) = \mathbb{V}\left(\frac{1}{nh} \sum_{i=1}^n K\left(\frac{X_i - x_0}{h}\right)\right) = \sum_{i=1}^n \mathbb{V}\left(\frac{1}{nh} K\left(\frac{X_i - x_0}{h}\right)\right) = \sum_{i=1}^n \mathbb{V}\left(\frac{1}{nh} K\left(\frac{X_i - x_0}{h}\right)\right) = \sum_{i=1}^n \frac{1}{n^2 h^2} \mathbb{V}\left(K\left(\frac{X_i - x_0}{h}\right)\right) = \frac{1}{nh^2} \mathbb{V}\left(K\left(\frac{X_i - x_0}{h}\right)\right)$$

Et enfin, on admet le résultat suivant:

il existe une constante positive $M(\beta, L)$ tel que $\|f\|_\infty \leq M(\beta, L)$. Ceci implique que

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

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 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 $x_0 \in \mathbb{R}$,

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

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

Démonstration. : On a

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

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

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

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 $c_1 = \left(\frac{L}{l!} \int |u|^\beta |K(u)| du \right)^2$ et $c_2 = \frac{M(\beta, L) \|K\|_2^2}{nh}$. 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 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'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 positive, on a

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

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

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

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'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).

-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. Dans le cas d'un estimateur 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 :

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

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

Comme \hat{f} est déterministe

$$2E_f(< \hat{f} - E_f(\hat{f}); E_f(\hat{f}) - f >) = 2 < 0, E_f(\hat{f}) - f > = 0$$

Ainsi $||E_f(\hat{f}) - f||$ est déterministe

On obtient :

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

On à bien trouver que

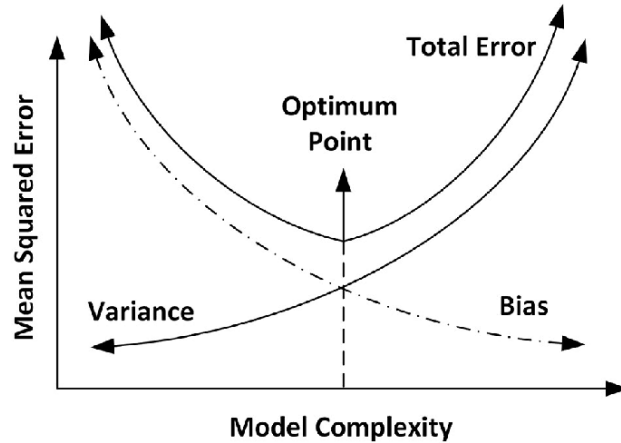
$$R(\hat{f}, f) = \text{biais}^2 + \text{Var}$$

Remarque 5. Plus la valeur de h est grande, plus le biais devient grand et la variance petite et vis-versa ; plus la valeur de h est petite plus le biais devient petit et la variance s'explode.

Donc afin de minimiser l'expression de 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 afin 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 effacée.



3.2 Méthodes adaptatives :

On a introduit précédement la notion de l'estimation de densité qui dépend d'un paramètre de lissage h . Soit $(\hat{f}_h)_{h \in \mathcal{H}}$ une famille des estimateurs de la vraie 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 répondre à la question qu'on a posée on va discuter au premier temps du choix du noyau. Ensuite, on va introduire deux méthodes pour le choix du paramètre de lissage h .

3.2.1 Choix du noyau

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, il existe 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 de 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 :

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

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 $R(\hat{f}_{n,h}, f)$ 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:

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

Le premier terme admet $\|\hat{f}_{n,h}\|_2^2$ 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

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

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

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

D'autre part, nous avons

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

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:

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

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 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 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 prenat on 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:

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

Remarque 6. Le terme de pénalisation choisi est proportionne a 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 $n^{-\frac{1}{2\alpha+1}}$