Vers une nouvelle validation croisée V-fold

Nelo Magalhães
(joint work with L. Birgé and P. Massart)

Journées MAS
Toulouse, 29 août 2014

Introduction

- $\mathbb{X} = \{X_1, \ldots, X_n\}$ où $X_i \in \mathbb{R}$ i.i.d. $\sim P$, de densité s.

Introduction

- $\mathbb{X} = \{X_1, \ldots, X_n\}$ où $X_i \in \mathbb{R}$ i.i.d. $\sim P$, de densité s.
- **But :** proposer un estimateur $\tilde{s} = \tilde{s}(\mathbb{X})$ de s.

Introduction

- $\mathbb{X} = \{X_1, \ldots, X_n\}$ où $X_i \in \mathbb{R}$ i.i.d. $\sim P$, de densité s.
- **But :** proposer un estimateur $\tilde{s} = \tilde{s}(\mathbb{X})$ de s.
-

possibilité :		
estimateurs à noyau		
histogrammes		
estimateurs par projection		
ondelettes		

Introduction

- $\mathbb{X} = \{X_1, \ldots, X_n\}$ où $X_i \in \mathbb{R}$ i.i.d. $\sim P$, de densité s.
- **But :** proposer un estimateur $\tilde{s} = \tilde{s}(\mathbb{X})$ de s.
-

possibilité :		**problème :**
estimateurs à noyau		sélection de la fenêtre
histogrammes	$\Longrightarrow$	sélection de partition
estimateurs par projection		sélection de modèle
ondelettes		choix du seuil

Introduction

- $\mathbb{X} = \{X_1, \dots, X_n\}$ où $X_i \in \mathbb{R}$ i.i.d. $\sim P$, de densité s.
- **But :** proposer un estimateur $\tilde{s} = \tilde{s}(\mathbb{X})$ de s.
-

possibilité :		**problème :**
estimateurs à noyau		sélection de la fenêtre
histogrammes	$\Longrightarrow$	sélection de partition
estimateurs par projection		sélection de modèle
ondelettes		choix du seuil

- $\{$noyaux, histogrammes, ondelettes, $\dots\}$
  $\Longrightarrow$ problème du choix d'une **méthode d'estimation**
- **méthode d'estimation** $\mathcal{A}_m : \mathbb{Y} \subseteq \mathbb{X} \mapsto \hat{s}_m(\mathbb{Y})$ estimateur de s

Cadre

$\mathbb{X} = \{X_1, \ldots, X_n\}$ où $X_i \in \mathbb{R}$ i.i.d. $\sim P$, de densité s,
$(\mathcal{A}_m)_{m \in \mathcal{M}}$ collection de méthodes d'estimation.

Cadre

$\mathbb{X} = \{X_1, \ldots, X_n\}$ où $X_i \in \mathbb{R}$ i.i.d. $\sim P$, de densité s,
$(\mathcal{A}_m)_{m \in \mathcal{M}}$ collection de méthodes d'estimation.

- **Risque de la méthode $\mathcal{A}_m$:** $\mathbb{E}[\ell(s, \hat{s}_m)]$, où $\hat{s}_m = \mathcal{A}_m(\mathbb{X})$,
  ℓ est une fonction de perte :
  Cas classique : $\ell(s,t) = P(\gamma(t,X) - \gamma(s,X)) \geq 0 \ \forall t$, avec
  1. $\gamma(t,x) = \|t\|^2 - 2t(x) \Longrightarrow \ell(s,t) = \|t - s\|^2$
  2. $\gamma(t,x) = -\log(t(x)) \Longrightarrow \ell(s,t) = K(s,t)$

  Ou : $\ell(s,t) = h^2(s,t) = 1/2 \int (\sqrt{s} - \sqrt{t})^2$

Cadre

$\mathbb{X} = \{X_1, \ldots, X_n\}$ où $X_i \in \mathbb{R}$ i.i.d. $\sim P$, de densité s,
$(\mathcal{A}_m)_{m \in \mathcal{M}}$ collection de méthodes d'estimation.

- **Risque de la méthode $\mathcal{A}_m$:** $\mathbb{E}[\ell(s, \hat{s}_m)]$, où $\hat{s}_m = \mathcal{A}_m(\mathbb{X})$, ℓ est une fonction de perte :
  Cas classique : $\ell(s,t) = P(\gamma(t,X) - \gamma(s,X)) \geq 0 \ \forall t$, avec
  1. $\gamma(t,x) = \|t\|^2 - 2t(x) \Longrightarrow \ell(s,t) = \|t-s\|^2$
  2. $\gamma(t,x) = -\log(t(x)) \Longrightarrow \ell(s,t) = K(s,t)$

  Ou : $\ell(s,t) = h^2(s,t) = 1/2 \int (\sqrt{s} - \sqrt{t})^2$
- **But :** proposer $\hat{m} = \hat{m}(\mathbb{X})$ t.q. $\tilde{s} = \mathcal{A}_{\hat{m}}(\mathbb{X})$ vérifie

$$\mathbb{E}[\ell(s,\tilde{s})] \sim \inf_{m \in \mathcal{M}} \mathbb{E}[\ell(s, \hat{s}_m)]$$

Table des Matières

Validation simple

- Évaluer la qualité de chaque méthode

Validation simple

- Évaluer la qualité de chaque méthode
- Problème : utiliser les mêmes données pour entraîner les méthodes et pour évaluer la qualité des estimateurs !

Validation simple

- Évaluer la qualité de chaque méthode
- Problème : utiliser les mêmes données pour entraîner les méthodes et pour évaluer la qualité des estimateurs !
- **Idée :** $\mathbb{X} = \mathbb{X}^t \bigsqcup \mathbb{X}^v$

Validation simple

- Évaluer la qualité de chaque méthode
- Problème : utiliser les mêmes données pour entraîner les méthodes et pour évaluer la qualité des estimateurs !
- **Idée :** $\mathbb{X} = \mathbb{X}^t \bigsqcup \mathbb{X}^v$
  1. $\mathbb{X}^t \Longrightarrow (\hat{s}_m^t = \mathcal{A}_m(\mathbb{X}^t))_{m \in \mathcal{M}}$

Validation simple

- Évaluer la qualité de chaque méthode
- Problème : utiliser les mêmes données pour entraîner les méthodes et pour évaluer la qualité des estimateurs !
- **Idée :** $\mathbb{X} = \mathbb{X}^t \bigsqcup \mathbb{X}^v$
  1. $\mathbb{X}^t \Longrightarrow (\hat{s}_m^t = \mathcal{A}_m(\mathbb{X}^t))_{m \in \mathcal{M}}$
  2. $\mathbb{X}^v \Longrightarrow$ définir $crit_{HO}(m)$ qui évalue la qualité de m

Validation simple

- Évaluer la qualité de chaque méthode
- Problème : utiliser les mêmes données pour entraîner les méthodes et pour évaluer la qualité des estimateurs !
- **Idée :** $\mathbb{X} = \mathbb{X}^t \bigsqcup \mathbb{X}^v$
  1. $\mathbb{X}^t \Longrightarrow (\hat{s}_m^t = \mathcal{A}_m(\mathbb{X}^t))_{m \in \mathcal{M}}$
  2. $\mathbb{X}^v \Longrightarrow$ définir $crit_{HO}(m)$ qui évalue la qualité de m

$$\hat{m} \in \arg\min_{m \in \mathcal{M}} crit_{HO}(m)$$

Validation simple

- Évaluer la qualité de chaque méthode
- Problème : utiliser les mêmes données pour entraîner les méthodes et pour évaluer la qualité des estimateurs !
- **Idée :** $\mathbb{X} = \mathbb{X}^t \bigsqcup \mathbb{X}^v$
  1. $\mathbb{X}^t \Longrightarrow (\hat{s}_m^t = \mathcal{A}_m(\mathbb{X}^t))_{m \in \mathcal{M}}$
  2. $\mathbb{X}^v \Longrightarrow$ définir $crit_{HO}(m)$ qui évalue la qualité de m

$$\hat{m} \in \arg\min_{m \in \mathcal{M}} crit_{HO}(m)$$

- **Cas classique :** $\ell(s, \hat{s}_m) = P(\gamma(\hat{s}_m, X) - \gamma(s, X))$,

choix idéal : $m^* \in \arg\min_{m \in \mathcal{M}} \ell(s, \hat{s}_m) = \arg\min_{m \in \mathcal{M}} P(\gamma(\hat{s}_m, X))$

$$\Longrightarrow crit_{HO}(m) = P_n^v \gamma(\hat{s}_m^t) := \frac{1}{|\mathbb{X}^v|} \sum_{X_i \in \mathbb{X}^v} \gamma(\hat{s}_m^t, X_i)$$

Validation croisée V-fold (VCVF)

Idée : $\mathbb{X} = \bigsqcup_{j=1}^{V} \mathbb{X}_j$ avec $|\mathbb{X}_j| = n/V \ \ \forall j \in \{1, \ldots, V\}$.
Pour chaque découpage j, $\mathbb{X} = \mathbb{X}_j^c \bigsqcup \mathbb{X}_j$,

1. $\mathbb{X}_j^c \Longrightarrow (\hat{s}_{m,j} := \mathcal{A}_m(\mathbb{X}_j^c))_{m \in \mathcal{M}}$
2. $\mathbb{X}_j \Longrightarrow$ définir $crit_j(m)$ qui évalue la qualité de m

$$\hat{m} \in \arg\min_{m \in \mathcal{M}} \frac{1}{V} \sum_{j=1}^{V} crit_j(m)$$

Validation croisée V-fold (VCVF)

Idée : $\mathbb{X} = \bigsqcup_{j=1}^{V} \mathbb{X}_j$ avec $|\mathbb{X}_j| = n/V \;\; \forall j \in \{1, \ldots, V\}$.
Pour chaque découpage j, $\mathbb{X} = \mathbb{X}_j^c \bigsqcup \mathbb{X}_j$,

1. $\mathbb{X}_j^c \Longrightarrow (\hat{s}_{m,j} := \mathcal{A}_m(\mathbb{X}_j^c))_{m \in \mathcal{M}}$
2. $\mathbb{X}_j \Longrightarrow$ définir $crit_j(m)$ qui évalue la qualité de m

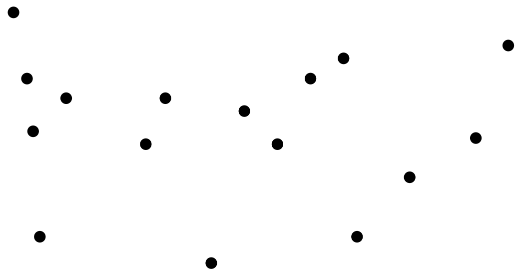
$$\hat{m} \in \arg\min_{m \in \mathcal{M}} \frac{1}{V} \sum_{j=1}^{V} crit_j(m)$$

Seule la définition de $crit_j(m)$ change, l'étape 1 est la même pour toutes les procédures de validation croisée !

Intuition

Pour chaque $j \in \{1, \ldots, V\}$, on a $\mathbb{X} = \mathbb{X}_j^c \bigsqcup \mathbb{X}_j$, où

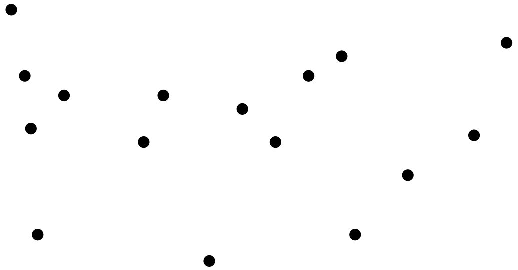
1. $\mathbb{X}_j^c \Longrightarrow (\hat{s}_{m,j} := \mathcal{A}_m(\mathbb{X}_j^c))_{m \in \mathcal{M}}$.
2. $\mathbb{X}_j \Longrightarrow$ comment choisir parmi une famille de points ?



Intuition

Pour chaque $j \in \{1, \ldots, V\}$, on a $\mathbb{X} = \mathbb{X}_j^c \bigsqcup \mathbb{X}_j$, où

1. $\mathbb{X}_j^c \Longrightarrow (\hat{s}_{m,j} := \mathcal{A}_m(\mathbb{X}_j^c))_{m \in \mathcal{M}}$.
2. $\mathbb{X}_j \Longrightarrow \psi_{l,m}(\mathbb{X}_j)$: test "‘robuste"’ entre $\hat{s}_{l,j}$ et $\hat{s}_{m,j}$



Test robuste

Propriété fondamentale : il existe des constantes $a > 0$, $\theta \in (0, 1/2)$, tel que pour deux densités quelconques t, u et $\forall z \in \mathbb{R}$, on peut trouver un test $\psi_{t,u}(\mathbb{X})$ qui satisfait :

$$\sup_{\{s | h(s,t) \leq \theta h(t,u)\}} \mathbb{P}[\psi_{t,u}(\mathbb{X}) = u] \leq \exp[-an(h^2(t,u) + z)];$$

$$\sup_{\{s | h(s,u) \leq \theta h(t,u)\}} \mathbb{P}[\psi_{t,u}(\mathbb{X}) = t] \leq \exp[-an(h^2(t,u) - z)].$$

Test robuste

Propriété fondamentale : il existe des constantes $a > 0$, $\theta \in (0, 1/2)$, tel que pour deux densités quelconques t, u et $\forall z \in \mathbb{R}$, on peut trouver un test $\psi_{t,u}(\mathbb{X})$ qui satisfait :

$$\sup_{\{s|h(s,t)\leq\theta h(t,u)\}} \mathbb{P}[\psi_{t,u}(\mathbb{X}) = u] \leq \exp[-an(h^2(t,u) + z)];$$

$$\sup_{\{s|h(s,u)\leq\theta h(t,u)\}} \mathbb{P}[\psi_{t,u}(\mathbb{X}) = t] \leq \exp[-an(h^2(t,u) - z)].$$

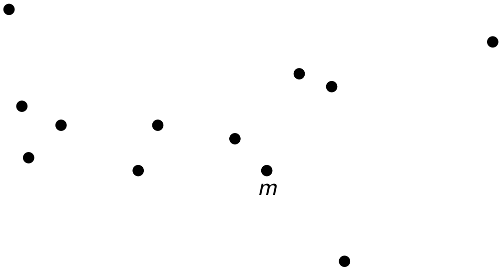
Le test de rapport de vraisemblance ne vérifie **pas** cette propriété de robustesse !

Nouvelle VCVF

$\forall j \in \{1, \ldots, V\},\ m \in \mathcal{M}$

$$crit_j(m) := \sup_{l \in \mathcal{R}_{m,j}} h^2(\hat{s}_{l,j}, \hat{s}_{m,j})$$

où $\mathcal{R}_{m,j} = \{l \in \mathcal{M},\ l \neq m \mid \psi_{l,m}(\mathbb{X}_j) = l\}$.

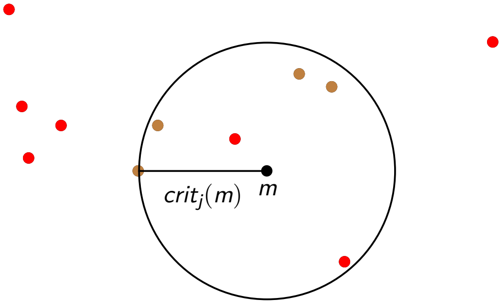


Nouvelle VCVF

$\forall j \in \{1, \ldots, V\}, \; m \in \mathcal{M}$

$$crit_j(m) := \sup_{l \in \mathcal{R}_{m,j}} h^2(\hat{s}_{l,j}, \hat{s}_{m,j})$$

où $\mathcal{R}_{m,j} = \{l \in \mathcal{M}, \; l \neq m \mid \psi_{l,m}(\mathbb{X}_j) = l\}$.



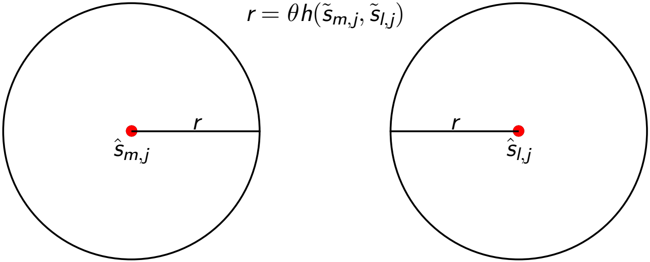
1ère possibilité : Test de boule (Birgé)

Cas classique avec $\gamma(t, x) = -\log(t(x))$

$\Longrightarrow$ rapport de vraisemblance entre $\hat{s}_{m,j}$ et $\hat{s}_{l,j}$: $\frac{\hat{s}_{m,j}}{\hat{s}_{l,j}}(\mathbb{X}_j)$

1ère possibilité : Test de boule (Birgé)

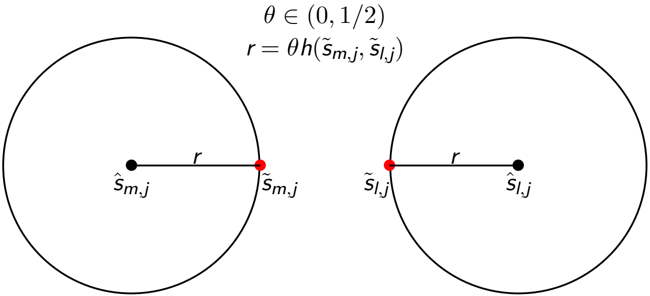
$$\theta \in (0, 1/2)$$
$$r = \theta h(\tilde{s}_{m,j}, \tilde{s}_{l,j})$$



Cas classique avec $\gamma(t, x) = -\log(t(x))$

$\Longrightarrow$ rapport de vraisemblance entre $\hat{s}_{m,j}$ et $\hat{s}_{l,j}$: $\frac{\hat{s}_{m,j}}{\hat{s}_{l,j}}(\mathbb{X}_j)$

1ère possibilité : Test de boule (Birgé)



Cas classique avec $\gamma(t, x) = -\log(t(x))$

$\Longrightarrow$ rapport de vraisemblance entre $\hat{s}_{m,j}$ et $\hat{s}_{l,j}$: $\frac{\hat{s}_{m,j}}{\hat{s}_{l,j}}(\mathbb{X}_j)$

$\Longrightarrow \psi_{l,m}(\mathbb{X}_j) = \frac{\tilde{s}_{m,j}}{\tilde{s}_{l,j}}(\mathbb{X}_j)$ test robuste

2ème possibilité : Formule variationnelle (Baraud)

Soient t et u deux densités, $\ell(s,t) = h^2(s,t) = 1 - \rho(s,t)$.

- Pour toute densité t : $\rho(s,t) = \inf_{\{v \text{ densité}\}} \ \rho_v(P,t)$,
  où $\rho_v(P,t) = \frac{1}{2}\left(\rho(t,v) + \int \sqrt{\frac{t}{v}} dP\right)$.
- Soit $v = \frac{t+u}{2}$ et $T(P,t,u) = \rho_v(P,t) - \rho_v(P,u)$.
  On a alors : $T(P,t,u) \geq 0 \Longrightarrow h^2(s,t) \leq \frac{\sqrt{2}+1}{\sqrt{2}-1} h^2(s,u)$.

2ème possibilité : Formule variationnelle (Baraud)

Soient t et u deux densités, $\ell(s,t) = h^2(s,t) = 1 - \rho(s,t)$.

- Pour toute densité t : $\rho(s,t) = \inf_{\{v \text{ densité}\}} \ \rho_v(P,t)$,
  où $\rho_v(P,t) = \frac{1}{2}\left(\rho(t,v) + \int \sqrt{\frac{t}{v}} dP\right)$.
- Soit $v = \frac{t+u}{2}$ et $T(P,t,u) = \rho_v(P,t) - \rho_v(P,u)$.
  On a alors : $T(P,t,u) \geq 0 \Longrightarrow h^2(s,t) \leq \frac{\sqrt{2}+1}{\sqrt{2}-1} h^2(s,u)$.

Idée : t meilleur que $u \Longleftrightarrow \rho(s,t) - \rho(s,u) \geq 0$

$$\leftrightsquigarrow T(P,t,u) \geq 0 \leftrightsquigarrow T(P_n,t,u) \geq 0$$

2ème possibilité : Formule variationnelle (Baraud)

Soient t et u deux densités, $\ell(s,t) = h^2(s,t) = 1 - \rho(s,t)$.

- Pour toute densité t : $\rho(s,t) = \inf_{\{v \text{ densité}\}} \; \rho_v(P,t)$,
  où $\rho_v(P,t) = \frac{1}{2}\left(\rho(t,v) + \int \sqrt{\frac{t}{v}} dP\right)$.
- Soit $v = \frac{t+u}{2}$ et $T(P,t,u) = \rho_v(P,t) - \rho_v(P,u)$.
  On a alors : $T(P,t,u) \geq 0 \Longrightarrow h^2(s,t) \leq \frac{\sqrt{2}+1}{\sqrt{2}-1} h^2(s,u)$.

Idée : t meilleur que $u \Longleftrightarrow \rho(s,t) - \rho(s,u) \geq 0$

$$\leftrightsquigarrow T(P,t,u) \geq 0 \leftrightsquigarrow T(P_n,t,u) \geq 0$$

Au découpage j, on utilise $\psi_{l,m}(\mathbb{X}_j) = l$ si $T(\mathbb{X}_j, \hat{s}_{m,j}, \hat{s}_{l,j}) \geq 0$.

Résumé

$$\hat{m}_{VF} = \operatorname*{argmin}_{m \in \mathcal{M}} \frac{1}{V} \sum_{j=1}^{V} \mathrm{crit}_j(m)$$

- Définition classique : $crit_j(m) := \frac{1}{|\mathbb{X}_j|} \sum_{X_i \in \mathbb{X}_j} \gamma(\hat{s}_{m,j}, X_i)$.
- Définition alternative : $crit_j(m) := \sup_{l \in \mathcal{R}_{m,j}} h^2(\hat{s}_{l,j}, \hat{s}_{m,j})$, avec $\mathcal{R}_{m,j} = \{l \in \mathcal{M} \;\; l \neq m \mid \psi_{l,m}(\mathbb{X}_j) = l\}$.
- différence dans l'étape de validation :
  - fonction de contraste (estimation du risque)
  - tests robustes (prend en compte les autres compétiteurs)

Comparaison au VF classique

Du point de vue

- pratique : semble performante en terme de risque, sa qualité augmente avec V
- algorithmique : nettement plus lente mais le coût n'est pas prohibitif
- théorique : une borne sur le risque Hellinger est possible sous une hypothèse très faible, MAIS les résultats restent insatisfaisants pour expliquer le choix de V

Perspectives : ne pas conclure !

Une autre procédure V-fold similaire peut être définie à l'aide du test de Baraud en effectuant la procédure V-fold sur le test !
Soit

- $\mathcal{T}(P_n, m, l) = \frac{1}{V}\sum_{j=1}^{V} T(\mathbb{X}_j, \hat{s}_{m,j}, \hat{s}_{l,j})$,
- $\mathcal{R}_m = \{l \in \mathcal{M},\ \ l \neq m \mid \mathcal{T}(P_n, l, m) \geq 0\}$,
- et $\mathcal{D}(m) = \sup_{l \in \mathcal{R}_m} h^2(\hat{s}_l, \hat{s}_m)$.

On choisit alors

$$\hat{m}_{\mathrm{TVF}} \in \underset{m \in \mathcal{M}}{\operatorname{argmin}}\ \mathcal{D}(m)\ \ .$$

Perspectives : ne pas conclure !

Une autre procédure V-fold similaire peut être définie à l'aide du test de Baraud en effectuant la procédure V-fold sur le test !
Soit

- $\mathcal{T}(P_n, m, l) = \frac{1}{V}\sum_{j=1}^{V} T(\mathbb{X}_j, \hat{s}_{m,j}, \hat{s}_{l,j})$,
- $\mathcal{R}_m = \{l \in \mathcal{M},\ l \neq m \mid \mathcal{T}(P_n, l, m) \geq 0\}$,
- et $\mathcal{D}(m) = \sup_{l \in \mathcal{R}_m} h^2(\hat{s}_l, \hat{s}_m)$.

On choisit alors

$$\hat{m}_{\mathrm{TVF}} \in \underset{m \in \mathcal{M}}{\operatorname{argmin}}\ \mathcal{D}(m)\ .$$

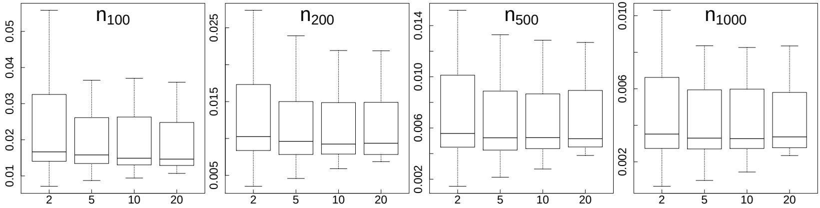
Avantages : temps de calcul nettement moins lourd et très bon en pratique également !
Espoir : inégalité oracle plus fine grâce aux outils classiques des inégalités de concentration.

Fin

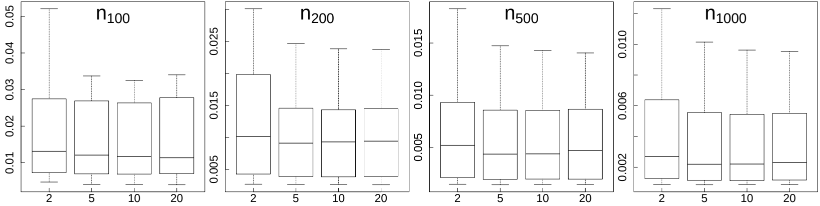
Merci

Influence de V ; $\ell = h^2$

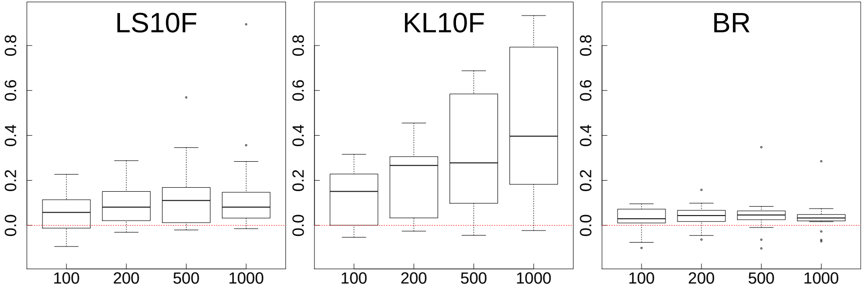
Histogrammes réguliers :



Estimateurs à noyau gaussien :



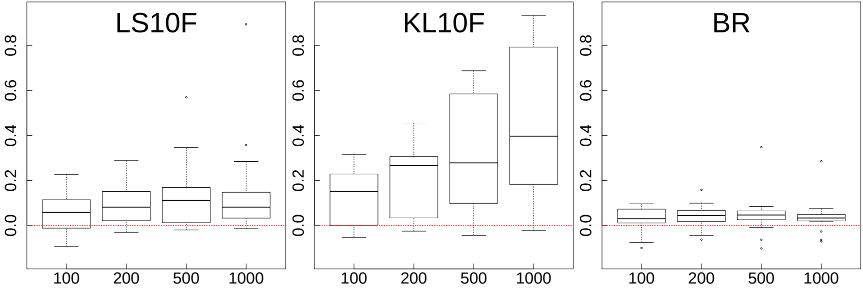
Histogrammes réguliers : $\ell = h^2$



La différence

- La différence augmente avec V pour les 3 procédures
- La différence augmente avec n pour le KLVF

Histogrammes réguliers : $\ell = h^2$



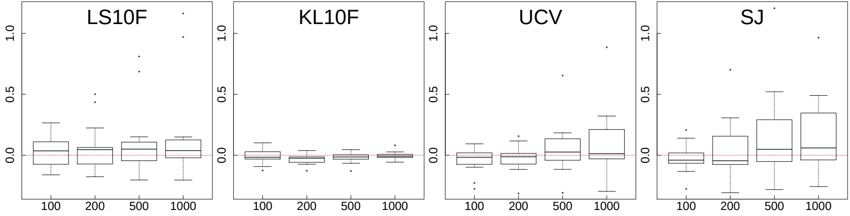
La différence

- La différence augmente avec V pour les 3 procédures
- La différence augmente avec n pour le KLVF

Même conclusions pour les autres pertes !

Estimateurs à noyau : $\ell = h^2$

- ucv unbiased cross-validation
- SJ implements the methods of Sheather & Jones (1991) to select the bandwidth using pilot estimation of derivatives



Histogrammes réguliers, irréguliers et estimateurs à noyau