Apprentissage Statistique
M2 Probabilités et Statistiques, Université Paris-Sud

Cours 6 : Sélection d'estimateurs par rééchantillonnage et validation croisée

SYLVAIN ARLOT ET FRANCIS BACH

1. Régression hétéroscédastique

Estimateurs linéaires, $Y = F + \varepsilon$, $\widehat{F} = A_m Y$.

Calcul de la pénalité idéale :

$$\begin{aligned}\mathrm{pen}_{\mathrm{id}}(m) &= \frac{1}{n}\left\|\widehat{F}_m - F\right\|^2 - \frac{1}{n}\left\|\widehat{F}_m - Y\right\|^2 + \frac{1}{n}\|\varepsilon\|^2 \\ &= \frac{2}{n}\langle A_m\varepsilon,\, \varepsilon\rangle + \frac{2}{n}\big\langle (A_m - I)F,\, \varepsilon\big\rangle\end{aligned}$$

Que vaut l'espérance si l'on suppose juste que les ε_i sont centrés, indépendants, et de variances respectives σ_i^2 (cadre hétéroscédastique) ?

$$\mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m)\right] = \mathbb{E}\left[\frac{2}{n}\langle A_m\varepsilon,\, \varepsilon\rangle\right] \quad .$$

Notons D_σ la matrice diagonale dont les coefficients diagonaux sont les σ_i. Alors, $\varepsilon = D_\sigma \xi$ où ξ est un vecteur centré, dont les coordonnées sont indépendantes et de même variance 1, et

$$\begin{aligned}\mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m)\right] = \mathbb{E}\left[\frac{2}{n}\langle A_m D_\sigma\xi,\, D_\sigma\xi\rangle\right] &= \mathbb{E}\left[\frac{2}{n}\left\langle D_\sigma^\top A_m D_\sigma\xi,\, \xi\right\rangle\right] \\ &= \frac{2}{n}\,\mathrm{tr}(D_\sigma^\top A_m D_\sigma) \\ &= \frac{2}{n}\sum_{i=1}^n \left(\sigma_i^2 (A_m)_{i,i}\right) \quad .\end{aligned}$$

Si les σ_i sont inconnus, on ne connaît donc plus la forme de la pénalité idéale. Comment l'estimer ?

Le même problème se poserait si on considérait des estimateurs $\widehat{F}_m$ non-linéaires.

Dans les deux cas, on peut utiliser le rééchantillonnage pour estimer une quantité telle que $\mathrm{pen}_{\mathrm{id}}(m)$.

Date: 16 Mars 2015.

2. Rééchantillonnage

Référence : troisième cours de [2].

2.1. Principe. Exemples.

– Heuristique du bootstrap [5] : on observe $\xi_1, \dots, \xi_n$ i.i.d. de loi commune P, on note P_n leur mesure empirique $n^{-1}\sum_{i=1}^n \delta_{\xi_i}$, et on cherche à estimer la distribution d'une quantité de la forme $F(P, P_n)$. On génère un rééchantillon $\xi_1^\star, \dots, \xi_n^\star$ i.i.d. de loi commune P_n, on définit sa mesure empirique

$$P_n^\star = P_n^W = \frac{1}{n}\sum_{i=1}^n \delta_{\xi_i^\star} \ .$$

L'heuristique du bootstrap est que $\mathcal{D}(F(P_n, P_n^W)|P_n)$ estime $\mathcal{D}(F(P, P_n))$.

– Bootstrap à poids échangeables :

$$P_n^W = \sum_{i=1}^n W_{i,n}\delta_{\xi_i}$$

où $W = (W_{i,n})_{1\leqslant i\leqslant n}$ est un vecteur aléatoire, indépendant de P_n, tel que $W_{i,n} \geqslant 0$ p.s., $\mathbb{E}[W_{i,n}] = 1$ pour tout i, et pour toute permutation τ de $\{1, \dots, n\}$, W a la même loi que $(W_{\tau(i),n})_{1\leqslant i\leqslant n}$. On note $\mathbb{E}_W[\cdot]$ l'espérance relative à l'aléa de W uniquement.

– Exemples :
  – bootstrap (W multinomial de paramètres $(n; 1/n, \dots, 1/n)$), bootstrap « m out of n » ($(m/n)W$ multinomial de paramètres $(m; 1/n, \dots, 1/n)$), bootstrap Poissonisé (m suit une loi de Poisson, ce qui rend les poids W_i indépendants)
  – sous-échantillonnage : $W_{i,n} = \kappa \mathbb{1}_{i\in I}$ pour une partie I de $\{1, \dots, n\}$ choisie aléaoirement. Par exemple, random hold-out ($\kappa = n/m$ et I choisie uniformément parmi les parties de taille m de $\{1, \dots, n\}$), ou Bernoulli (p) ($pW_{i,n}$ indépendants de loi de Bernoulli de paramètre $p \in]0; 1[$; en particulier, $p = 1/2$ correspond aux poids « Rademacher »).

– Heuristique de rééchantillonnage : $\mathcal{D}(F(P_n, P_n^W)|P_n)$ estime (à un facteur d'échelle C_W près) $\mathcal{D}(F(P, P_n))$.

2.2. Utilisations en statistique.

– estimation ou correction du biais de $\widehat{s}(P_n) \in \mathbb{R}$ comme estimateur de $\mu(P)$ [5] :

$$\mathbb{E}\big[\widehat{s}(P_n) - \mu(P)\big] \quad \text{estimé par} \quad C_W \mathbb{E}_W\big[\widehat{s}(P_n^W) - \mu(P_n)\big] = C_W \mathbb{E}\big[\widehat{s}(P_n^W) - \mu(P_n) \,\big|\, P_n\big] \ .$$

– estimation de la variance [5] :

$$\operatorname{var}\big[\widehat{s}(P_n)\big] \quad \text{estimé par} \quad C_W \operatorname{var}_W\big[\widehat{s}(P_n^W)\big] = C_W \operatorname{var}\big[\widehat{s}(P_n^W) \,\big|\, P_n\big] \ .$$

– estimation du risque quadratique de $\widehat{s}(P_n) \in \mathbb{R}$ comme estimateur de $\mu(P)$:

$$\mathbb{E}\left[\big(\widehat{s}(P_n) - \mu(P)\big)^2\right] \quad \text{estimé par} \quad C_W \mathbb{E}_W\left[\big(\widehat{s}(P_n^W) - \mu(P_n)\big)^2\right] \ .$$

– construction d'un intervalle de confiance pour la moyenne $\mu(P)$ de $\xi_1, \ldots, \xi_n \in \mathbb{R}$: on cherche un t_α tel que

$$\big[\mu(P_n) - t_\alpha, \mu(P_n) + t_\alpha\big]$$

est un intervalle confiance de probabilité de couverture $1-\alpha$ pour $\mu(P)$, où

$$\mu(P_n) = \frac{1}{n}\sum_{i=1}^n \xi_i$$

est la moyenne empirique de l'échantillon. Autrement dit, on voudrait avoir

$$\mathbb{P}\Big(\big|\mu(P_n) - \mu(P)\big| \leqslant t_\alpha\Big) \geqslant 1 - \alpha \ .$$

Le choix idéal serait $t_\alpha^\star$, le quantile d'ordre $(1-\alpha)$ de $\mathcal{L}\big(|\mu(P_n) - \mu(P)|\big)$. On peut l'estimer par rééchantillonnage par le quantile d'ordre $(1-\alpha)$ de

$$\mathcal{L}\big(C_W\big|\mu(P_n^W) - \mu(P_n)\big| \,|\, P_n\big) \ .$$

2.3. **Avantages et limites.**

– Avantages : généralité, adaptativité. Théorie asymptotique (processus empiriques et bootstrap à poids échangeables) : voir le chapitre 3.6 de [11].
Propriété de stabilisation [8], utilisée notamment pour le « bagging » (bootstrap aggregating).

– Limites : (universalité/abus d'utilisation).
Que faire si les ξ_i ne sont pas i.i.d. ? Bootstrapper les résidus dans le cas d'un design fixe avec des erreurs i.i.d.
Bootstrap par blocs pour gérer la dépendance (à courte portée) de données stationnaires.

2.4. **Étude d'un cas : un estimateur du risque quadratique.** Référence : Section 4 du troisième cours de [2].

Soient $\xi_1, \ldots, \xi_n$ des variables aléatoires i.i.d. de moyenne μ et de variance σ^2. Un estimateur naturel de μ est la moyenne empirique $n^{-1}\sum_{i=1}^n \xi_i$, dont le risque quadratique vaut

$$\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n \xi_i - \mu\right)^2\right] = \frac{\sigma^2}{n} \ . \tag{1}$$

L'estimateur par rééchantillonnage du risque quadratique vaut alors (en renormalisant les poids W_i pour que leur somme soit toujours égale à n)

$$\mathbb{E}_W\left[\left(\frac{1}{n}\sum_{i=1}^n \frac{W_i}{\overline{W}}\xi_i - \frac{1}{n}\sum_{i=1}^n \xi_i\right)^2\right] \qquad \text{avec} \quad \overline{W} = \frac{1}{n}\sum_{i=1}^n W_i \ .$$

Au vu de (1), on note cet estimateur $\widehat{\sigma^2_W}/n$, puisque

$$\widehat{\sigma^2_W} = n\mathbb{E}_W\left[\left(\frac{1}{n}\sum_{i=1}^n \frac{W_i}{\overline{W}}\xi_i - \frac{1}{n}\sum_{i=1}^n \xi_i\right)^2\right]$$

peut être vu comme un estimateur de σ^2.

2.4.1. *Formule close.* Calculons $\widehat{\sigma^2_W}$ pour des poids W échangeables généraux :

$$\begin{aligned}\widehat{\sigma^2_W} &= n\mathbb{E}_W\left[\left(\frac{1}{\sum_{k=1}^n W_k}\sum_{i=1}^n (W_i\xi_i) - \frac{1}{n}\sum_{i=1}^n \xi_i\right)^2\right] \\ &= \frac{1}{n}\mathbb{E}_W\left[\left(\sum_{i=1}^n\left[\left(\frac{nW_i}{\sum_{k=1}^n W_k} - 1\right)\xi_i\right]\right)^2\right]\end{aligned}$$

Ainsi,

$$\begin{aligned}\widehat{\sigma^2_W} &= \frac{1}{n}\mathbb{E}_W\left[\sum_{i,j}\left[\left(\frac{nW_i}{\sum_{k=1}^n W_k} - 1\right)\left(\frac{nW_j}{\sum_{k=1}^n W_k} - 1\right)\xi_i\xi_j\right]\right] \\ &= \frac{1}{n}\sum_{i=1}^n\left[R_V^{(W)}\xi_i^2\right] + \frac{1}{n}\sum_{i\neq j}\left[R_C^{(W)}\xi_i\xi_j\right]\end{aligned}$$

en posant

$$\begin{aligned} R_V^{(W)} &:= \mathbb{E}_W\left[\left(\frac{nW_i}{\sum_{k=1}^n W_k} - 1\right)^2\right] \\ \text{et} \quad R_C^{(W)} &:= \mathbb{E}_W\left[\left(\frac{nW_i}{\sum_{k=1}^n W_k} - 1\right)\left(\frac{nW_j}{\sum_{k=1}^n W_k} - 1\right)\right]\end{aligned}$$

pour des $i \neq j$ quelconques (ces quantités ne dépendent pas de (i,j) si $i \neq j$ car W est échangeable). Remarquons maintenant que

$$0 = \mathbb{E}_W\left[\left(\sum_{i=1}^n\left(\frac{nW_i}{\sum_{k=1}^n W_k} - 1\right)\right)^2\right] = nR_V^{(W)} + n(n-1)R_C^{(W)}$$

et donc, en supposant $n \geqslant 2$,

$$R_C^{(W)} = \frac{-1}{n-1}R_V^{(W)} \ .$$

Comme $R_V^{(W)} = 0$ lorsque $n = 1$, on en déduit que

$$\widehat{\sigma_W^2} = \frac{R_V^{(W)}}{n} \mathbb{1}_{n \geqslant 2} \left[\sum_{i=1}^n \xi_i^2 - \frac{1}{n-1} \sum_{i \neq j} \xi_i \xi_j \right] \ . \tag{2}$$

Notons également que $\widehat{\sigma_W^2}$ est invariant par translation des données car

$$\sum_{i=1}^n \left(\frac{W_i}{\sum_{j=1}^n W_j} - 1 \right) = 0 \ ,$$

la formule (2) est donc encore valable en remplaçant $(\xi_i)_i$ par $(\xi_i - \mu)_i$ (par exemple).

2.4.2. *Comparaison avec l'estimateur classique.* L'estimateur sans biais classique de la variance s'écrit

$$\begin{aligned} \widehat{\sigma^2} = \frac{1}{n-1} \sum_{i=1}^n \left(\xi_i - \frac{1}{n} \sum_{k=1}^n \xi_k \right)^2 &= \frac{n}{n-1} \left[\frac{1}{n} \sum_{i=1}^n \xi_i^2 - \left(\frac{1}{n} \sum_{i=1}^n \xi_i \right)^2 \right] \\ &= \frac{1}{n} \left(\sum_{i=1}^n \xi_i^2 - \frac{1}{n-1} \sum_{i \neq j} \xi_i \xi_j \right) \ . \end{aligned} \tag{3}$$

Ainsi,

$$\widehat{\sigma_W^2} = R_V^{(W)} \widehat{\sigma^2} \ ,$$

d'où l'on déduit notamment que

$$\mathbb{E}\left[\widehat{\sigma_W^2} \right] = R_V^{(W)} \sigma^2 \ . \tag{4}$$

2.5. **Pénalités par rééchantillonnage.**

- Définition générale [6, 1] : estimateur par rééchantillonnage de (l'espérance de) la pénalité idéale $(P - P_n)\gamma(\widehat{s}_m(P_n))$, soit :

$$\mathrm{pen}_{\mathrm{reech}}(m) = C_W \mathbb{E}_W \left[(P_n - P_n^W) \gamma\big(\widehat{s}_m\left(P_n^W\right)\big) \right] \ .$$

- Lien avec les pénalités Rademacher (globales) [7], qui sont l'estimateur par rééchantillonnage (avec des poids Rademacher) de $\mathrm{pen}_{\mathrm{id,g}}(m)$.
- Il existe des résultats (inégalités-oracle) en régression dans le cas hétéroscédastique [1] et en estimation de densité [9].

3. Validation croisée

La référence principale de cette section est [3]. Voir aussi le quatrième cours de [2].

3.1. **Calibration/Sélection d'algorithmes.**

- Définition d'un algorithme statistique (règle d'apprentissage / de prédiction) $\mathcal{A} : \bigcup_{k \in \mathbb{N}} (\mathcal{X} \times \mathcal{Y})^k \to \mathbb{S}$.

– Problème du choix d'algorithme (cadre de la prédiction, fonction de contraste γ) : étant donnée $(\mathcal{A}_m)_{m\in\mathcal{M}}$ et un échantillon D_n, choisir un $\widehat{m}(D_n)$ tel que
$$P\gamma\left(\mathcal{A}_{\widehat{m}(D_n)}(D_n)\right)$$
est minimal. Oracle, inégalité oracle.
– Exemples : sélection de modèles, choix d'un estimateur linéaire en régression, choix de k pour les k plus proches voisins en classification, choix du paramètre de régularisation ou d'un noyau pour les SVM, etc.
– Compromis entre sur-apprentissage et sous-apprentissage.

3.2. **Principe de la validation croisée.**
– Heuristique générale : besoin de « nouvelles données » pour estimer sans biais la perte de $\mathcal{A}(D_n)$ (contrairement au risque empirique $P_n\gamma(\mathcal{A}(D_n))$, qui est en général trop optimiste). Idée : découper l'échantillon D_n en un échantillon d'entraînement $D_n^{(e)}$ et un échantillon de validation $D_n^{(v)}$ indépendants.
– Définitions : estimateur par validation (hold-out) du risque :
$$\widehat{\mathcal{R}}^{\mathrm{val}}\big(\mathcal{A}; D_n; I^{(e)}\big) = P_n^{(v)}\gamma\Big(\mathcal{A}\big(D_n^{(e)}\big)\Big)$$
$$\text{où}\quad D_n^{(e)} = \big\{(X_i,Y_i)\big\}_{i\in I^{(e)}} \quad \text{et} \quad P_n^{(v)} = \frac{1}{\mathrm{Card}(I^{(v)})}\sum_{i\in I^{(v)}}\delta_{(X_i,Y_i)}\ ,$$
estimateur par validation croisée (VC) du risque :
$$\widehat{\mathcal{R}}^{\mathrm{vc}}\left(\mathcal{A}; D_n; \left(I_j^{(e)}\right)_{1\leqslant j\leqslant B}\right) = \frac{1}{B}\sum_{j=1}^{B}\widehat{\mathcal{R}}^{\mathrm{val}}\left(\mathcal{A}; D_n; I_j^{(e)}\right)\ .$$
– Hypothèses :
$\left(I_j^{(e)}\right)_{1\leqslant j\leqslant B}$ choisis indépendamment de D_n (si ce choix est aléatoire), et pour tout j, $\mathrm{Card}(I_j^{(e)}) = n_e \in \{1,\dots,n-1\}$.

3.3. **Exemples.**
– Exploration exhaustive : Leave-one-out ($n_e = n-1$), Leave-p-out ($n_e = n-p$).
– Exploration partielle : VC Monte-Carlo (MCCV), Apprentissage-test répété (RLT), VC « V-fold » (VFCV ; $n_e = n(V-1)/V$, $B = V$).

3.4. **Estimation du risque par validation croisée.**

3.4.1. *Biais.*
– Calcul du biais de la VC pour l'estimation du risque :
$$\begin{aligned}\mathbb{E}\left[\widehat{\mathcal{R}}^{\mathrm{vc}}\left(\mathcal{A}; D_n; \left(I_j^{(e)}\right)_{1\leqslant j\leqslant B}\right)\right] &= \mathbb{E}\left[\widehat{\mathcal{R}}^{\mathrm{val}}\left(\mathcal{A}; D_n; I_1^{(e)}\right)\right]\\ &= \mathbb{E}\left[P\gamma\left(\mathcal{A}(D_{n,1}^{(e)})\right)\right]\\ &= \mathbb{E}\Big[P\gamma\big(\mathcal{A}(D_{n_e})\big)\Big]\ .\end{aligned}$$

Donc, si $\mathcal{R}(n,\mathcal{A}) = \mathbb{E}\left[P\gamma\left(\mathcal{A}(D_n)\right)\right]$, alors le biais de l'estimateur du risque par VC vaut

$$\mathbb{E}\left[\widehat{\mathcal{R}}^{\mathrm{vc}}\left(\mathcal{A}; D_n; \left(I_j^{(e)}\right)_{1\leqslant j\leqslant B}\right)\right] - \mathbb{E}\Big[P\gamma\big(\mathcal{A}(D_n)\big)\Big] = \mathcal{R}(n_e,\mathcal{A}) - \mathcal{R}(n,\mathcal{A}) \ .$$

– Règle intelligente [4] : $\mathcal{A}$ telle que $(\mathcal{R}(n,\mathcal{A}))_{n\in\mathbb{N}}$ décroissante pour toute distribution P. Alors, le biais de la VC est positif, et d'autant plus petit que n_e est proche de n.
Exemple : $\mathcal{A}$ estimateur par projection sur un modèle linéaire en régression, si γ est le contraste des moindres carrés, car

$$\mathcal{R}(n,\mathcal{A}) = a(\mathcal{A},P) + \frac{b(\mathcal{A},P)}{n} \quad \text{avec} \quad a(\mathcal{A},P), b(\mathcal{A},P) \geqslant 0 \ .$$

Contre-exemple : règle du plus proche voisin en classification (voir [4]).

3.4.2. *Variance.*

– Lemme : pour toutes variables aléatoires X, Z,

$$\mathrm{var}(Z) = \mathbb{E}\big[\mathrm{var}(Z \,|\, X)\big] + \mathrm{var}\big(\mathbb{E}[Z \,|\, X]\big) \ .$$

– Calcul général pour la validation simple :

$$\mathrm{var}\left(P_n^{(v)}\gamma\left(\widehat{s}_m^{(e)}\right)\right) = \frac{1}{n_v}\mathbb{E}\Big[\mathrm{var}\Big(\gamma(u;\xi)\,\Big|\, u=\widehat{s}_m^{(e)}\Big)\Big] + \mathrm{var}\Big(P\gamma\big(\widehat{s}_m(D_{n_e})\big)\Big)$$

– Calcul général pour la validation croisée « Monte Carlo » ($n_e = n-p$) :

$$\mathrm{var}\left(\widehat{\mathcal{R}}^{\mathrm{vc}}\left(\widehat{s}_m; D_n; \left(I_j^{(e)}\right)_{1\leqslant j\leqslant B}\right)\right) = \mathrm{var}\Big(\widehat{\mathcal{R}}^{\ell\mathrm{po}}\big(\widehat{s}_m; D_n\big)\Big) + \frac{1}{B}\mathbb{E}\left[\mathrm{var}_{I^{(e)}}\left(P_n^{(v)}\gamma\Big(\widehat{s}_m^{(e)}\Big)\,\Big|\, D_n\right)\right]$$

– Facteurs de variabilité : taille n_v de l'échantillon de validation (l'augmenter fait diminuer la variance, à n_e fixe du moins), « stabilité » de $\mathcal{A}$ (pour un échantillon de taille n_e), nombre B de découpages considéré.
– En général, la variance est difficile à quantifier précisément, car n_e et n_v sont toujours liés ($n_e + n_v = n$), et parfois B leur est lié également (e.g., VFCV).
– Le type de VC le moins variable dépend du cadre dans lequel on se trouve.

3.5. **Propriétés de la VC en sélection de modèles.**

$$\widehat{m}^{\mathrm{vc}}(D_n) \in \operatorname*{argmin}_{m\in\mathcal{M}}\left\{\widehat{\mathcal{R}}^{\mathrm{vc}}\left(\mathcal{A}_m; D_n; \left(I_j^{(e)}\right)_{1\leqslant j\leqslant B}\right)\right\}$$

– Résultats asymptotiques (voir par exemple [10], en régression) : le comportement de la VC dépend essentiellement de n_e (et reste le même pour la plupart des méthodes de VC, avec peu d'hypothèses sur B) :

(1) Si $n_e \sim n$, alors, on a une inégalité-oracle avec une constante $C = C_n \to 1$ quand $n \to \infty$.

(2) Si $n_e \sim \kappa n$ avec $\kappa \in \,]0;1[$, alors, on a une inégalité-oracle avec une constante $C = C_n \to C(\kappa) > 1$ quand $n \to \infty$, et l'excès

de risque de l'estimateur sélectionné est sous-optimal (perte d'un facteur multiplicatif $C'(\kappa) > 1$).

(3) Si $n_e \ll n$, alors, on n'a pas d'inégalité-oracle avec une constante $C = \mathcal{O}(1)$ quand $n \to \infty$ (perte d'un facteur multiplicatif tendant vers l'infini avec n).

– Que se passe-t-il à n et σ^2 fixés ?
Les performances sont souvent (légèrement) meilleures quand on surpénalise légèrement (d'où l'intérêt à choisir n_e un peu plus petit que ce que préconise les résultats asymptotiques).
La variance de la VC joue un rôle majeur (le hold-out donnant des résultats bien plus mauvais que la VFCV avec $V \geqslant 10$).
– Comment choisir une méthode de VC pour un problème de choix d'algorithme donné ? Compromis entre biais, variance et complexité algorithmique.

Références

[1] Sylvain Arlot. Model selection by resampling penalization. *Electron. J. Stat.*, 3 :557–624 (electronic), 2009.

[2] Sylvain Arlot. Sélection de modèles et sélection d'estimateurs pour l'apprentissage statistique, January 2011. Cours Peccot. Collège de France. `http://www.di.ens.fr/~arlot/peccot.htm`.

[3] Sylvain Arlot and Alain Celisse. A survey of cross-validation procedures for model selection, July 2009. arXiv :0907.4728v1.

[4] Luc P. Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*, volume 31 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, 1996.

[5] Bradley Efron. Bootstrap methods : another look at the jackknife. *Ann. Statist.*, 7(1) :1–26, 1979.

[6] Bradley Efron. Estimating the error rate of a prediction rule : improvement on cross-validation. *J. Amer. Statist. Assoc.*, 78(382) :316–331, 1983.

[7] Magalie Fromont. Model selection by bootstrap penalization for classification. *Mach. Learn.*, 66(2–3) :165–207, 2007.

[8] Peter Hall. *The Bootstrap and Edgeworth Expansion*. Springer Series in Statistics. Springer-Verlag, New York, 1992.

[9] Matthieu Lerasle. Optimal model selection in density estimation. arXiv :0910.1654v2, 2009.

[10] Jun Shao. An asymptotic theory for linear model selection. *Statist. Sinica*, 7(2) :221–264, 1997. With comments and a rejoinder by the author.

[11] Aad W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996. With applications to statistics.