Sélection de modèles et sélection d'estimateurs pour l'Apprentissage statistique (Cours Peccot)

Quatrième cours: Validation croisée et pénalités reliées

SYLVAIN ARLOT
CNRS – ÉQUIPE SIERRA

Table des matières

1. Validation croisée

Nous renvoyons à l'article de survol [6] au sujet des différentes méthodes de validation croisée, à la fois vues comme moyens d'estimer le risque d'un estimateur, et comme procédures de sélection de modèles (ou d'estimateurs).

Date: 31 Janvier 2011.

Le cadre et les notations utilisées dans cette section (ainsi que dans le reste de ce cours) sont celles introduites en Sections 1–2 du premier cours ; voir aussi [6, Section 1].

Au vu du principe d'estimation sans biais du risque, les propriétés de la validation croisée pour la sélection d'estimateurs dépend fortement de ses propriétés en tant que procédure d'estimation du risque d'un estimateur isolé. Nous supposerons donc dans cette section un algorithme statistique $\mathcal{A}_m$ fixé, et considérerons la validation croisée comme une méthode d'estimation du risque

$$\mathbb{E}_{D_n \sim P^{\otimes n}} \left[\ell \left(s^\star, \mathcal{A}_m(D_n) \right) \right] \quad .$$

1.1. **Principe et définition générale.** Voir [6, Sections 4.1–4.2].

Le principe général des méthodes de validation croisée est de découper l'échantillon D_n en deux sous-échantillons disjoints : l'échantillon d'entraînement $D_n^{(e)} = (\xi_i)_{i \in I^{(e)}}$ et l'échantillon de validation $D_n^{(v)} = (\xi_i)_{i \in I^{(v)}}$, où $I^{(e)}$ est un sous-ensemble propre [1] de $\{1, \dots, n\}$ et $I^{(v)} = (I^{(e)})^c = \{1, \dots, n\} \setminus I^{(e)}$. On utilise alors $D_n^{(e)}$ pour «entraîner» l'algorithme $\mathcal{A}_m$, puis $D_n^{(v)}$ pour mesurer l'erreur de prédiction de $\mathcal{A}_m(D_n^{(e)})$. On obtient alors l'estimateur par *validation simple* [16, hold-out] du risque de $\mathcal{A}_m$:

$$\begin{aligned} \widehat{\mathcal{R}}^{\mathrm{val}} \left(\mathcal{A}_m; D_n; I^{(e)} \right) &:= P_n^{(v)} \gamma \left(\mathcal{A}_m(D_n^{(e)}) \right) \\ &= \frac{1}{n_v} \sum_{i \in D_n^{(v)}} \gamma \left(\mathcal{A}_m(D_n^{(e)}); \xi_i \right) \quad . \end{aligned} \tag{1}$$

En répétant ce procédé de découpage en deux sous-échantillons, on obtient la forme générale des estimateurs par *validation croisée* du risque de $\mathcal{A}_m$ [20]. Soit $B \geq 1$ un entier et $I_1^{(e)}, \dots, I_B^{(e)}$ une suite de sous-ensembles propres de $\{1, \dots, n\}$. Alors, l'estimateur par validation croisée du risque de $\mathcal{A}_m(D_n)$, avec ensembles d'entraînement $(I_j^{(e)})_{1 \leq j \leq B}$, est défini par

$$\widehat{\mathcal{R}}^{\mathrm{vc}} \left(\mathcal{A}_m; D_n; \left(I_j^{(e)} \right)_{1 \leq j \leq B} \right) := \frac{1}{B} \sum_{j=1}^{B} \widehat{\mathcal{R}}^{\mathrm{val}} \left(\mathcal{A}_m; D_n; I_j^{(e)} \right) \quad . \tag{2}$$

Tous les estimateurs par validation croisée habituels sont de la forme (2). Chacun est défini de manière unique par $(I_j^{(e)})_{1 \leq j \leq B}$, c'est-à-dire, par la manière de découper successivement l'échantillon.

1.2. **Exemples.** Voir [6, Section 4.3].

Le plus souvent, on considère des estimateurs par validation croisée avec une taille d'échantillon d'entraînement fixe, c'est-à-dire, de la forme (2) avec $\mathrm{Card}(I_j^{(e)}) = n_e$ pour tout $j \in \{1, \dots, B\}$. On peut alors distinguer deux familles principales d'estimateurs par validation croisée.

1. Propre : non-vide et de complémentaire non-vide

Découpages exhaustifs. On peut tout d'abord procéder à une exploration *exhaustive* des découpages de l'échantillon, en prenant $\{I_j^{(e)} \text{ t.q. } 1 \leq j \leq B\} = \mathfrak{P}_{n_e}(\{1,\dots,n\})$ l'ensemble des parties de $\{1,\dots,n\}$ à n_e éléments. Lorsque $n_e = n-1$, on obtient l'estimateur par «*leave-one-out*» [29, 1, 20, LOO], défini par (2) avec $B = n$ et $I_j^{(e)} = \{j\}^c$ pour $j = 1,\dots,n$:

$$\widehat{\mathcal{R}}^{\text{loo}}(\mathcal{A}_m; D_n) = \frac{1}{n}\sum_{j=1}^{n} \gamma\left(\mathcal{A}_m\left(D_n^{(-j)}\right); \xi_j\right) \tag{3}$$

où $D_n^{(-j)} = (\xi_i)_{i\neq j}$.

Lorsque $n_e = n-p$ avec $p \in \{1,\dots,n-1\}$, on obtient l'estimateur par «*leave-p-out*» [27, LPO], défini par (2) avec $B = \binom{n}{p}$ et $(I_j^{(e)})_{1\leq j\leq B}$ est l'ensemble des parties de $\{1,\dots,n\}$ de taille $n-p$.

Découpages partiels. Pour des raisons algorithmiques, il est souvent difficile d'entraîner $\binom{n}{p}$ fois l'algorithme $\mathcal{A}_m$. C'est pourquoi plusieurs alternatives ont été proposées aux LOO et LPO, reposant sur une *exploration partielle* de $\mathfrak{P}_{n_e}(\{1,\dots,n\})$.

La méthode la plus utilisée est probablement la validation croisée «V-fold» [20, VFCV], définie comme suit pour un $V \in \{1,\dots,n\}$ quelconque. On commence par choisir (indépendamment des données) une partition $\mathcal{B} = (B_j)_{1\leq j\leq V}$ de $\{1,\dots,n\}$ en V sous-ensembles (approximativement) de même taille n/V. Ensuite, pour tout $j = 1,\dots,V$, on va utiliser B_j^c pour l'entraînement et B_j pour la validation. Formellement, l'estimateur par validation croisée «V-fold» du risque de $\mathcal{A}_m$ est défini par (2) avec $B = V$ et $I_j^{(e)} = B_j^c$ pour $j = 1,\dots,B$:

$$\begin{aligned}
\widehat{\mathcal{R}}^{\text{vf}}(\mathcal{A}_m; D_n; \mathcal{B}) &= \frac{1}{V}\sum_{j=1}^{V} P_n^{(B_j)}\gamma\left(\mathcal{A}_m\left(D_n^{(-B_j)}\right)\right) \\
&= \frac{1}{V}\sum_{j=1}^{V}\left[\frac{1}{\text{Card}(B_j)}\sum_{i\in B_j}\gamma\left(\mathcal{A}_m\left(D_n^{(-B_j)}\right); \xi_i\right)\right] \\
\text{où}\quad D_n^{(-B_j)} &:= (\xi_i)_{i\in B_j^c} \quad \text{et} \quad P_n^{(B_j)} := \frac{1}{\text{Card}(B_j)}\sum_{i\in B_j}\delta_{\xi_i}\ .
\end{aligned}$$

La complexité algorithmique de VFCV est seulement V fois celle de l'entraînement de $\mathcal{A}_m$ sur un échantillon de taille $n - n/V$, soit beaucoup moins que le LOO ou le LPO si $V \ll n$. Notons que la VFCV avec $V = n$ donne le LOO.

On peut mentionner trois autres méthodes de validation croisée reposant sur une exploration partielle de $\mathfrak{P}_{n_e}(\{1,\dots,n\})$:

- «Balanced Incomplete Cross Validation» [27, BICV], apparentée à la VFCV avec notamment la possibilité de prendre n_e plus petit que $n/2$.
- «Repeated learning-testing» [12, RLT] : on choisit $(I_j^{(e)})_{1\leq j\leq B}$ uniformément parmi les suites de B éléments *distincts* de $\mathfrak{P}_{n_e}(\{1,\dots,n\})$.

– «Monte-Carlo CV» [25, MCCV] est très proche de RLT : les $(I_j^{(e)})_{1\leq j\leq B}$ sont i.i.d. de loi uniforme sur $\mathfrak{P}_{n_e}(\{1,\dots,n\})$. La seule différence avec RLT est donc que l'on autorise à choisir plusieurs fois le même découpage $I_j^{(e)}$. Ceci simplifie l'analyse théorique (grâce à l'indépendance des $I_j^{(e)}$), mais vraisemblablement sans changer grand chose en pratique.

Méthodes reliées. Mentionnons enfin plusieurs autres méthodes apparentées à la validation croisée, mais ne suivant pas exactement la définition (2) :

– des versions de VFCV ou RLT avec correction du biais [13, 14].
– la validation croisée généralisée [15, GCV], qui semble plus proche de C_L que des autres méthodes de validation croisée [18].
– APCV, une approximation analytique du LPO [27] pour le cas des modèles linéaires.
– LOO bootstrap [17], .632 bootstrap [17] et .632+ bootstrap [19].

1.3. **Estimation du risque par validation croisée.** Les deux caractéristiques principales d'un estimateur d'un réel tel que le risque sont son biais et sa variance[2]. Considérons les l'un après l'autre.

Biais. Voir [6, Section 5.1].

En utilisant l'indépendance de $D_n^{(e)}$ et $D_n^{(v)}$, on a

$$\begin{aligned}
\mathbb{E}\left[P_n^{(v)}\gamma\left(\widehat{s}_m(D_n^{(e)})\right)\right] &= \mathbb{E}\left[\frac{1}{n_v}\sum_{i\in D_n^{(v)}}\gamma\left(\widehat{s}_m(D_n^{(e)});\xi_i\right)\right] \\
&= \frac{1}{n_v}\sum_{i\in D_n^{(v)}}\mathbb{E}\left[\gamma\left(\widehat{s}_m(D_n^{(e)});\xi_i\right)\right] \\
&= \mathbb{E}_{\xi\sim P}\left[\gamma\left(\widehat{s}_m(D_n^{(e)});\xi\right)\right] \\
&= \mathbb{E}\left[P\gamma\left(\widehat{s}_m(D_n^{(e)})\right)\right] \\
&= \mathbb{E}\left[P\gamma\left(\widehat{s}_m(D_{n_e})\right)\right] \quad .
\end{aligned}$$

Ce calcul se généralise directement aux estimateurs par validation croisée dès lors que tous les échantillons d'entraînement ont la même taille n_e.

En particulier, si le risque s'écrit sous la forme

$$\mathbb{E}\left[P\gamma\left(\widehat{s}_m(D_n)\right)\right] = \alpha_m + \frac{\beta_m}{n} \quad , \tag{4}$$

comme dans le cas des régressogrammes par exemple (voir troisième cours), l'espérance d'un estimateur par validation croisée du risque vaut

$$\alpha_m + \frac{\beta_m}{n_e} = \alpha_m + \frac{n}{n_e}\frac{\beta_m}{n} \quad .$$

2. On fera bien attention ici à ne pas confondre les termes «biais» et «variance» avec la décomposition du risque en erreur d'approximation et erreur d'estimation. C'est pourquoi nous n'utilisons pas dans ce cours l'expression «compromis biais-variance» qui pourrait porter à confusion.

Le biais d'estimation du risque est donc positif (si $\beta_m > 0$), et consiste à surestimer l'erreur d'estimation β_m/n d'un facteur $n/n_e > 1$, d'autant plus grand que l'échantillon d'entraînement est petit. Ainsi, pour avoir un estimateur asymptotiquement sans biais du risque au premier ordre, il faut que $n_e \sim n$ quand $n \to +\infty$.

Variance. Voir [6, Section 5.2].

Pour le calcul de la variance de l'estimateur du risque par validation simple, on utilise le fait que pour toutes variables aléatoires X, Z,

$$\begin{aligned} \operatorname{var}(Z) &= \mathbb{E}\left[\operatorname{var}\left(Z \,|\, X\right)\right] + \operatorname{var}\left(\mathbb{E}\left[Z \,|\, X\right]\right) \quad , \\ \text{où} \quad \operatorname{var}\left(Z \,|\, X\right) &= \mathbb{E}\left[Z^2 \,\middle|\, X\right] + \left(\mathbb{E}\left[Z \,|\, X\right]\right)^2 \quad . \end{aligned}$$

En prenant $Z = P_n^{(v)}\gamma\left(\widehat{s}_m(D_n^{(e)})\right)$ et $X = D_n^{(e)}$, on obtient (en utilisant l'indépendance de $D_n^{(e)}$ et $D_n^{(v)}$) que

$$\begin{aligned} \operatorname{var}\left(P_n^{(v)}\gamma\left(\widehat{s}_m(D_n^{(e)})\right)\right) &= \mathbb{E}\left[\operatorname{var}\left(P_n^{(v)}\gamma\left(\widehat{s}_m(D_n^{(e)})\right) \,\middle|\, D_n^{(e)}\right)\right] \\ &\quad + \operatorname{var}\left(\mathbb{E}\left[P_n^{(v)}\gamma\left(\widehat{s}_m(D_n^{(e)})\right) \,\middle|\, D_n^{(e)}\right]\right) \\ &= \mathbb{E}\left[\frac{1}{n_v}\operatorname{var}_{\xi\sim P}\left(\gamma\left(\widehat{s}_m(D_n^{(e)});\xi\right) \,\middle|\, D_n^{(e)}\right)\right] \\ &\quad + \operatorname{var}\left(P\gamma\left(\widehat{s}_m(D_n^{(e)})\right)\right) \quad . \end{aligned}$$

En oubliant dans un premier temps que n_e et n_v sont reliées par la relation $n = n_e + n_v$, on constate que le premier terme de cette somme varie comme $1/n_v$ (avec un facteur multiplicatif dépendant de n_e, sans doute faiblement). Le deuxième terme de cette somme est quant à lui fonction de n_e uniquement, et dépend en particulier de la *stabilité* de $\mathcal{A}_m$: si $\widehat{s}_m(D_n^{(e)})$ dépend varie peu quand le sous-échantillon $D_n^{(e)}$ varie, alors la variance de la perte de $\widehat{s}_m(D_n^{(e)})$ sera faible.

2. Validation croisée pour la sélection d'estimateurs

On suppose désormais donnée une famille d'algorithmes statistiques $(\mathcal{A}_m)_{m\in\mathcal{M}_n}$, les estimateurs correspondants étant notés $(\widehat{s}_m(D_n))_{m\in\mathcal{M}_n}$. On cherche à sélectionner un $\widehat{m}(D_n) \in \mathcal{M}_n$ avec l'un des deux objectifs de la sélection d'estimateur : la prédiction ou l'identification (voir Section 3 du premier cours ou [6, Section 2]).

2.1. **Sélection d'estimateurs pour la prédiction.** Voir [6, Section 6].

Pour toute méthode de validation croisée, on considère dans cette section la procédure de sélection d'estimateurs associée

$$\widehat{m} \in \operatorname{argmin}_{m\in\mathcal{M}_n}\left\{\widehat{\mathcal{R}}^{\mathrm{vc}}\left(\mathcal{A}_m; D_n; \left(I_j^{(e)}\right)_{1\le j\le B}\right)\right\} \quad . \tag{5}$$

Dans plusieurs cadres où (4) est vérifiée (c'est-à-dire, l'erreur d'estimation varie comme n^{-1}), la performance de prédiction de la procédure définie par (5) est essentiellement fonction de la taille n_e de l'échantillon d'entraînement :

- Lorsque $n_e \sim n$, la validation croisée est asymptotiquement équivalente à C_p, donc asymptotiquement optimale.
- Lorsque $n_e \sim \lambda n$ avec $\lambda \in]0,1[$, la validation croisée est asymptotiquement équivalente à GIC_κ avec $\kappa = 1 + \lambda^{-1}$, qui est définie comme C_p avec une pénalité multipliée par $\kappa/2$. En d'autres termes, la validation croisée peut être vue comme surpénalisant d'un facteur $(1+\lambda)/(2\lambda) > 1$.

Les résultats ci-dessus ont été prouvés en régression linéaire pour le LOO [23], le LPO [28], et pour le RLT [37] en supposant $B \gg n^2$.

Plus généralement, pour la sélection parmi des estimateurs par minimum de contraste, des inégalités-oracle ont été montrées dans une série d'articles [30, 31, 32, 33], montrant que l'estimateur sélectionné par la validation croisée fait aussi bien (à une constante $C_n = 1 + o(1)$ près) que l'oracle avec n_e observations $m^\star(D_{n_e})$. Le plus souvent, ceci implique l'optimalité asymptotique de la validation croisée si $n_e/n = \mathcal{O}(1)$. Lorsque $n_e \sim \lambda n$ avec $\lambda \in]0,1[$, ceci généralise les résultats ci-dessus.

À propos de la sous-optimalité de la validation croisée «V-fold» dans le cas des régressogrammes, voir [2, Section 2].

2.2. **Sélection d'estimateurs pour l'identification.** Voir [6, Section 7], et en particulier ses deux références principales [35, 34].

2.3. **Choix d'une méthode de validation croisée.** Voir [6, Section 10].

2.4. **Limites de la validation croisée.** Voir [6, Section 8].

3. Détection de ruptures par validation croisée

À titre d'illustration des capacités naturelles d'adaptation de la validation croisée, nous montrons dans cette section comment l'utiliser pour un problème de détection de ruptures avec bruit hétéroscédastique. L'objectif est le suivant : on observe les valeurs successives d'un signal bruité $Y_i = \eta(x_i) + \varepsilon_i \in \mathbb{R}$ pour $i = 1, \ldots, n$ avec $x_1 < \cdots < x_n$. On suppose les variables ε_i indépendantes et centrées, mais pas forcément de même loi (cadre hétéroscédastique). On suppose de plus la fonction de régression $\eta : \mathcal{X} = [0,1] \mapsto \mathbb{R}$ constante par morceaux, et l'on cherche à détecter les «ruptures» de η.

En comparaison du problème classique de détection de ruptures (où l'on observe $Y_1, \ldots, Y_n$ indépendantes et de loi constante par morceaux, et l'on cherche à détecter les instants où la loi change), on se focalise ici sur la détection de *ruptures de la moyenne*. Il y a également une difficulté supplémentaire dans le fait que les données sont hétéroscédastiques et qu'on ne veut rien supposer sur les variations du niveau de bruit. En particulier, celui-ci n'est pas supposé constant sur chaque intervalle où η est constante.

Les résultats présentés dans cette section proviennent principalement de [5].

3.1. **Détection de ruptures et sélection de modèles.** La sélection de modèles a été employée avec succès comme un moyen de détecter les ruptures de la moyenne $\eta(x_i)$ du signal Y_i [22, 21]. L'idée est de se placer dans le cadre de la régression sur un plan d'expérience $x_1, \ldots, x_n$ déterministe, utiliser le contraste des moindres carrés, et considérer les régressogrammes $(\widehat{s}_m)_{m\in\mathcal{M}_n}$ associés à toutes les partitions de $\{x_1, \ldots, x_n\}$ en intervalles (on peut réduire $\mathcal{X}$ au plan d'expérience $\{x_1, \ldots, x_n\}$ puisque celui-ci est déterministe). Choisir un modèle S_m revient alors à choisir une partition m en intervalles, et donc un ensemble de ruptures. Il semble alors raisonnable de penser qu'une procédure de choix de modèles $\widehat{m}$ satisfaisant une inégalité-oracle (pour la perte des moindres carrés) est également une bonne procédure pour la détection de ruptures. Voir aussi [5, Section 2].

Un exemple de telle procédure a été étudié dans [22], où il est supposé que le signal est homoscédastique ($\sigma(x_i) = \sigma$ pour tout i). Elle consiste à utiliser une pénalité du type de celles proposées par Birgé et Massart [10], en tenant compte du fait que la collection $\mathcal{M}_n$ est de type exponentiel car elle contient 2^n modèles au total, et $\binom{n-1}{D-1} \approx \exp(cD\ln(n))$ modèles de dimension D :

$$\widehat{m} \in \operatorname{argmin}_{m\in\mathcal{M}_n} \left\{ P_n\gamma\left(\widehat{s}_m\right) + \frac{C\sigma^2 D_m}{n}\left(5 + 2\ln\left(\frac{n}{D_m}\right)\right)\right\} \qquad (6)$$

Il est intéressant de noter que les modèles de même dimension étant pénalisés de la même manière, ceci revient à aggréger les modèles de même dimension : si pour tout D on définit le modèle aggrégé

$$\widetilde{S}_D := \bigcup_{m\in\mathcal{M}_n\,,\,D_m=D} S_m$$

et

$$\widehat{s}_D \in \operatorname{argmin}_{t\in\widetilde{S}_D}\left\{P_n\gamma\left(t\right)\right\} \qquad (7)$$

l'estimateur des moindres carrés correspondant, alors

$$\widehat{s}_D = \widehat{s}_{\widehat{m}(D)} \quad \text{avec} \quad \widehat{m}(D) \in \operatorname{argmin}_{m\in\mathcal{M}_n\,,\,D_m=D}\left\{P_n\gamma\left(t\right)\right\}$$

et le modèle sélectionné par (6) coïncide avec $\widehat{m}(\widehat{D})$ où

$$\widehat{D} \in \operatorname{argmin}_{1\le D\le n}\left\{P_n\gamma\left(\widehat{s}_{\widehat{m}(D)}\right) + \frac{C\sigma^2 D}{n}\left(5 + 2\ln\left(\frac{n}{D}\right)\right)\right\} \ .$$

Notons au passage que le problème de minimisation (7) est a priori très coûteux algorithmiquement (si l'on explore tous les modèles S_m de dimension D), mais peut être résolu exactement de manière efficace par programmation dynamique [9, 26].

3.2. **Choix de modèle à D fixé.** Voir [5, Section 3].

3.3. **Choix du nombre de ruptures.**

En minimisant le risque empirique lorsque D est fixée. Voir [5, Section 4].

En utilisant la validation croisée lorsque D est fixée. Voir [5, Section 5]. Les méthodes alternatives mentionnées à l'oral proviennent des articles suivants :

- BM (pénalité «Birgé-Massart») : pénalités proposées dans un cadre général par [10] et utilisées en détection de ruptures par [22, 21], combinées ou non à l'heuristique de pente [11].
- BGH : pénalité multiplicative (voir Section 1.2 du deuxième cours) proposée par Baraud, Giraud et Huet [8].
- ZS : pénalité BIC modifiée proposée par Zhang et Siegmund [36].
- PML : maximum de (log-)vraisemblance pénalisée, avec un modèle de données indépendantes Gaussiennes et $(\eta(x_i), \sigma(x_i))$ est constant par morceaux, proposé dans le cadre de la détection de ruptures par [24] pour l'analyse de données de bioinformatique.

4. Pénalités V-fold

Nous avons vu deux approches principales pour la sélection d'estimateurs dans des situations complexes : les pénalités par rééchantillonnage (Section 5 du troisième cours) dont le coût de calcul peut être très élevé, et la validation croisée (Section 1) dont les versions algorithmiquement efficaces (validation croisée «V-fold») sont biaisées.

L'objectif de cette section est de montrer une façon d'éviter ces deux écueils, en combinant l'heuristique de rééchantillonnage et la procédure de sous-échantillonnage «V-fold» : c'est la pénalisation «V-fold» [2].

4.1. Principe. Définition.

Voir [2, Section 3.1.1].

En appliquant l'heuristique de rééchantillonnage avec des poids de sous-échantillonnage calqués sur la validation croisée «V-fold», on obtient l'estimateur suivant de la pénalité idéale

$$\mathrm{pen}_{\mathrm{id}}(\mathcal{A}_m; D_n) = \mathrm{pen}_{\mathrm{id}}(m; D_n) = (P - P_n)\gamma\left(\mathcal{A}_m(D_n)\right)$$

que l'on appelle *pénalité «V-fold»* [2] :

$$\mathrm{pen}_{\mathrm{VF}}(\mathcal{A}_m; D_n; C; \mathcal{B}) := \frac{C}{V}\sum_{j=1}^{V}\left[\left(P_n - P_n^{(-B_j)}\right)\left(\gamma\left(\widehat{s}_m^{(-j)}\right)\right)\right] \qquad (8)$$

en posant

$$\widehat{s}_m^{(-j)} = \mathcal{A}_m\left(D_n^{(-B_j)}\right) \ .$$

La constante multiplicative $C > 0$ est à choisir (voir Section 4.2), et la partition $\mathcal{B} = (B_j)_{1\leq j\leq V}$ de $\{1, \ldots, n\}$ est à prendre aussi régulière que possible. Idéalement, $\mathcal{B}$ doit vérifier :

$$\left.\begin{array}{r} (B_j)_{1\leq j\leq V} \text{ partition de } \{1, \ldots, n\} \\ \text{et} \quad \forall j \in \{1, \ldots, V\} \ , \qquad \mathrm{Card}(B_j) = \dfrac{n}{V} \end{array}\right\} \qquad \textbf{(RegPart)}$$

4.2. Espérances.

Pour le cas des régressogrammes, voir [2, Section 3.2, notamment Proposition 2]. Le cas général (qui repose sur la même technique de preuve) apparaîtra prochainement dans [4].

4.3. **Concentration.** Pour le cas des régressogrammes, voir [2, Proposition 11]. Le cas général (qui repose sur la même technique de preuve) apparaîtra prochainement dans [4].

4.4. **Inégalité oracle.** Pour le cas des régressogrammes, voir [2, Section 3.3, notamment le Théorème 2]. Pour le cas général, voir [4].

4.5. **Choix de V et comparaison à la validation croisée.** Voir [2, Sections 4–5]. Les résultats mentionnés en estimation de densité avec le contraste des moindres carrés apparaîtront prochainement dans [7].

5. Conclusion

La figure mentionnée comme réponse à la question «Quelle procédure de sélection de modèles pour quel problème ?» (dans un cadre particulier) provient de [3].

Références

[1] David M. Allen. The relationship between variable selection and data augmentation and a method for prediction. *Technometrics*, 16 :125–127, 1974.

[2] Sylvain Arlot. V-fold cross-validation improved : V-fold penalization, February 2008. arXiv :0802.0566v2.

[3] Sylvain Arlot. Choosing a penalty for model selection in heteroscedastic regression, June 2010. arXiv :0812.3141v2.

[4] Sylvain Arlot. V-fold cross-validation improved : V-fold penalization, 2011. Work in progress, to appear soon. arXiv :0802.0566v3.

[5] Sylvain Arlot and Alain Celisse. Segmentation of the mean of heteroscedastic data via cross-validation. *Statistics and Computing*, pages 1–20, 2010. 10.1007/s11222-010-9196-x.

[6] Sylvain Arlot and Alain Celisse. A survey of cross-validation procedures for model selection. *Statist. Surv.*, 4 :40–79, 2010.

[7] Sylvain Arlot and Matthieu Lerasle. V-fold penalization for least-squares density estimation, 2011. Work in progress, to appear soon.

[8] Yannick Baraud, Christophe Giraud, and Sylvie Huet. Gaussian model selection with an unknown variance. *Ann. Statist.*, 37(2) :630–672, 2009.

[9] Richard E. Bellman and Stuart E. Dreyfus. *Applied dynamic programming*. Princeton University Press, Princeton, N.J., 1962.

[10] Lucien Birgé and Pascal Massart. Gaussian model selection. *J. Eur. Math. Soc. (JEMS)*, 3(3) :203–268, 2001.

[11] Lucien Birgé and Pascal Massart. Minimal penalties for Gaussian model selection. *Probab. Theory Related Fields*, 138(1-2) :33–73, 2007.

[12] Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. *Classification and Regression Trees*. Wadsworth Statistics/Probability Series. Wadsworth Advanced Books and Software, Belmont, CA, 1984.

[13] Prabir Burman. A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods. *Biometrika*, 76(3) :503–514, 1989.

[14] Prabir Burman. Estimation of optimal transformations using v-fold cross validation and repeated learning-testing methods. *Sankhyā Ser. A*, 52(3) :314–345, 1990.

[15] Peter Craven and Grace Wahba. Smoothing noisy data with spline functions. Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numer. Math.*, 31(4) :377–403, 1978/79.

[16] Luc P. Devroye and Terry J. Wagner. Distribution-Free performance Bounds for Potential Function Rules. *IEEE Transaction in Information Theory*, 25(5) :601–604, 1979.

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

[18] Bradley Efron. How biased is the apparent error rate of a prediction rule ? *J. Amer. Statist. Assoc.*, 81(394) :461–470, 1986.

[19] Bradley Efron and Robert J. Tibshirani. Improvements on cross-validation : the .632+ bootstrap method. *J. Amer. Statist. Assoc.*, 92(438) :548–560, 1997.

[20] Seymour Geisser. The predictive sample reuse method with applications. *J. Amer. Statist. Assoc.*, 70 :320–328, 1975.

[21] Marc Lavielle. Using penalized contrasts for the change-point problem. *Signal Proces.*, 85(8) :1501–1510, 2005.

[22] Émilie Lebarbier. Detecting multiple change-points in the mean of a Gaussian process by model selection. *Signal Proces.*, 85 :717–736, 2005.

[23] Ker-Chau Li. Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation : discrete index set. *Ann. Statist.*, 15(3) :958–975, 1987.

[24] Franck Picard, Stéphane Robin, Marc Lavielle, Christian Vaisse, and Jean-Jacques Daudin. A statistical approach for array CGH data analysis. *BMC Bioinformatics*, 27(6) :electronic access, 2005.

[25] Richard R. Picard and R. Dennis Cook. Cross-validation of regression models. *J. Amer. Statist. Assoc.*, 79(387) :575–583, 1984.

[26] Guillem Rigaill. Pruned dynamic programming for optimal multiple change-point detection. arXiv :1004.0887, April 2010.

[27] Jun Shao. Linear model selection by cross-validation. *J. Amer. Statist. Assoc.*, 88(422) :486–494, 1993.

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

[29] Mervyn Stone. Cross-validatory choice and assessment of statistical predictions. *J. Roy. Statist. Soc. Ser. B*, 36 :111–147, 1974. With discussion by G. A. Barnard, A. C. Atkinson, L. K. Chan, A. P. Dawid, F. Downton, J. Dickey, A. G. Baker, O. Barndorff-Nielsen, D. R. Cox, S. Giesser, D. Hinkley, R. R. Hocking, and A. S. Young, and with a reply by the authors.

[30] Mark J. van der Laan and Sandrine Dudoit. Unified Cross-Validation Methodology For Selection Among Estimators and a General Cross-Validated Adaptive Epsilon-Net Estimator : Finite Sample Oracle Inequalities and Examples. Working Paper Series Working Paper 130, U.C. Berkeley Division of Biostatistics, November 2003. available at http ://www.bepress.com/ucbbiostat/paper130.

[31] Mark J. van der Laan, Sandrine Dudoit, and Sunduz Keles. Asymptotic optimality of likelihood-based cross-validation. *Stat. Appl. Genet. Mol. Biol.*, 3 :Art. 4, 27 pp. (electronic), 2004.

[32] Mark J. van der Laan, Sandrine Dudoit, and Aad W. van der Vaart. The cross-validated adaptive epsilon-net estimator. *Statist. Decisions*, 24(3) :373–395, 2006.

[33] Aad W. van der Vaart, Sandrine Dudoit, and Mark J. van der Laan. Oracle inequalities for multi-fold cross validation. *Statist. Decisions*, 24(3) :351–371, 2006.

[34] Yuhong Yang. Comparing learning methods for classification. *Statist. Sinica*, 16(2) :635–657, 2006.

[35] Yuhong Yang. Consistency of cross validation for comparing regression procedures. *Ann. Statist.*, 35(6) :2450–2473, 2007.

[36] N. R. Zhang and D. O. Siegmund. Modified Bayes Information Crietrion with Application to the Analysis of Comparative Genomic Hybridization Data. *Biometrics*, 63 :22–32, 2007.

[37] Ping Zhang. Model selection via multifold cross validation. *Ann. Statist.*, 21(1) :299–313, 1993.

Cnrs – Équipe Sierra, Laboratoire d'Informatique de l'École Normale Supérieure, (CNRS/ENS/INRIA UMR 8548), INRIA - 23 avenue d'Italie - CS 81321, 75214 PARIS Cedex 13 - France

E-mail address: sylvain.arlotRETIRERCECI@ens.fr

URL: http://www.di.ens.fr/~arlot/