⋆⋆⋆ SELECTION DE MODELES ⋆⋆⋆
Estimateur par Histogramme.
⋆⋆⋆ Estimation de Densités ⋆⋆⋆

Alexandre LEKINA

Groupe de travail MISTIS, INRIA Rhône-Alpes

Mémoire pour l'obtention du Master 2R MASS
Sous la direction de Y. BARAUD (Laboratoire JA. Dieudonné (CNRS))
Soutenu le 11/09/2007 - Version longue et révisée du 19/10/07

Généralités

- *L'histogramme est un moyen simple et rapide pour représenter la distribution d'un paramètre obtenu lors d'une fabrication.*
  - *concentration d'un élément dans la composition d'alliages produit par une fonderie;*
  - *masse de préparation alimentaire dans une boîte de conserve.*

Généralités

- *L'histogramme est un moyen simple et rapide pour représenter la distribution d'un paramètre obtenu lors d'une fabrication.*
  - *concentration d'un élément dans la composition d'alliages produit par une fonderie;*
  - *masse de préparation alimentaire dans une boîte de conserve.*
- *Son utilisation est aussi courante en archéologie. Il permet entre autre d'évaluer l'âge des fossiles humains - de détecter certaines anomalies - ou de faire un diagnostic.*

Généralités

- *L'histogramme est un moyen simple et rapide pour représenter la distribution d'un paramètre obtenu lors d'une fabrication.*
  - *concentration d'un élément dans la composition d'alliages produit par une fonderie;*
  - *masse de préparation alimentaire dans une boîte de conserve.*
- *Son utilisation est aussi courante en archéologie. Il permet entre autre d'évaluer l'âge des fossiles humains - de détecter certaines anomalies - ou de faire un diagnostic.*

Définition

- *Il est défini comme étant l'estimateur le plus naïf de la fonction de densité.*

Plan

- Nous allons nous restreindre à l'étude de l'estimation d'une densité s sur $[0, 1]$.

- Nous allons nous restreindre à l'étude de l'estimation d'une densité s sur $[0,1]$.
- Nous supposerons que nos observations sont modélisées par n variables aléatoires $X_1, X_2, .., X_n$ $i.i.d$ de densité s sur $[0,1]$.

- Nous allons nous restreindre à l'étude de l'estimation d'une densité s sur $[0,1]$.
- Nous supposerons que nos observations sont modélisées par n variables aléatoires $X_1, X_2, .., X_n$ $i.i.d$ de densité s sur $[0,1]$.
- On définira N comme étant la mesure empirique associée à notre n-échantillon

$$\forall\ I \subseteq [0,1],\ N(I) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{X_i \in I}. \qquad (1)$$

- Nous allons nous restreindre à l'étude de l'estimation d'une densité s sur $[0,1]$.
- Nous supposerons que nos observations sont modélisées par n variables aléatoires $X_1, X_2, .., X_n$ $i.i.d$ de densité s sur $[0,1]$.
- On définira N comme étant la mesure empirique associée à notre n-échantillon

$$\forall\ I \subseteq [0,1],\ N(I) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{X_i \in I}. \tag{1}$$

- On définira $\mathcal{L}$ comme étant l'ensemble des fonctions t mesurables et positives sur $[0,1]$ tel que:

$$\int_0^1 t d\lambda < +\infty. \tag{2}$$

Plan

Définition

Etant donné une partition m de $[0,1]$*, on définit le modèle histogramme* S_m *par*

$$S_m = \left\{ t = \sum_{I \in m} t_I \mathbb{1}_I \ \backslash \ t_I \geq 0 \ \text{ et pour } I \in m, \lambda(I) > 0 \right\}. \quad (3)$$

Définition

La fonction t définie sur $[0,1]$ *est une fonction constate par morceaux si elle est constante sur un certain nombre d'intervalles de* $[0,1]$ *puis fait des sauts de temps en temps sur les autres intervalles. Elle a alors pour forme*

$$t = \sum_{k=1}^{D} t_k \mathbb{1}_{[x_{k-1}, x_k]} \quad \text{avec} \quad 0 = x_0 < x_1 < \ldots < x_D = 1. \quad (4)$$

Définition

On appelle histogramme (ou estimateur par histogramme) basé sur la partition m, l'élément $\hat{s}_m$ de S_m défini par

$$\hat{s}_m = \sum_{I \in m} \frac{N(I)}{\lambda(I)} \mathbb{1}_I. \tag{5}$$

Remarque

$\hat{s}_m$ peut s'annuler si un élément de la partition m ne contient pas d'observations.

Plan

Définition

La distance de Hellinger entre deux élements t et $t' \in \mathcal{L}$ est définie par

$$H(t,t') = \left[\int_0^1 \left(\sqrt{t} - \sqrt{t'}\right)^2 d\lambda\right]^{1/2}. \qquad (6)$$

Définition

La distance de Hellinger entre deux élements t et $t' \in \mathcal{L}$ est définie par

$$H(t,t') = \left[\int_0^1 \left(\sqrt{t} - \sqrt{t'}\right)^2 d\lambda\right]^{1/2}. \qquad (6)$$

On définit $\bar{s}_m \in S_m$ par

$$\bar{s}_m = \mathbb{E}[\hat{s}_m] = \sum_{\mathrm{I}\in\mathrm{m}} \frac{s_I}{\lambda(I)} \mathbb{1}_I \quad \text{avec} \quad s_I = \int_I s d\lambda.$$

Définition

La distance de Hellinger entre deux élements t et $t' \in \mathcal{L}$ est définie par

$$H(t,t') = \left[\int_0^1 \left(\sqrt{t} - \sqrt{t'}\right)^2 d\lambda\right]^{1/2}. \qquad (6)$$

On définit $\bar{s}_m \in S_m$ par

$$\bar{s}_m = \mathbb{E}[\hat{s}_m] = \sum_{I \in m} \frac{s_I}{\lambda(I)} \mathbb{1}_I \quad \text{avec} \quad s_I = \int_I s d\lambda.$$

Lemme

Pour tout $s \in \mathcal{L}$, nous avons:

$$H^2(s,\bar{s}_m) \leq 2H^2(s,S_m) \quad \textit{avec} \quad \bar{s}_m = \sum_{I \in m} \left(\int_I s \frac{d\lambda}{\lambda(I)}\right) \mathbb{1}_I.$$

Remarque

Au vu lemme, $\bar{s}_m$ réalise presque le minimum de s à S_m car

$$H^2(s, \bar{s}_m) \leq 2H^2(s, S_m) = 2 \inf_{t \in S_m} H^2(s, t)$$

C'est donc à juste titre que $\bar{s}_m$ jouera le rôle de l'approximateur naturel de s dans S_m.

Définition

Soit $\hat{q}$ un estimateur de q et L une fonction de perte, la fonction de risque de l'estimateur $\hat{q}$ associé à L est $R(q, \hat{q}) = \mathbb{E}[L(q, \hat{q})]$.

Pour évaluer le risque, le statisticien recourt deux approches.

Pour évaluer le risque, le statisticien recourt deux approches.

1. L'approche paramétrique qui consiste à faire l'hypothèse que la densité s appartient à un modèle S_m connu par avance. Cette approche se révèle par conséquent insuffisante quand une telle information n'est pas disponible.

Pour évaluer le risque, le statisticien recourt deux approches.

1. L'approche paramétrique qui consiste à faire l'hypothèse que la densité s appartient à un modèle S_m connu par avance. Cette approche se révèle par conséquent insuffisante quand une telle information n'est pas disponible.
2. L'approche non paramétrique permet quant à elle de tenir compte du fait qu'il est possible de commettre deux erreurs lorsqu'on fait l'hypothèse que la densité s appartient à un modèle S_m:

Pour évaluer le risque, le statisticien recourt deux approches.

1. L'approche paramétrique qui consiste à faire l'hypothèse que la densité s appartient à un modèle S_m connu par avance. Cette approche se révèle par conséquent insuffisante quand une telle information n'est pas disponible.
2. L'approche non paramétrique permet quant à elle de tenir compte du fait qu'il est possible de commettre deux erreurs lorsqu'on fait l'hypothèse que la densité s appartient à un modèle S_m:
   - l'erreur d'approximation (induite de l'hypothèse que $s \in S_m$) qui traduira la qualité de l'ajustement au modèle S_m; elle sera représente par la distance de s à S_m;

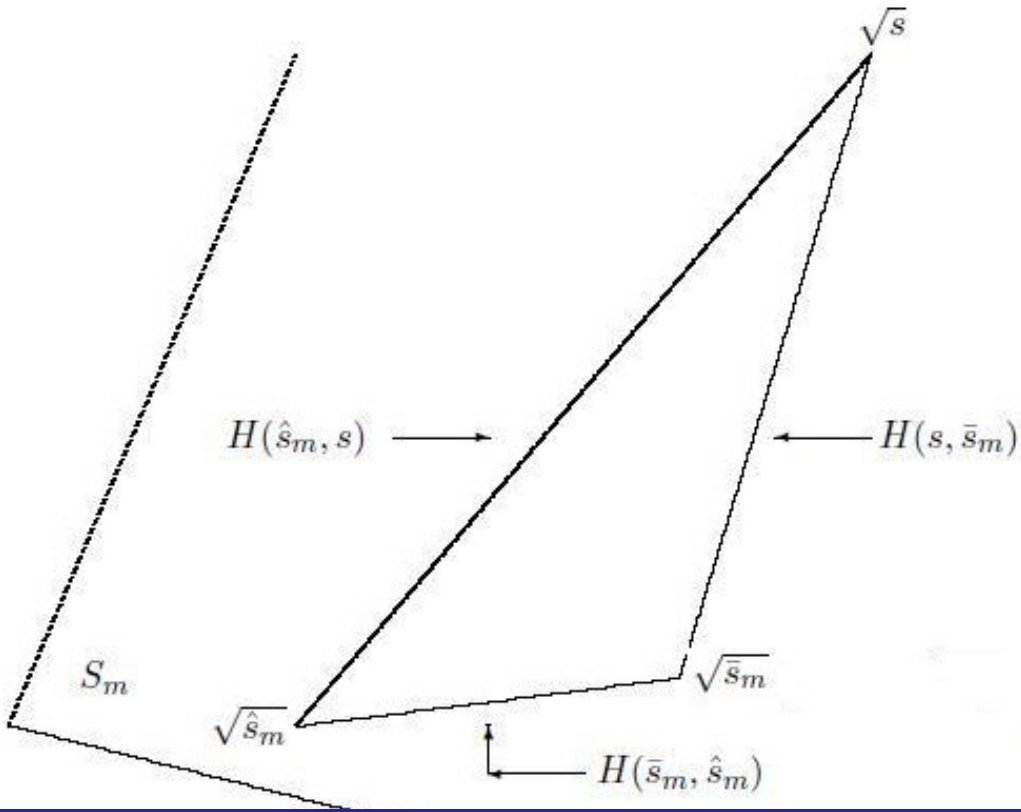
Pour évaluer le risque, le statisticien recourt deux approches.

1. L'approche paramétrique qui consiste à faire l'hypothèse que la densité s appartient à un modèle S_m connu par avance. Cette approche se révèle par conséquent insuffisante quand une telle information n'est pas disponible.
2. L'approche non paramétrique permet quant à elle de tenir compte du fait qu'il est possible de commettre deux erreurs lorsqu'on fait l'hypothèse que la densité s appartient à un modèle S_m:
   - l'erreur d'approximation (induite de l'hypothèse que $s \in S_m$) qui traduira la qualité de l'ajustement au modèle S_m; elle sera représente par la distance de s à S_m;
   - l'erreur d'estimation dans le modèle S_m représentée elle, par la distance de $\bar{s}_m$ à $\hat{s}_m$.

- Quelle approche adopter ?

♣ Quelle approche adopter ?

Afin d'analyser le risque de notre histogramme, notre méthode consistera à faire une projection au sens de Hellinger sur le modèle $S_m = vect\{\mathbb{1}_I, I \in m\}$.



- l'inégalité triangulaire: $H(s, \hat{s}_m) \leq H(s, \bar{s}_m) + H(\bar{s}_m, \hat{s}_m)$;

- l'inégalité triangulaire: $H(s, \hat{s}_m) \leq H(s, \bar{s}_m) + H(\bar{s}_m, \hat{s}_m)$;
- lemme $\oplus$ l'identité $\forall\ a > 0,\ 2xy \leq ax^2 + a^{-1}y^2$:

- l'inégalité triangulaire: $H(s,\hat{s}_m) \leq H(s,\bar{s}_m) + H(\bar{s}_m,\hat{s}_m)$;
- lemme $\oplus$ l'identité $\forall\ a>0,\ 2xy \leq ax^2 + a^{-1}y^2$:

$$H^2(s,\hat{s}_m) \leq 2H^2(s,\bar{s}_m)+2H^2(\bar{s}_m,\hat{s}_m) \leq 4H^2(s,S_m)+2H^2(\bar{s}_m,\hat{s}_m)$$

- l'inégalité triangulaire: $H(s,\hat{s}_m) \leq H(s,\bar{s}_m) + H(\bar{s}_m,\hat{s}_m)$;
- lemme $\oplus$ l'identité $\forall\ a>0,\ 2xy \leq ax^2 + a^{-1}y^2$:

$$H^2(s,\hat{s}_m) \leq 2H^2(s,\bar{s}_m)+2H^2(\bar{s}_m,\hat{s}_m) \leq 4H^2(s,S_m)+2H^2(\bar{s}_m,\hat{s}_m)$$

- passage à l'expérence:

$$\begin{aligned}
\mathbb{E}\left[H^2(\hat{s}_m,s)\right] &\leq 4H^2\left(s,S_m\right) + 2\mathbb{E}\left[H^2(\hat{s}_m,\bar{s}_m)\right] \\
&\leq 4H^2\left(s,S_m\right) + \frac{2}{n}\left[|m|-1\right]. \\
&\leq 4H^2\left(s,S_m\right) + 2|m|.
\end{aligned}$$

On a donc

$$\mathbb{E}\left[H^2(\hat{s}_m, s)\right] \leq 4H^2\left(s, S_m\right) + 2\frac{|m|}{n}. \tag{7}$$

On a donc

$$\mathbb{E}\left[H^2(\hat{s}_m, s)\right] \leq 4H^2\left(s, S_m\right) + 2\frac{|m|}{n}. \tag{7}$$

1. le terme non linaire $H^2\left(s, S_m\right)$ s'appelle le terme du biais

On a donc

$$\mathbb{E}\left[H^2(\hat{s}_m, s)\right] \leq 4H^2\left(s, S_m\right) + 2\frac{|m|}{n}. \tag{7}$$

1. le terme non linaire $H^2\left(s, S_m\right)$ s'appelle le terme du biais
   - il traduit la qualité d'approximation du modèle S_m;
   - il décroît avec la dimension du modèle S_m notée $|m|$;

On a donc

$$\mathbb{E}\left[H^2(\hat{s}_m, s)\right] \leq 4H^2\left(s, S_m\right) + 2\frac{|m|}{n}. \tag{7}$$

1. le terme non linaire $H^2\left(s, S_m\right)$ s'appelle le terme du biais
   - il traduit la qualité d'approximation du modèle S_m;
   - il décroît avec la dimension du modèle S_m notée $|m|$;
2. le terme $\mathbb{E}[H^2(\hat{s}_m, \bar{s}_m)]$ s'appelle le terme de la variance

On a donc

$$\mathbb{E}\left[H^2(\hat{s}_m, s)\right] \leq 4H^2\left(s, S_m\right) + 2\frac{|m|}{n}. \tag{7}$$

1. le terme non linaire $H^2\left(s, S_m\right)$ s'appelle le terme du biais
   - il traduit la qualité d'approximation du modèle S_m;
   - il décroît avec la dimension du modèle S_m notée $|m|$;

2. le terme $\mathbb{E}[H^2(\hat{s}_m, \bar{s}_m)]$ s'appelle le terme de la variance
   - il traduit l'erreur d'estimation dans le modèle;
   - il croît avec $|m|$ - il est proportionel à $|m|/n$.

On a donc

$$\mathbb{E}\left[H^2(\hat{s}_m, s)\right] \leq 4H^2\left(s, S_m\right) + 2\frac{|m|}{n}. \tag{7}$$

1. le terme non linaire $H^2\left(s, S_m\right)$ s'appelle le terme du biais
   - il traduit la qualité d'approximation du modèle S_m;
   - il décroît avec la dimension du modèle S_m notée $|m|$;

2. le terme $\mathbb{E}[H^2(\hat{s}_m, \bar{s}_m)]$ s'appelle le terme de la variance
   - il traduit l'erreur d'estimation dans le modèle;
   - il croît avec $|m|$ - il est proportionel à $|m|/n$.

Remarque

- *Si $s \in S_m$, alors $s = \bar{s}_m$ et par conséquent $H^2\left(s, S_m\right) = 0$.*
- *L'approche non paramétrique englobe l'approche paramétrique.*

Proof.

$$\mathbb{E}\left[H^2(\hat{s}_m, \bar{s}_m)\right] = \mathbb{E}\left[\sum_{I \in m}\left(\sqrt{N(I)} - \sqrt{\mathbb{E}[N(I)]}\right)^2\right]$$

Proof.

$$\begin{aligned}\mathbb{E}\left[H^2(\hat{s}_m, \bar{s}_m)\right] &= \mathbb{E}\left[\sum_{I\in m}\left(\sqrt{N(I)} - \sqrt{\mathbb{E}[N(I)]}\right)^2\right] \\ &\leq \sum_{I\in m}\frac{Var(N(I))}{\mathbb{E}[N(I)]}\end{aligned}$$

Proof.

$$\begin{aligned}
\mathbb{E}\left[H^2(\hat{s}_m, \bar{s}_m)\right] &= \mathbb{E}\left[\sum_{I\in m}\left(\sqrt{N(I)} - \sqrt{\mathbb{E}[N(I)]}\right)^2\right] \\
&\leq \sum_{I\in m}\frac{Var(N(I))}{\mathbb{E}[N(I)]} \\
&= \sum_{I\in m}\left\{\sum_{i=1}^{n}\frac{P(X_i\in I)(1-P(X_i\in I))}{n}\right\}\frac{1}{P(X_i\in I)}
\end{aligned}$$

Proof.

$$
\begin{aligned}
\mathbb{E}\left[H^2(\hat{s}_m, \bar{s}_m)\right] &= \mathbb{E}\left[\sum_{I\in m}\left(\sqrt{N(I)} - \sqrt{\mathbb{E}[N(I)]}\right)^2\right] \\
&\leq \sum_{I\in m}\frac{Var(N(I))}{\mathbb{E}[N(I)]} \\
&= \sum_{I\in m}\left\{\sum_{i=1}^{n}\frac{P(X_i\in I)(1-P(X_i\in I))}{n}\right\}\frac{1}{P(X_i\in I)} \\
&= \frac{1}{n}\sum_{I\in m}(1-P(X_i\in I)) \\
&= \frac{1}{n}\left[|m|-1\right] \leq |m|.
\end{aligned}
$$

- la variance et le biais varie en sens contraire;

- la variance et le biais varie en sens contraire;

But

Nous aimerions avoir un modèle dit idéal (oracle) qui puisse réaliser le meilleur compromis possible entre les deux termes. Or ce dernier (modèle idéal) dépend de s, nous ne pouvons donc pas l'utiliser pour construire l'histogramme.
Notre exercice consistera à construire partir de notre n-échantillon et sans trop faire d'hypothses sur s, un bon modèle, qui se comportera tout aussi bien que le modèle idéal. Pour y parvenir, nous allons construire une série de modèles (famille de modèles) et à chaque modèle construit, nous associerons un estimateur $\hat{s}_m$.
Grâce à un outil de comparaison, algorithmique en l'occurrence, nous confronterons ces estimateurs les uns par rapport aux autres, afin d'en choisir le meilleur.
C'est ce qu'on appelle faire de la Sélection de Modèle.

Plan

Dans tout ce qui suit, on considère que $\mathcal{M}$ est une famille de partitions et $\{S_m, m \in \mathcal{M}\}$ une famille de modèles à laquelle on associe les estimateurs par projection au sens de la distance de Hellinger $\{\hat{s}_m, m \in \mathcal{M}\}$.

Dans tout ce qui suit, on considère que $\mathcal{M}$ est une famille de partitions et $\{S_m, m \in \mathcal{M}\}$ une famille de modèles à laquelle on associe les estimateurs par projection au sens de la distance de Hellinger $\{\hat{s}_m, m \in \mathcal{M}\}$.

But

Idéalement on aimerait avoir $m = m^*$ dans S_m t.q

$$\begin{aligned} \mathbb{E}\left[H^2\left(s, \hat{s}_{m^*}\right)\right] &= \inf_{m \in \mathcal{M}} \mathbb{E}\left[H^2\left(s, \hat{s}_m\right)\right] \\ &= \inf_{m \in \mathcal{M}} \mathbb{E}\left[H^2\left(s, \bar{s}_m\right) + \frac{\chi^2(m)}{n}\right] \end{aligned}$$

Par définition, $\chi^2(m) = nH^2\left(\hat{s}_m, \bar{s}_m\right)$

Problème

m^* est inconnue et dépend de s.

Solution

Trouver un moyen de sélection parmi $\mathcal{M}$ qui ne dépend que $X_1, X_2, .., X_n$ de manière ce que $\hat{m} = \hat{m}(X_1, X_2, .., X_n)$ vérifie

$$\begin{aligned} \mathbb{E}\left[H^2\left(s, \hat{s}_{\hat{m}}\right)\right] &\leq C \inf_{m \in \mathcal{M}} \mathbb{E}\left[H^2\left(s, \hat{s}_m\right)\right] \\ &\leq C \inf_{m \in \mathcal{M}} \left(H^2\left(s, S_m\right) + C_0 \frac{|m|}{n}\right) \end{aligned}$$

avec C_0 et C des constantes indépendantes de s et n.

♢ Enoncé de la statistique de test

- *on considère une famille de classes*
  $\bar{\mathcal{M}} = \{m \vee m' \text{ pour } m, m' \in \mathcal{M}\}$ *avec*

  $m \vee m' = \{I \cap I' | I \in m,\ I' \in m'\}$ *une partition finie de* $[0,1]$;

- $m \vee m'$ *on associe le modèle* $S_{m \vee m'}$;
- $S_{m \vee m'}$ *on associe l'estimateur* $\hat{s}_{m \vee m'}$;

♢ Enoncé de la statistique de test

- *on considère une famille de classes* $\bar{\mathcal{M}} = \{m \vee m' \text{ pour } m, m' \in \mathcal{M}\}$ *avec*

  $m \vee m' = \{I \cap I' | I \in m,\ I' \in m'\}$ *une partition finie de* $[0,1]$;

- $m \vee m'$ *on associe le modèle* $S_{m \vee m'}$;
- $S_{m \vee m'}$ *on associe l'estimateur* $\hat{s}_{m \vee m'}$;
- **H:** $\exists\ \delta \geq 1$ *t.q* $|m \vee m'| \leq \delta\,(|m| + |m'|) \quad \forall\ (m, m') \in \mathcal{M}^2$;
- **H':** $\exists$ *trois constantes positives a, b et* $c \geq 0$ *t.q* $\forall\ m \in \mathcal{M}$,

  $$\mathbb{P}\left[H^2\left(\hat{s}_m, \bar{s}_m\right) \geq c|m| + bz\right] \leq ae^{-z} \quad \text{pour tout} \quad z \geq 0;$$

♢ Enoncé de la statistique de test

♢ Enoncé de la statistique de test

- *on introduit une fonction* $pen : \mathcal{M} \longrightarrow \mathbb{R}^+$ *t.q*

$$pen(m) = c\delta|m| + b\Delta_m \tag{8}$$

avec $\{\Delta_m,\ m \in \mathcal{M}\}$ *une famille de poids (positifs) choisie*

$$\sum_{m \in \mathcal{M}} e^{-\Delta_m} = \Sigma < +\infty. \tag{9}$$

♢ Enoncé de la statistique de test

♢ Enoncé de la statistique de test

$$\forall\ m \neq m',$$
$$T(N) = H^2\left(\hat{s}_{\mathrm{m}}, \hat{s}_{\mathrm{m}\vee\mathrm{m}'}\right) - H^2\left(\hat{s}_{\mathrm{m}'}, \hat{s}_{\mathrm{m}\vee\mathrm{m}'}\right) + 16\left[pen\left(m\right) - pen\left(m'\right)\right]$$

- *si* $T_{\mathrm{m},\mathrm{m}'} < 0$ *préférer le modèle* m*;*
- *si* $T_{\mathrm{m},\mathrm{m}'} > 0$ *préférer le modèle* m'*;*
- *sinon préférer alatoirement le modèle* m *ou* m'*.*

♢ Enoncé de la statistique de test

$\forall\ m \neq m'$,

$$T(N) = H^2\left(\hat{s}_m, \hat{s}_{m\vee m'}\right) - H^2\left(\hat{s}_{m'}, \hat{s}_{m\vee m'}\right) + 16\left[pen(m) - pen\left(m'\right)\right]$$

- *si* $T_{m,m'} < 0$ *préférer le modèle* m;
- *si* $T_{m,m'} > 0$ *préférer le modèle* m';
- *sinon préférer alatoirement le modèle* m *ou* m'.

Soit $\mathcal{R}_m$ *l'ensemble des* m *non préférés*

$$\mathcal{R}_m = \left\{m' \in \mathcal{M},\ m \neq m' |\ m \text{ n'est pas préféré d'après le test } T(N)\right\}$$

et, $\forall\ \epsilon > 0$ *(donné), nous définissons notre* $\hat{m} \in \mathcal{M}$ *par*

$$\mathcal{D}(\hat{m}) \leq \inf_{m\in\mathcal{M}} \mathcal{D}(m) + \epsilon/3 \quad \text{avec} \quad \mathcal{D}(m) = \sup_{m'\in\mathcal{R}_m} \left\{H^2\left(\hat{s}_m, \hat{s}_{m'}\right)\right\}.$$

Théorème

On suppose que les assertions H et H′ sont vraies.
Soit une fonction de pénalité $pen : \mathcal{M} \longrightarrow \mathbb{R}^+$ *telle que*

$$pen(m) \geq c\delta|m| + b\Delta_m \tag{10}$$

où $\{\Delta_m,\ m \in \mathcal{M}\}$ *est une famille de poids positifs satisfaisant l'équation (9). Si on choisit* $\hat{m} \in \mathcal{M}$ *telle que*

$$\mathcal{D}(\hat{m}) \leq \inf_{m \in \mathcal{M}} \mathcal{D}(m) + \epsilon/3 \quad avec\ \mathcal{D}(m) = \sup_{m' \in \mathcal{R}_m} \left\{ H^2(\hat{s}_m, \hat{s}_{m'}) \right\}$$

alors l'estimateur par histogramme pénalisé noté $\tilde{s} = \hat{s}_{\hat{m}}$ *vérifie*

$$\mathbb{E}\left[H^2(\tilde{s}, s)\right] \leq \left[390 \left(\inf_{m \in \mathcal{M}} \left(H^2(s, S_m) + pen(m) \right) + \frac{ab\Sigma^2}{2} \right) + \epsilon \right] \wedge 2. \tag{11}$$

Plan

Les motivations pour le choix d'une famille sont doubles.

Les motivations pour le choix d'une famille sont doubles.

1. $\mathcal{M}$ doit obligatoirement vérifier H.
   1. 1^{er} cas: si famille de partitions est totalement ordonnée pour l'inclusion, alors $\forall\ (m, m') \in \mathcal{M}^2$, on a $m \vee m' = m$ ou $m \vee m' = m'$ et H est satisfaite pour $\delta = 1$; et $\bar{\mathcal{M}} = \mathcal{M}$.
   2. $2^{ème}$ cas: comme $[0, 1]$ est un intervalle de $\mathbb{R}$, si $m \in \mathcal{M}$ est une partition finie de $[0, 1]$ alors l'assertion H est satisfaite avec $\delta = 1$.

Les motivations pour le choix d'une famille sont doubles.

1. $\mathcal{M}$ doit obligatoirement vérifier H.
   1. 1^{er} cas: si famille de partitions est totalement ordonnée pour l'inclusion, alors $\forall\ (m, m') \in \mathcal{M}^2$, on a $m \vee m' = m$ ou $m \vee m' = m'$ et H est satisfaite pour $\delta = 1$; et $\bar{\mathcal{M}} = \mathcal{M}$.
   2. $2^{ème}$ cas: comme $[0, 1]$ est un intervalle de $\mathbb{R}$, si $m \in \mathcal{M}$ est une partition finie de $[0, 1]$ alors l'assertion H est satisfaite avec $\delta = 1$.

2. La deuxième motivation est liée aux propriétés d'approximation des modèles. Par rapport à la croyance qu'on peut avoir sur la forme de la densité s, il est raisonnable de choisir une famille de modèles qui se rapproche le mieux de la forme de la fonction s.

On distinguera deux types de familles de partitions.

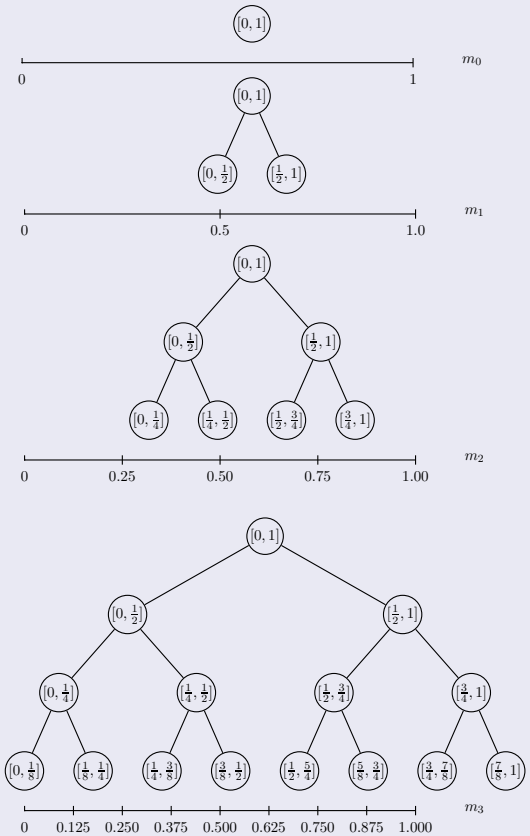
On distinguera deux types de familles de partitions.

Premier type de famille

- *Le premier type de famille est celui où le nombre de modèles de même dimension n'est pas très gros (i.e peu de modèles par dimension). Il est utilisé si la famille de partitions est régulière sur* $[0,1]$ *ou si les partitions sont formées de pavés sur* $[0,1]^k$ *construits à partir de partitions régulières sur chaque axe.*
- *La méthode consiste à partitionner de manière récursive et dyadique l'intervalle* $[0,1]$ *en partitions de plus en plus fines.*
- *Chaque modèle construit peut être représenté par un arbre binaire de hauteur* h.
- *Le nombre de modèles de même dimension est égal à* 1.

Premier type de famille

Famille de modèles associée à une famille d'arbres binaires



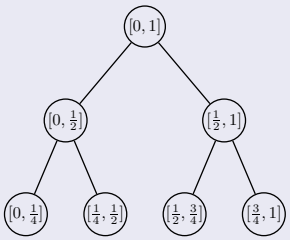
Deuxième type de famille

- *Le deuxième type de famille est celui où la famille de partitions est une grille fine régulière sur* $[0,1]$ *de pas fixé (dépendant de n en général). Les partitions de la famille sont alors des partitions dont les intervalles ont pour extrémités des points de la grille.*
- *Notre grille fine et régulière, de pas fixé* 2^{-h} *peut aussi être vue comme un arbre binaire de hauteur h.*
- *Pour ce type de famille, le nombre de modèles (resp. d'arbres) de même dimension (resp. de même hauteur) est exponentiel et borné par* $C_{2^h-1}^{|m|-1}$.

Deuxième type de famille

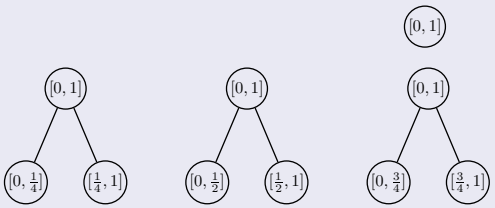
Famille de modèles associée à un arbre binaire de hauteur initiale h

Si notre arbre (ou grille) initial est

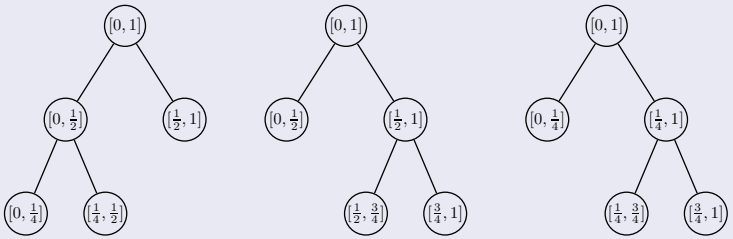


alors, comme sous-arbres ou modèles de la famille nous aurons

-



-



-

Chaque famille de modèles issue d'un arbre binaire de hauteur h(resp. grille de pas fixe égale à 2^{-h})

Plan

Définition

Si $\{V_m,\ m \in \mathcal{M}\}$ *est une famille de sous-espaces vectoriels de* $\mathbb{R}^n$ *à laquelle on associe les estimateurs par projection* $\left\{\hat{f}_m = \Pi_{V_m},\ m \in \mathcal{M}\right\}$, *on dira que* $\{V_m,\ m \in \mathcal{M}\}$ *admet* $(M, a) \in (\mathbb{R}_+ \times \mathbb{R}_+)$ *comme indice de complexité si*

$$Card\left\{m \in \mathcal{M} \mid D_m = D\right\} \leq Me^{aD} \quad \forall\ D \in \{1, .., n\}, \qquad (12)$$

où D_m *représente la dimension du sous-espace vectoriel* V_m.

Définition

Si $\{V_m,\ m \in \mathcal{M}\}$ *est une famille de sous-espaces vectoriels de* $\mathbb{R}^n$ *à laquelle on associe les estimateurs par projection* $\left\{\hat{f}_m = \Pi_{V_m},\ m \in \mathcal{M}\right\}$, *on dira que* $\{V_m,\ m \in \mathcal{M}\}$ *admet* $(M, a) \in (\mathbb{R}_+ \times \mathbb{R}_+)$ *comme indice de complexité si*

$$Card\left\{m \in \mathcal{M} \mid D_m = D\right\} \leq Me^{aD} \quad \forall\ D \in \{1, .., n\}, \qquad (12)$$

où D_m *représente la dimension du sous-espace vectoriel* V_m.

- Choisir des poids en fonction de la dimension des modèles

Premier type de famille

Remarques

- *La dimension des modèles croit avec la hauteur des arbres; ceci suggère de réécrire par exemple* Δ_m *sous la forme* $\Gamma_{\mathrm{m}} D_{\mathrm{m}}$.
- *on a un modèle par dimension* $\longrightarrow$ *famille de poids constants*

Comme $\{S_{\mathrm{m}},\ m \in \mathcal{M}\}$ *admet* $(1,0)$ *comme indice de complexité, on peut choisir* $\forall\ m \in \mathcal{M}$, $\Gamma_{\mathrm{m}} = \Gamma > 0$.

Premier type de famille

Remarques

- *La dimension des modèles croit avec la hauteur des arbres; ceci suggère de réécrire par exemple* Δ_m *sous la forme* $\Gamma_{\mathrm{m}} D_{\mathrm{m}}$.
- *on a un modèle par dimension* $\longrightarrow$ *famille de poids constants*

Comme $\{S_{\mathrm{m}},\ m \in \mathcal{M}\}$ *admet* $(1,0)$ *comme indice de complexité, on peut choisir* $\forall\ m \in \mathcal{M}$, $\Gamma_{\mathrm{m}} = \Gamma > 0$. *On a donc*

$$\sum_{\mathrm{m}\in\mathcal{M}} e^{-\Delta_m} = \sum_{\mathrm{m}\in\mathcal{M}} e^{-\Gamma_{\mathrm{m}} D_{\mathrm{m}}} \leq \sum_{\mathrm{D}\geq 1} M e^{-\Gamma D} \leq \frac{M}{1-e^{-\Gamma}}$$

Premier type de famille

- *Par conséquent, et conformément à l'équation* (10) *du théorème, notre pénalité devient pour* $D_m = |m|$,

$$\begin{aligned} pen(m) &= c\delta|m| + b\Gamma|m| \\ &= c|m| + b\Gamma|m| \quad car \quad \delta = 1 \\ &= |m|(c + b\Gamma) \\ &= \alpha|m|, \end{aligned}$$

$$avec\ \alpha = (c + b\Gamma) > 0.$$

Premier type de famille

- *Par conséquent, et conformément à l'équation* (10) *du théorème, notre pénalité devient pour* $D_m = |m|$,

$$\begin{aligned} pen(m) &= c\delta|m| + b\Gamma|m| \\ &= c|m| + b\Gamma|m| \quad car \quad \delta = 1 \\ &= |m|(c + b\Gamma) \\ &= \alpha|m|, \end{aligned}$$

$$avec\ \alpha = (c + b\Gamma) > 0.$$

- *Ainsi après renormalisation, notre théorème peut s'appliquer avec une pénalité de la forme*

$$pen(m) = \alpha\frac{|m|}{n} \quad avec\ \alpha > 0. \tag{13}$$

Deuxième type de famille

Pour ce type de famille, on va chercher à calibrer une pénalité de la forme

$$pen_\alpha(m) = \begin{cases} \alpha \dfrac{|m|}{n} & \text{si } m \text{ est régulière;} \\ \alpha \dfrac{|m|}{n} \left[1 + \dfrac{\log\left(|m| C_{2^h-1}^{|m|-1} \right)}{|m|} \right] & \text{si } m \text{ est irrégulière;} \end{cases} \tag{14}$$

avec α une constante positive à déterminer.
Ce choix est étroitement lié au fait qu'on aimerait borner la série de poids indépendamment de n.
Comme famille de poids associée à la forme de pénalité (14) *on a*

$$\Delta_m = \alpha \log\left(|m| C_{2^h-1}^{|m|-1} \right).$$

Plan

Pour calibrer au mieux une fonction de pénalité, on mesure sa qualité d'estimation. La qualité d'une procédure de sélection de modèles est définie par l'application

$$\alpha \longmapsto R_\alpha = \frac{\mathbb{E}\left[H^2\left(s, \tilde{s}_\alpha\right)\right]}{\inf_{\mathrm{m}} \mathbb{E}\left[H^2\left(s, \hat{s}_{\mathrm{m}}\right)\right]}.$$

Pour calibrer au mieux une fonction de pénalité, on mesure sa qualité d'estimation. La qualité d'une procédure de sélection de modèles est définie par l'application

$$\alpha \longmapsto R_\alpha = \frac{\mathbb{E}\left[H^2\left(s, \tilde{s}_\alpha\right)\right]}{\inf_{\mathrm{m}} \mathbb{E}\left[H^2\left(s, \hat{s}_{\mathrm{m}}\right)\right]}.$$

L'évaluation de ce ration se fait par la méthode de Monté Carlo. Ce qui nous permet de réécrire

$$R_\alpha = \frac{\sum_{\mathrm{j}=1}^{N} H^2(s, \hat{s}^{(j)}_{\hat{\mathrm{m}}(\mathrm{j},\alpha)})}{\inf_{\mathrm{m}} \sum_{\mathrm{j}=1}^{N} H^2(s, \hat{s}^{(j)}_{\mathrm{m}})}.$$

Pour calibrer au mieux une fonction de pénalité, on mesure sa qualité d'estimation. La qualité d'une procédure de sélection de modèles est définie par l'application

$$\alpha \longmapsto R_\alpha = \frac{\mathbb{E}\left[H^2\left(s, \tilde{s}_\alpha\right)\right]}{\inf_{\mathrm{m}} \mathbb{E}\left[H^2\left(s, \hat{s}_{\mathrm{m}}\right)\right]}.$$

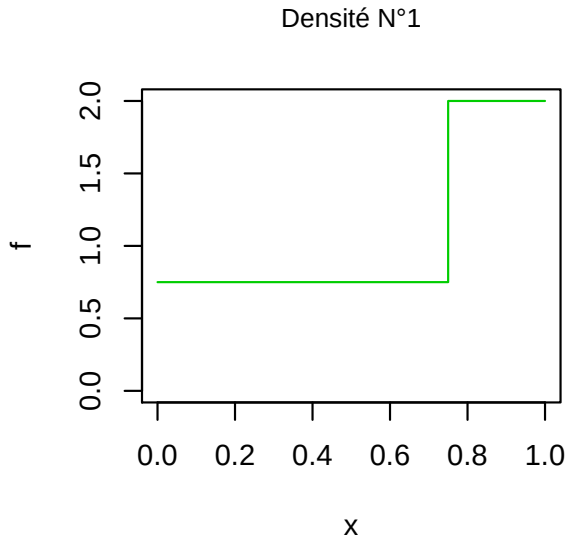
L'évaluation de ce ration se fait par la méthode de Monté Carlo. Ce qui nous permet de réécrire

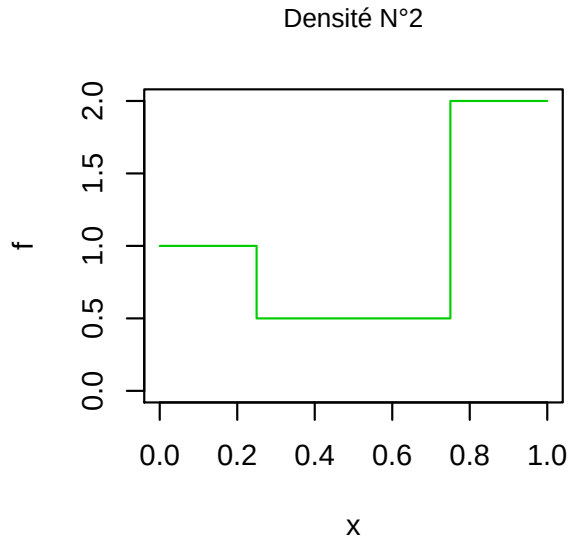
$$R_\alpha = \frac{\sum_{\mathrm{j}=1}^{N} H^2(s, \hat{s}^{(j)}_{\hat{\mathrm{m}}(\mathrm{j},\alpha)})}{\inf_{\mathrm{m}} \sum_{\mathrm{j}=1}^{N} H^2(s, \hat{s}^{(j)}_{\mathrm{m}})}.$$

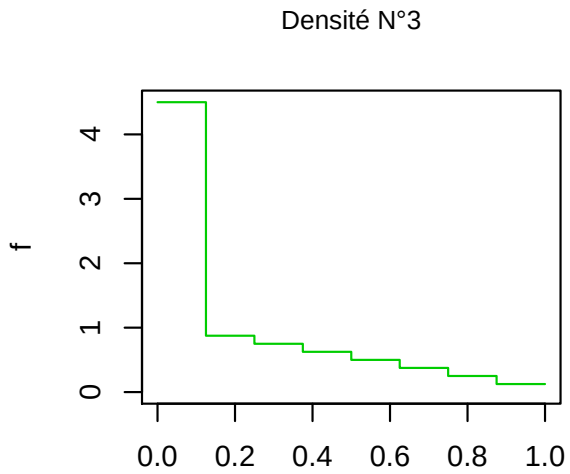
R_α dépend de nombreux facteurs:

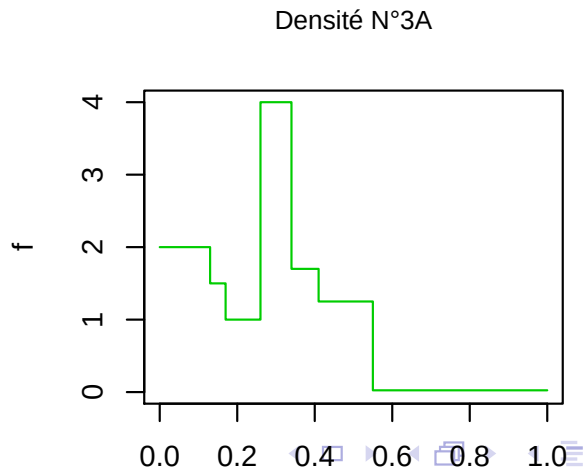
- $\tilde{s}$ est une fonction inconnue en pratique

Calibration de la fonction de pénalité

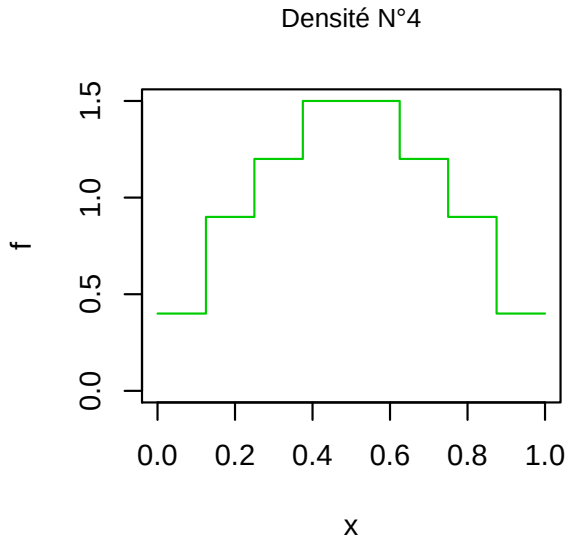


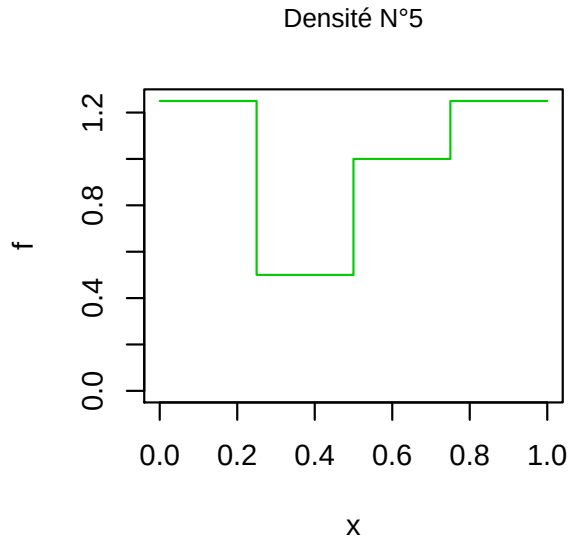


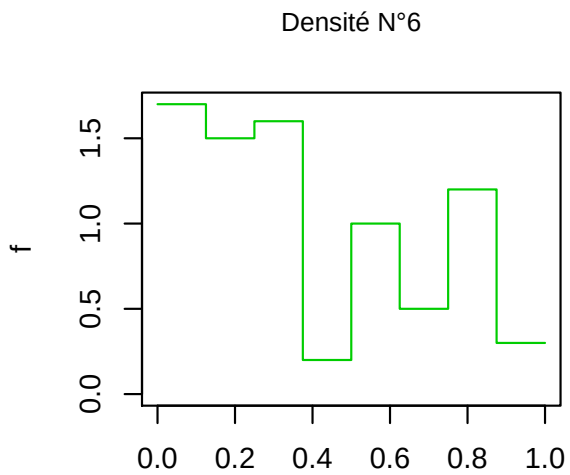


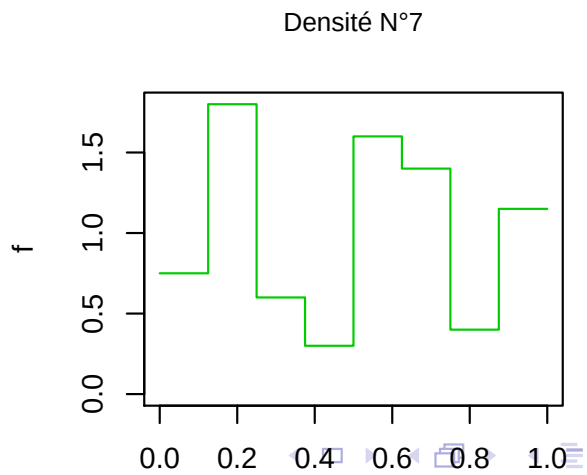


Calibration de la fonction de pénalité

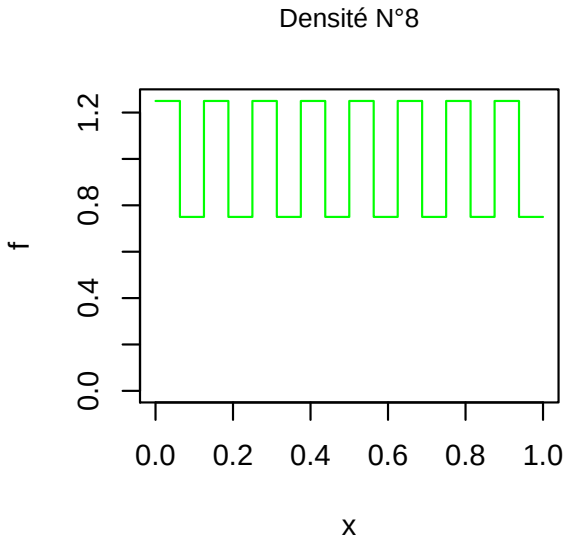


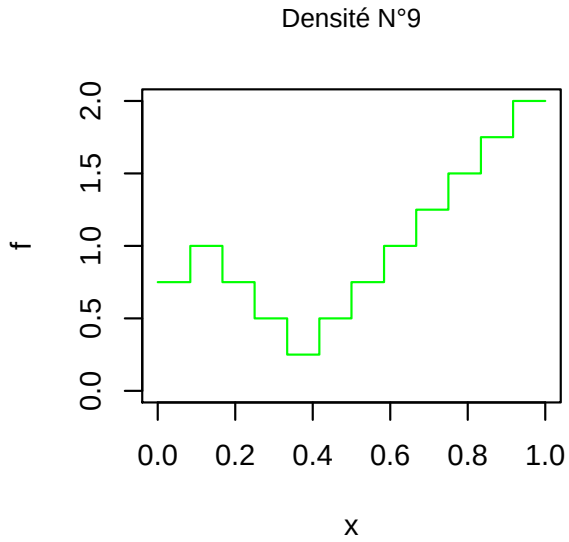




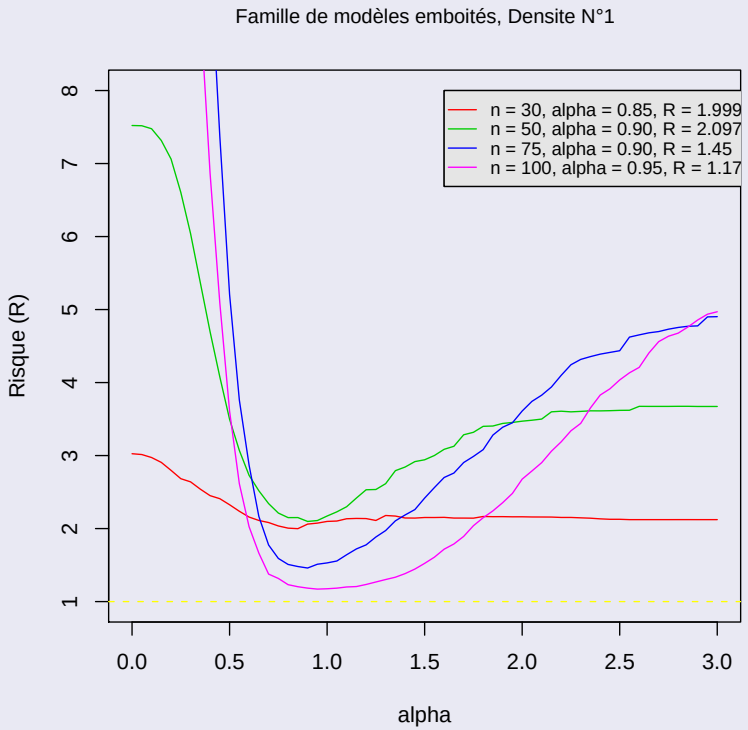


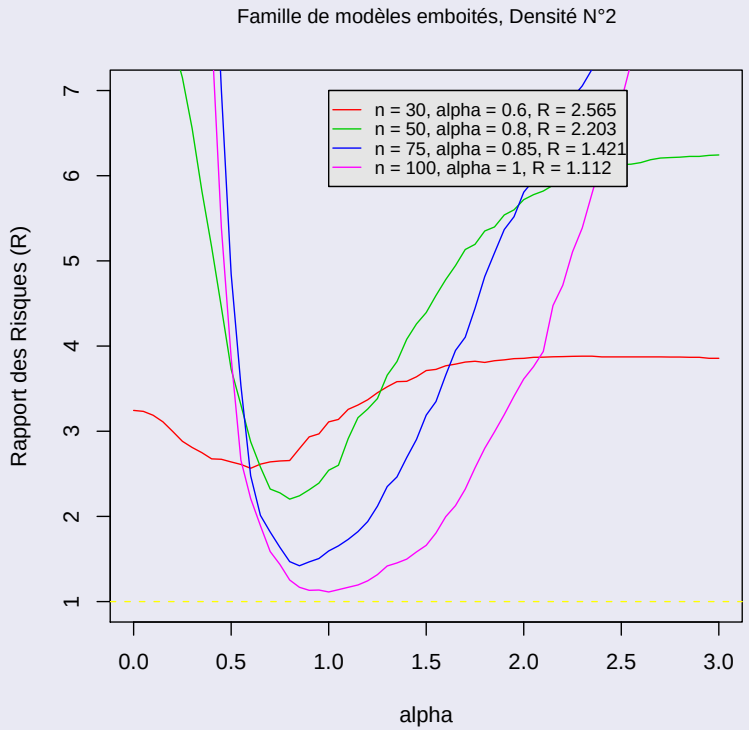
Calibration de la fonction de pénalité



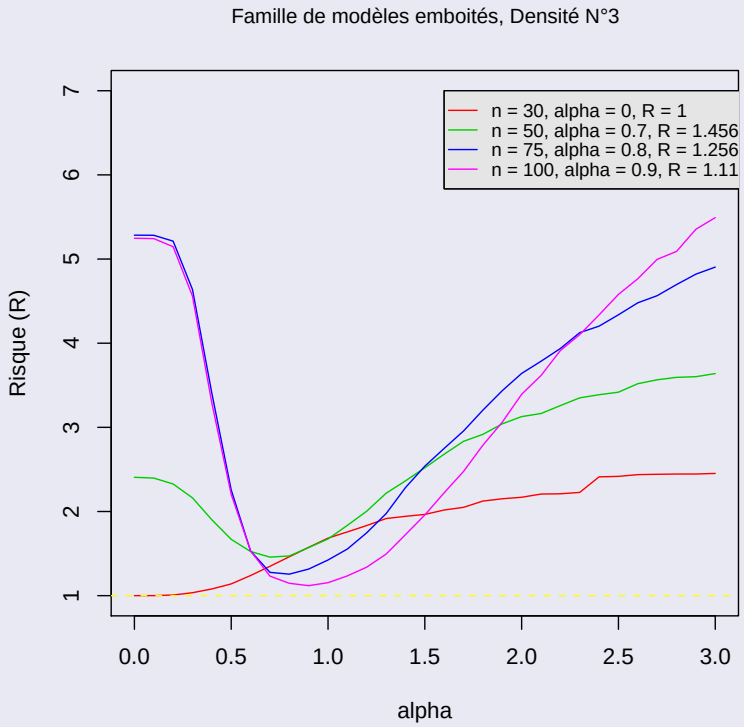


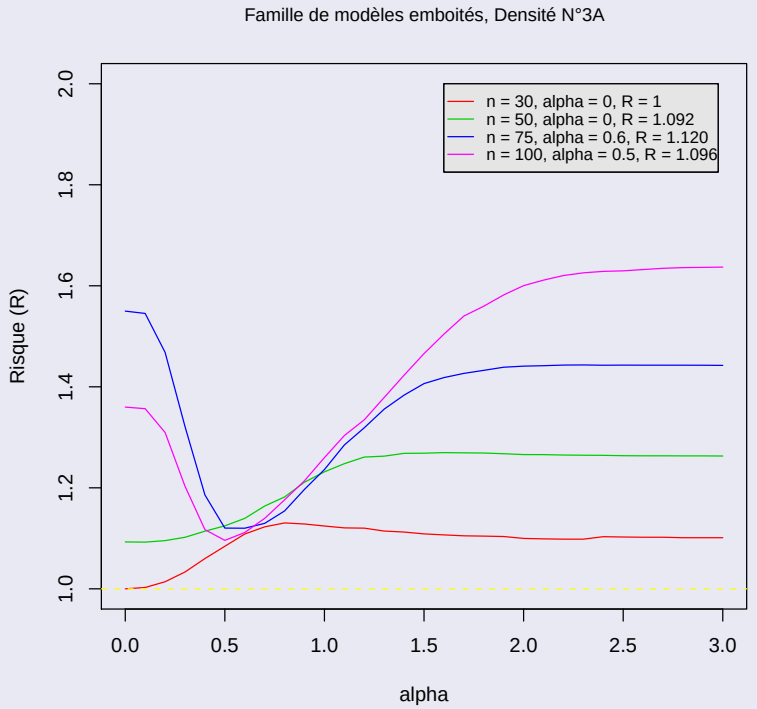
Premier type de famille



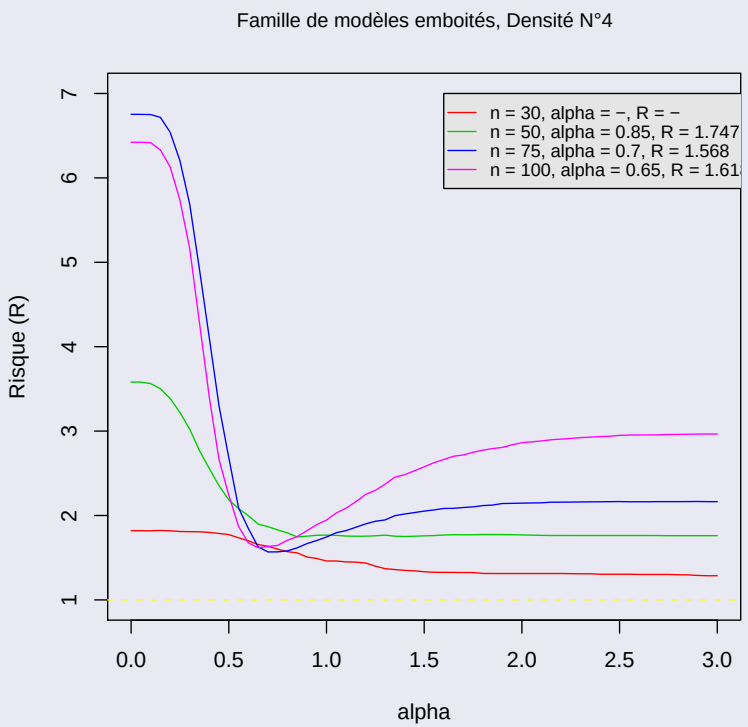


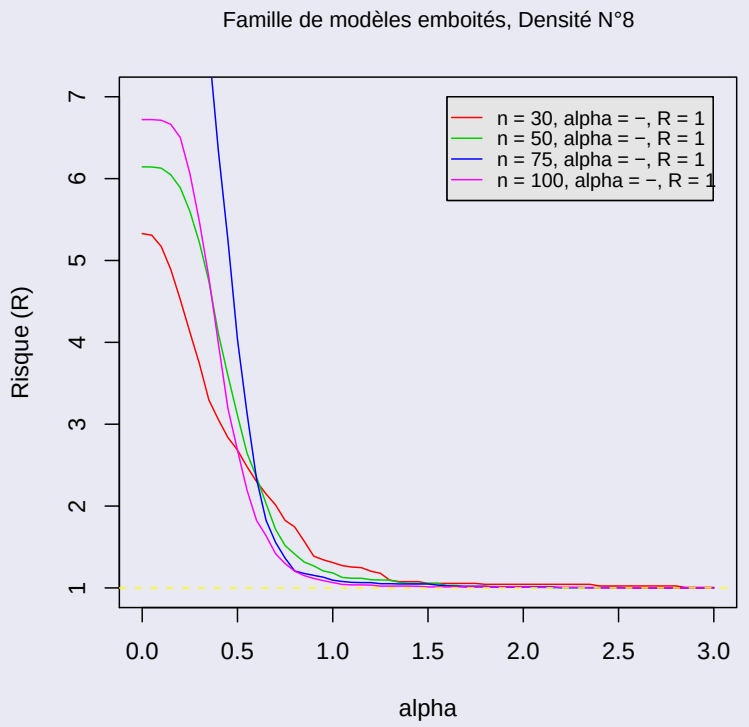
Premier type de famille





Premier type de famille





Premier type de famille

Remarque

Les simulations faites orientent notre choix vers une constante $\alpha > 1/2$. *Mieux encore, nos constatations sur la valeur moyenne de* α *montre qu'une constante* α *autour de* $3/4$ *contribuerait certainement au choix d'un bon estimateur.*

Premier type de famille, Résultats obtenus pour $\alpha = 3/4$

Densité N°1, D = 2, -- non régulière

	1	2	4	8	16	32	64	128
30	20%	26%	44%	10%	0%	0%	0%	0%
50	6%	13%	73%	6%	2%	0%	0%	0%
75	0.4%	4.5%	91.3%	3.1%	0.7%	0%	0%	0%
100	0%	0.4%	96.2%	1.7%	1.6%	0.1%	0%	0%

Densité N°2, D = 3, -- non régulière

	1	2	4	8	16	32	64	128
30	27%	7%	56%	10%	0%	0%	0%	0%
50	9.9%	1.7%	79.1%	7.6%	1.7%	0%	0%	0%
75	0.5%	0.5%	93.9%	4.2%	0.9%	0%	0%	0%
100	0.2%	0%	96.6%	2.5%	0.7%	0%	0%	0%

Densité N°3, D = 8

	1	2	4	8	16	32	64	128
30	0.1%	13.9%	33.3%	52.7%	0%	0%	0%	0%
50	0%	2.1%	17.3%	75.2%	5.4%	0%	0%	0%
75	0%	0%	4.7%	90.9%	0.2%	0%	0%	0%
100	0%	0%	1%	94.8%	4%	0.2%	0%	0%

Densité N°3A, D = 7, -- non régulière

	1	2	4	8	16	32	64	128
30	0%	69.9%	17.7%	19.4%	0%	0%	0%	0%
50	0%	34.5%	18.9%	43.5%	3.1%	0%	0%	0%
75	0%	13.4%	10.1%	67.1%	9.4%	0%	0%	0%
100	0%	5.4%	6.1%	66.7%	21.7%	0.1%	0%	0%

Densité N°8, D = 16

	1	2	4	8	16	32	64	128
30	86.1%	7.7%	2.3%	3.9%	0%	0%	0%	0%
50	89.5%	6.4%	0.9%	0.5%	2.7%	0%	0%	0%
75	87.5%	8.4%	1.7%	0.2%	1.8%	0.4%	0%	0%
100	89%	6.5%	1.4%	0.3%	2.5%	0.3%	0%	0%

Densité N°9, D = 12, -- non régulière

	1	2	4	8	16	32	64	128
30	21.1%	47.4%	22.2%	5.3%	0%	0%	0%	0%
50	11.5%	41.3%	40.2%	5.3%	1.7%	0%	0%	0%
75	1.4%	38.2%	53.1%	5.5%	1.5%	0.2%	0%	0%
100	0.2%	25.8%	68.2%	4.3%	1.3%	0.3%	0%	0%

Partition **Oracle**.

Partition la plus préférée en % quand **pen(m) = 0.75*|m|/n.**

Intersection des couleurs et .

Premier type de famille

- *pour des densités suffisamment régulières, on n'arrive à faire aussi mieux que l'oracle;*
- *la probabilité de faire presqu'aussi bien que l'oracle augmente avec la taille de l'échantillon quelque soit la densité;*
- *bien que la densité N°4 soit régulière, la chance de faire aussi bien que l'oracle reste relativement faible même pour un échantillon de taille* $n = 100$*;*
- *la partition oracle trouvée pour la densité N°8 n'est pas la bonne. Il faudrait prendre un échantillon de taille plus grande.*

Premier type de famille, Rapport des Risques pour $\alpha = \{3/4, 1\}$

TABLEAUX PARTIE I

Densité N°1, D = 2, -- non régulière

	30	50	75	100
3/4	2.085528	1.998098	1.39379	1.220069
1	2.166660	2.101907	1.409275	1.170014

Densité N°2, D = 3, -- non régulière

	30	50	75	100
3/4	2.532256	2.285677	1.445892	1.315855
1	2.939251	2.599307	1.446452	1.096044

Densité N°3, D = 8

	30	50	75	100
3/4	1.411709	1.481943	1.24236	1.107728
1	1.715156	1.765460	1.456735	1.164081

Densité N°3A, D = 7, -- non régulière

	30	50	75	100
3/4	1.122839	1.182148	1.143967	1.14893
1	1.115612	1.223856	1.229254	1.243921

Densité N°4, D = 8

	30	50	75	100
3/4	1.638404	1.735403	1.621475	1.671086
1	1.491712	1.767753	1.695997	2.03235

Densité N°8, D = 16

	30	50	75	100
3/4	1.762976	1.547181	1.325269	1.180567
1	1.234462	1.187753	1.099913	1.063308

Rapport des Risques évalué pour 1000 échantillons de tailles respectives 30, 50, 75 et 100.

Premier type de famille

1. *Si* $\alpha = 3/4$, *alors la qualité d'estimation décroît sensiblement et tend presque vers* 1 *lorsque la taille de l'échantillon augmente. Ce qui traduit une amélioration notable de la qualité de sélection des modèles.*
2. *Si* $\alpha = 1$,
   - *la qualité d'estimation tend à augmenter avec la taille de l'échantillon pour la densité N°*4*. Ce qui traduit une mauvaise estimation;*
   - *la qualité d'estimation des densités N°*2 *et N°*8 *est très intéressante. On a un rapport des risques de l'ordre de* 1.097 *pour la densité N°*2 *et* 1.063 *pour la densité N°*8*;*
   - *pour les autres densités, on a un résultat identique à* $\alpha = 3/4$*;*

Premier type de famille

Cependant, l'étude de la courbe de risque montre qu'il faut éviter des valeurs très petites ou très grandes de α. En effet

- *lorsque $\alpha < 0.5$, on aura toujours tendance à sélectionner le modèle de plus grande dimension. Ce qui est en général un mauvais choix;*
- *lorsque $\alpha > 1$, bien que la remontée de la courbe de risque soit moins prononcée, on est tout de même conduit à faire de mauvais choix. On a tendance à choisir le modèle de plus petite dimension.*

Premier type de famille, Rapport des Risques pour $\alpha = \{1/4, 3/2\}$

TABLEAUX PARTIE II

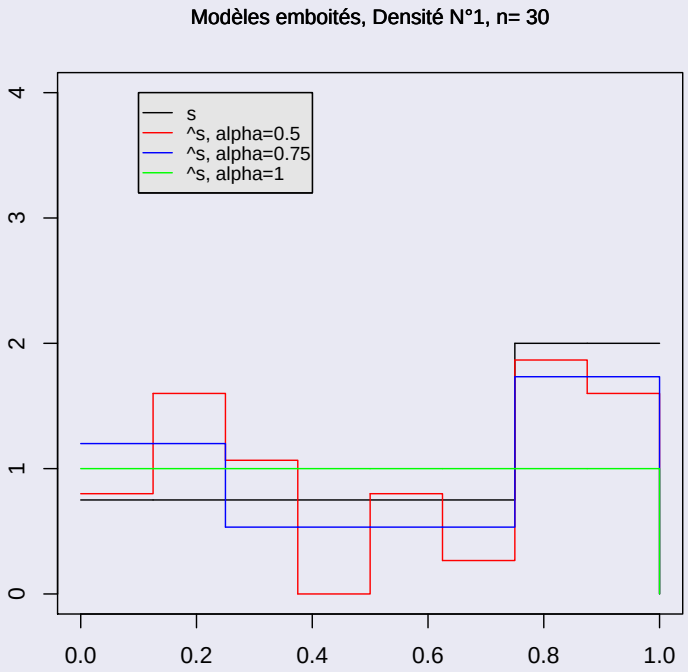
DENSITE N°1

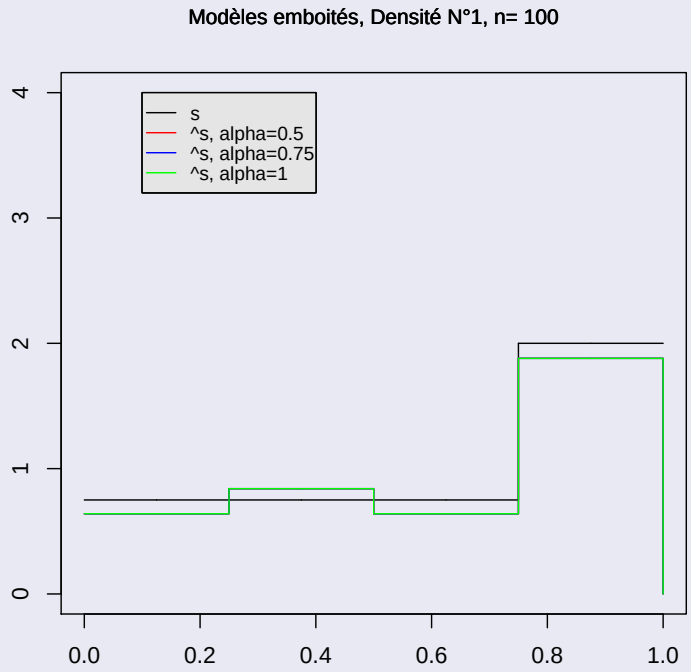
	30	50	75	100
3/4	2.085528	1.998098	1.39379	1.220069
1	2.166660	2.101907	1.409275	1.170014

	30	50	75	100
3/4	2.467966	3.322419	4.919703	3.618984
1	2.108257	2.282606	2.026623	1.477418

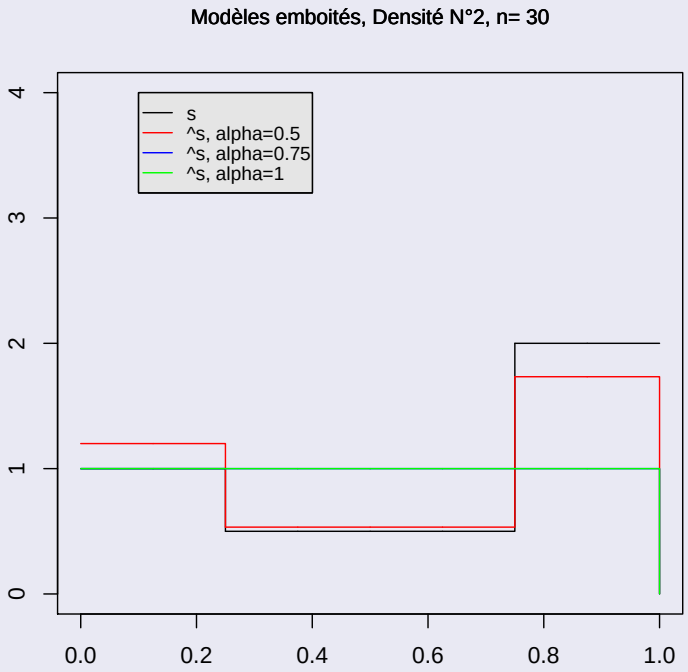
Rapport des Risques évalué pour 1000 échantillons de tailles respectives 30, 50, 75 et 100.

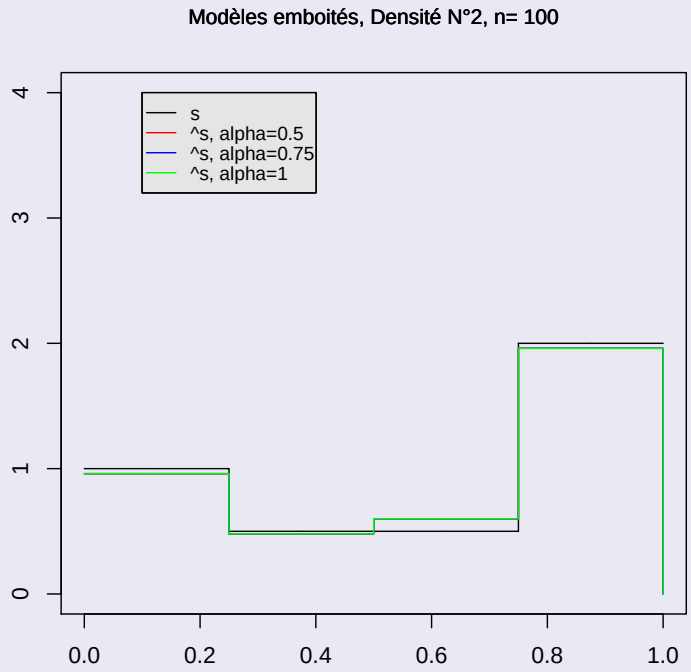
Premier type de famille



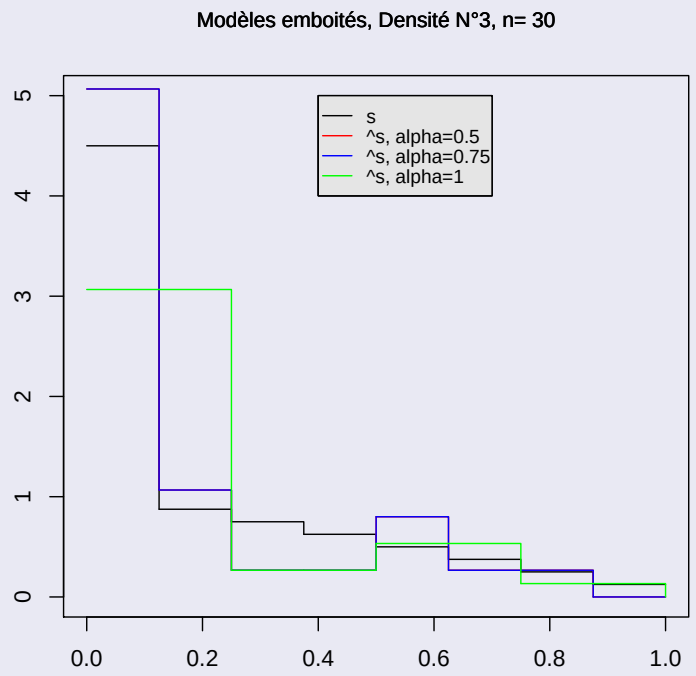


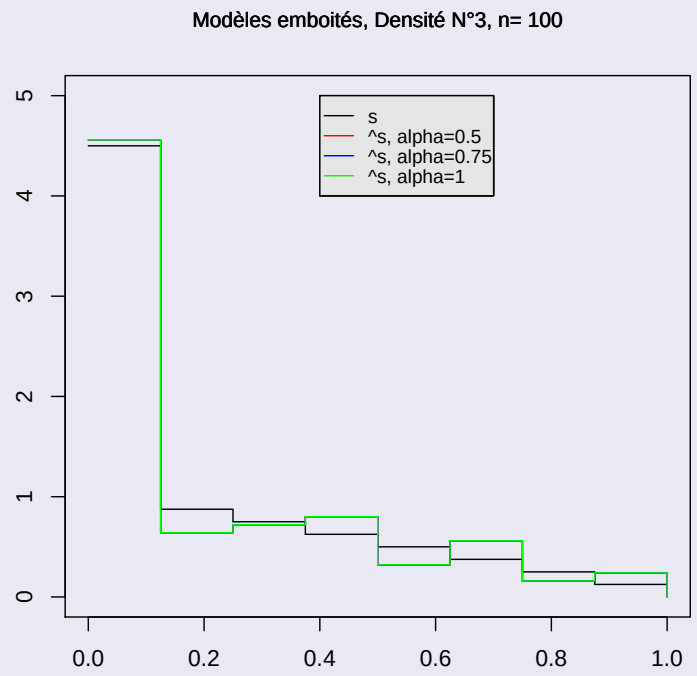
Premier type de famille



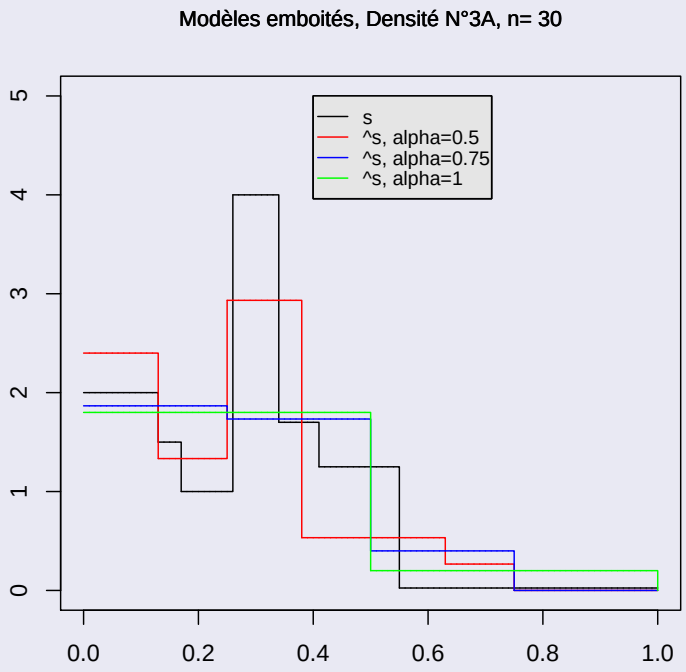


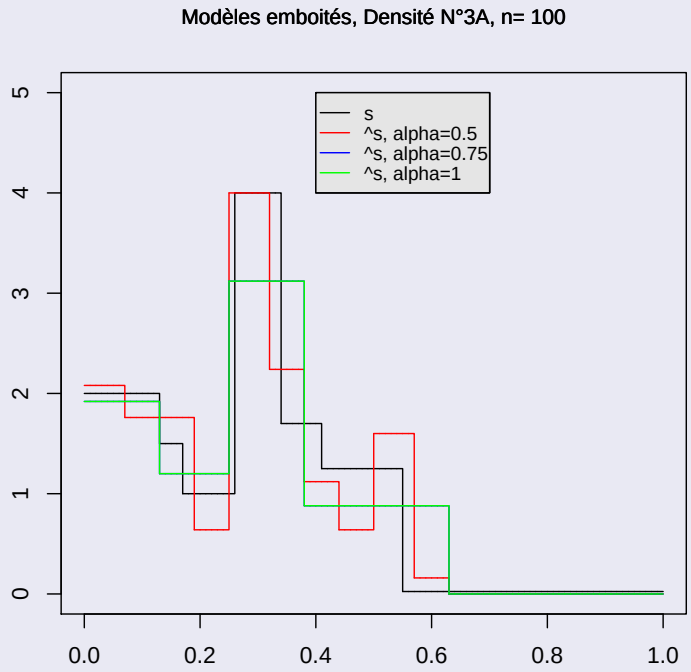
Premier type de famille



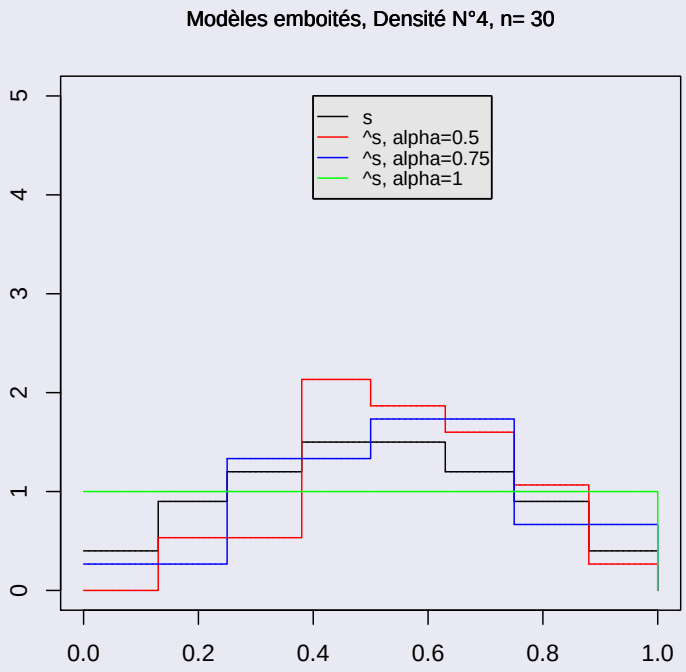


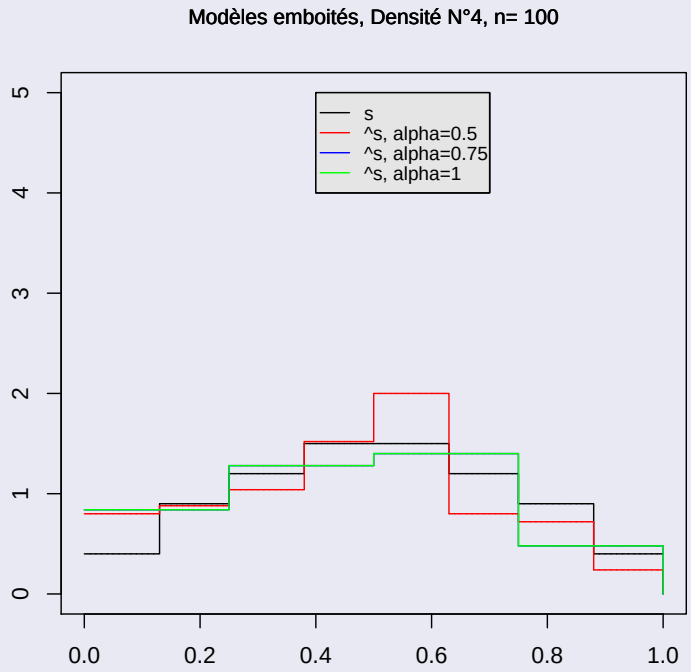
Premier type de famille



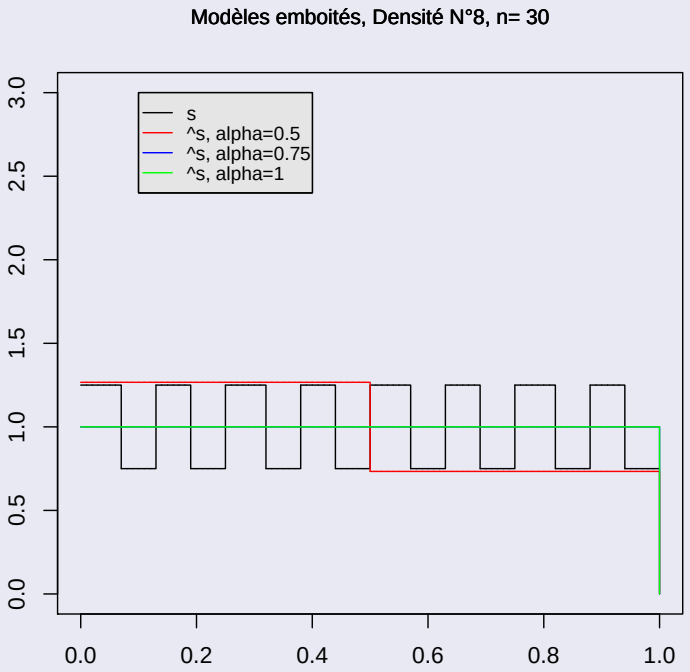


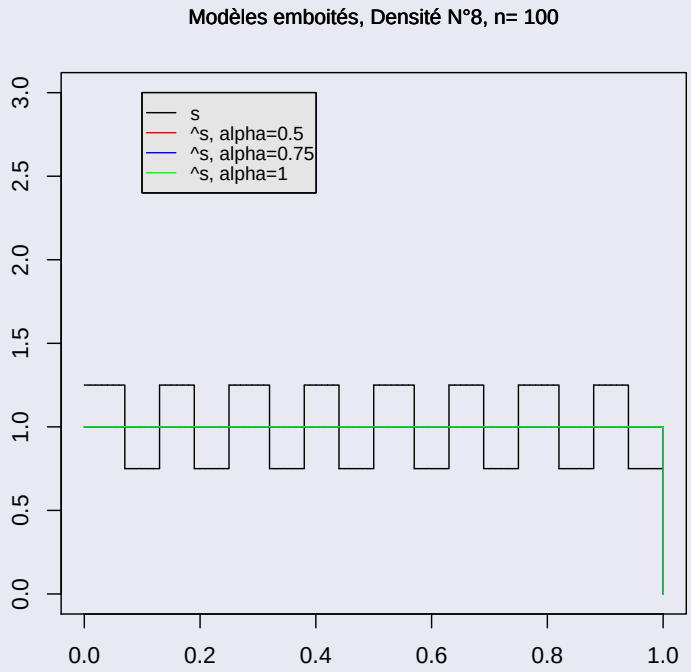
Premier type de famille





Premier type de famille





Premier type de famille

- *On constate qu'une constante* $\alpha \approx 3/4$ *semble bien marcher.*

Premier type de famille

- *On constate qu'une constante* $\alpha \approx 3/4$ *semble bien marcher.*
- *Une pénalité autour de* $\frac{3}{4}\frac{|m|}{n}$ *fonctionne très bien, ou tout au moins, contribue à sélectionner un estimateur pas trop mauvais, dès lors que nous disposons de suffisamment de données* ($n \geq 75$).

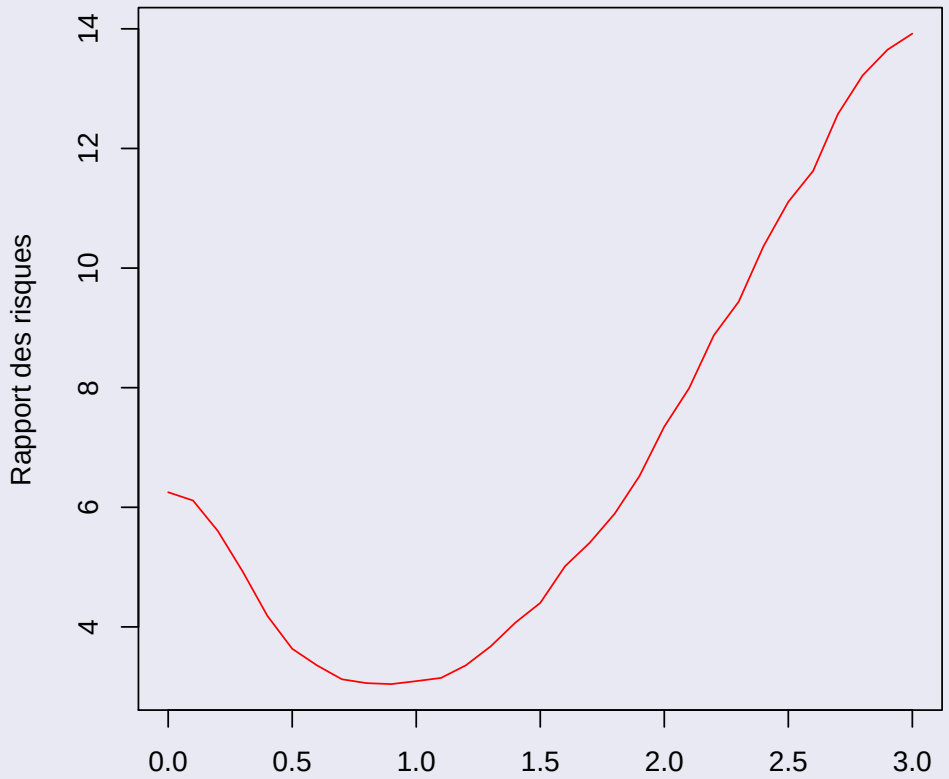
Premier type de famille

- *On constate qu'une constante* $\alpha \approx 3/4$ *semble bien marcher.*
- *Une pénalité autour de* $\frac{3}{4}\frac{|m|}{n}$ *fonctionne très bien, ou tout au moins, contribue à sélectionner un estimateur pas trop mauvais, dès lors que nous disposons de suffisamment de données* ($n \geq 75$).
- *Pour ce type de famille, nous conseillons une pénalité de la forme*

$$pen(m) = \frac{3}{4}\frac{|m|}{n}$$

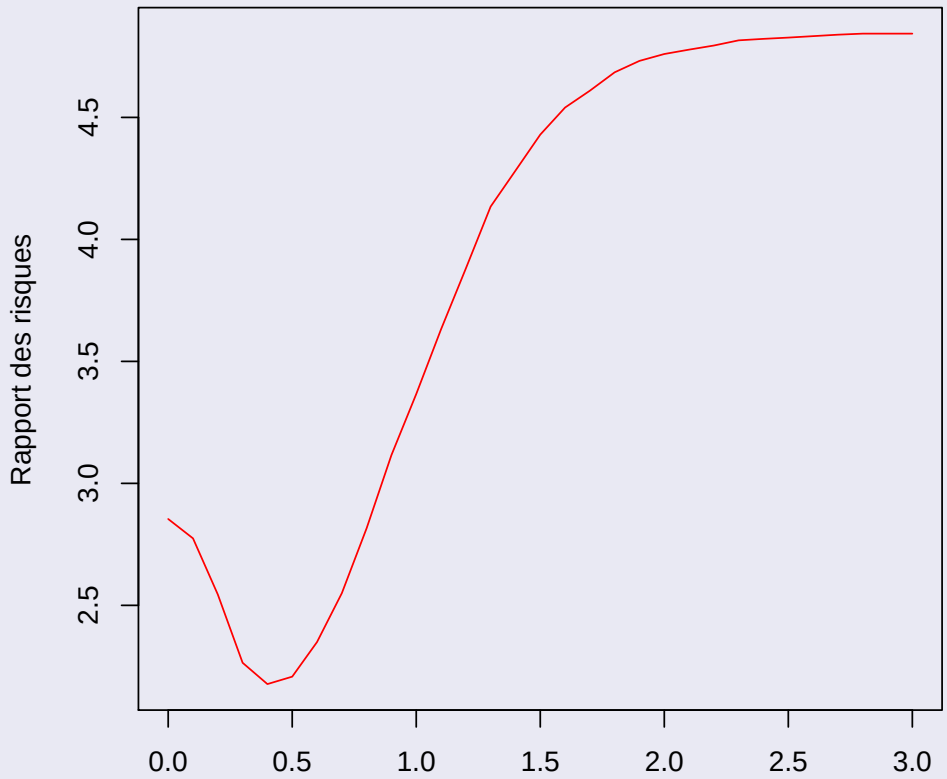
Deuxième type de famille

Figure: Densité N^o1. $n = 100$, $\alpha = 0.9$



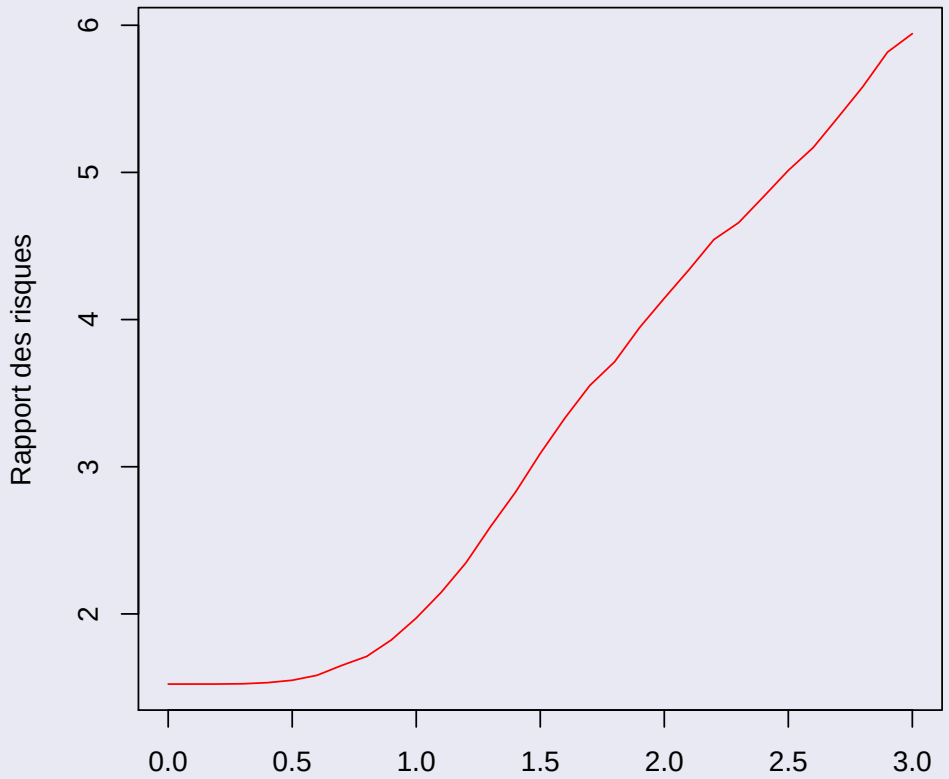
Deuxième type de famille

Figure: Densité N^o5. $n = 100$, $\alpha = 0.4$



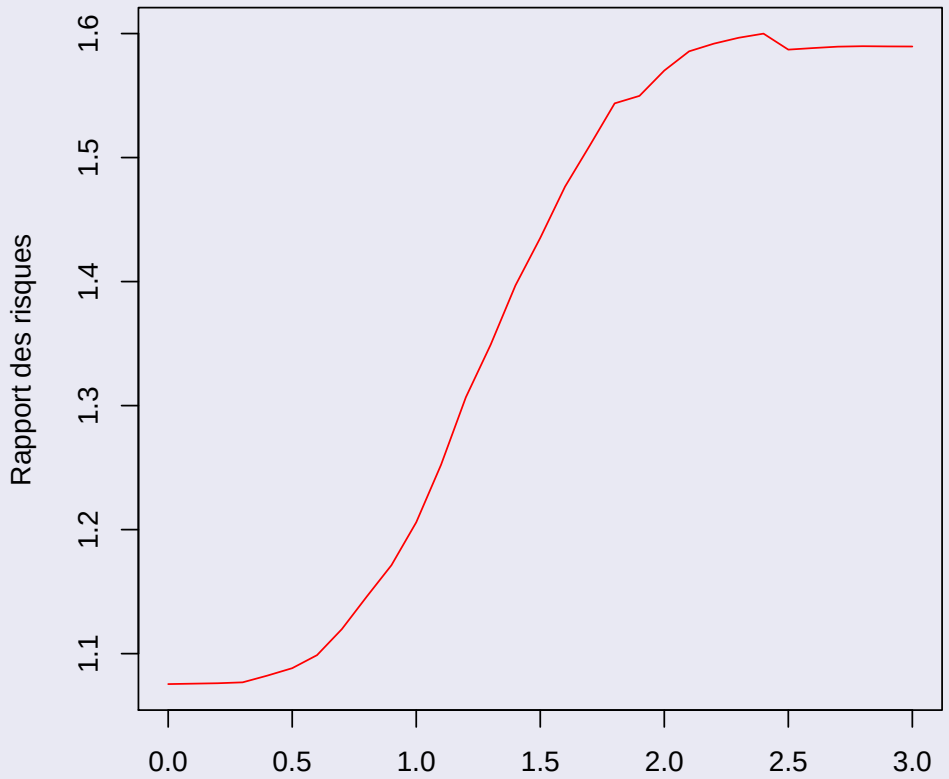
Deuxième type de famille

Figure: Densité $N^o 3$. $n = 100$, $\alpha = 0$



Deuxième type de famille

Figure: Densité $N^o 3A$. $n = 100$, $\alpha = 0$



Deuxième type de famille

INSUFFISANT POUR CONCLURE !!!!

- Le résultat des simulations obtenu pour le premier type de famille d'histogrammes s'est averé très encourageant. On remarque qu'une pénalité autour de $pen(m) = \frac{3}{4}\frac{|m|}{n}$ semble fonctionner très bien pour cette famille.

- Le résultat des simulations obtenu pour le premier type de famille d'histogrammes s'est averé très encourageant. On remarque qu'une pénalité autour de $pen(m) = \frac{3}{4}\frac{|m|}{n}$ semble fonctionner très bien pour cette famille.
- Toutefois, il serait bien d'affiner la fonction de pénalité en rajoutant par exemple un terme correctif afin de stabiliser les bornes de risques pour toutes les valeurs de n.

- Le résultat des simulations obtenu pour le premier type de famille d'histogrammes s'est averé très encourageant. On remarque qu'une pénalité autour de $pen(m) = \frac{3}{4}\frac{|m|}{n}$ semble fonctionner très bien pour cette famille.
- Toutefois, il serait bien d'affiner la fonction de pénalité en rajoutant par exemple un terme correctif afin de stabiliser les bornes de risques pour toutes les valeurs de n.
- Il faudrait faire une étude plus approfondie afin de déterminer notamment la loi asymptotique et la vitesse de convergence de nos estimateurs.

- Le résultat des simulations obtenu pour le premier type de famille d'histogrammes s'est averé très encourageant. On remarque qu'une pénalité autour de $pen(m) = \frac{3}{4}\frac{|m|}{n}$ semble fonctionner très bien pour cette famille.
- Toutefois, il serait bien d'affiner la fonction de pénalité en rajoutant par exemple un terme correctif afin de stabiliser les bornes de risques pour toutes les valeurs de n.
- Il faudrait faire une étude plus approfondie afin de déterminer notamment la loi asymptotique et la vitesse de convergence de nos estimateurs.
- Il serait en outre très intéressant de regarder et d'étudier le comportement de nos estimateurs sur des données réelles provenant de l'archéologie, l'agriculture et pourquoi pas l'imagerie.

- Les résultats ne sont pas assez concluants pour la famille d'histogrammes à pas irrégulier.

- Les résultats ne sont pas assez concluants pour la famille d'histogrammes à pas irrégulier.
- La grille initiale n'était pas suffisamment fine et la procédure ayant conduit à nos résultats est assez lourde.

- Les résultats ne sont pas assez concluants pour la famille d'histogrammes à pas irrégulier.
- La grille initiale n'était pas suffisamment fine et la procédure ayant conduit à nos résultats est assez lourde.
- L'idéal serait de ré-implémenter la procédure en C ou C++ pour leur robustesse.

Bibliographie

- *Y. BARAUD and L. BIRGÉ (2005). Histogram Type Estimators Based on Non-Negative Random Variables. Prépublication no 715*
- *Y. BARAUD and L. BIRGÉ (2006). Estimating the Intensity of a Randon Measure by Histogram Type Estimators. Prépublication no ?*
- *L. BIRGÉ and Y. ROZENHOLC (2002). How many bins should be put in a regular histogram. Prépublication no 721*
- *G. CASTELLAN (2000). Sélection d'histogrammes à l'aide d'un critère de type Akaike.C.R.A.S. 330, 729-732*

Bibliographie

- *G. CASTELLAN (1999). Modified Akaike's criterion for histogram density estimation.Technical Report 99.61. Université Paris-Sud, Orsay*
- *P. Deheuvels (1977).Estimation non paramétrique de la densité par histogrammes généralisés. Revue de Statistique Appliquée, 25 no. 3 (1977), p. 5-42*
- *C. TULEAU.(2005). Sélection de Variables pour la Discrimination en Grande Dimension et Classification de Données Fonctionnelles. Thèse Université Paris-Sud, Orsay*

♣QUESTIONS ????