Validation croisée

Sylvain Arlot

sylvain.arlot@universite-paris-saclay.fr

http://www.imo.universite-paris-saclay.fr/~arlot/

9 septembre 2025

Table des matières

Ce texte aborde deux questions fondamentales qui se posent en apprentissage supervisé.

D'une part, une fois que l'on a entraîné une règle d'apprentissage $\widehat{f}$ sur un jeu de données D_n, on dispose d'un prédicteur $\widehat{f}(D_n)$ mais d'aucune information sur sa qualité. Il est nécessairement imparfait, mais à quel point ? Avant de s'appuyer sur les prévisions qu'il fournit, il est indispensable d'évaluer son risque.

D'autre part, il est extrêmement rare que l'on ne puisse utiliser qu'une règle d'apprentissage pour un problème donné. Chaque méthode dépend en général d'un ou plusieurs paramètres (un modèle sous-jacent, un niveau de régularisation, un nombre de plus proches voisins, etc.) dont le choix impacte fortement le risque du prédicteur final. Et bien souvent, on dispose de nombreuses méthodes pour un problème donné (par exemple, en classification on peut hésiter entre des partitions cubiques, les k plus proches voisins, les SVM et les forêts aléatoires). Comment choisir la meilleur méthode et ses paramètres ?

La validation croisée fournit une réponse à ces deux questions, dans un cadre très général [1]. Ce texte vise à définir cette famille de procédures et à donner les principaux éléments de compréhension disponibles à ce jour à son sujet. Nous donnons volontairement peu de références bibliographiques ; une bibliographie complète se trouve dans l'article de survol de Arlot et Celisse (2010) et l'article récent de Arlot et Lerasle (2016).

1 Sélection d'estimateurs

Ce texte se place dans le cadre général de la prévision, tel que présenté par Arlot (2018a), dont on utilise les notations et auquel on fait régulièrement référence par la suite.

On peut alors formaliser le problème de sélection d'estimateurs [2] comme suit.

On dispose d'une collection de règles d'apprentissage $(\widehat{f}_m)_{m\in\mathcal{M}_n}$ et d'un échantillon D_n. On souhaite pouvoir choisir l'une de ces règles $\widehat{f}_{\widehat{m}}$ à l'aide des données uniquement. Cette question générale recouvre de nombreux exemples :

- sélection de modèles : pour tout $m \in \mathcal{M}_n$, $\widehat{f}_m$ est une règle par minimisation du risque empirique sur un modèle S_m.
- choix d'un hyperparamètre : m désigne alors un ou plusieurs paramètres réels dont dépend la règle $\widehat{f}_m$ (par exemple, le nombre de voisins k pour les k plus proches voisins, ou bien le paramètre de régularisation λ pour les SVM).
- choix entre des méthodes de natures diverses (par exemple, entre les k plus proches voisins, les SVM et les forêts aléatoires).

1. La validation croisée ne s'applique malheureusement pas systématiquement. La parenthèse 1 en section 2.1 précise des conditions suffisantes pour sa mise en œuvre.

2. Compte-tenu de la terminologie introduite par Arlot (2018a), il serait plus logique de parler de sélection de règles d'apprentissage. Nous utilisons néanmoins ici l'expression « sélection d'estimateurs », plus courante et plus concise.

Les enjeux du problème et approches pour le résoudre sont essentiellement les mêmes que pour le problème de sélection de modèles, que Arlot (2018a, section 3.9) décrit en détail. Nous n'en rappelons donc ici que les grandes lignes.

Tout d'abord, il faut préciser l'objectif (prévision ou identification de la « meilleure » règle $\widehat{f}_m$). Ce texte se focalise sur l'objectif de prévision : on veut minimiser le risque de l'estimateur final $\widehat{f}_{\widehat{m}(D_n)}(D_n)$ — entraîné sur l'ensemble des données, celles-là mêmes qui ont servi à choisir $\widehat{m}(D_n)$.

Atteindre un tel objectif nécessite d'éviter deux défauts principaux : le surapprentissage (lorsqu'un prédicteur « colle » excessivement aux observations, ce qui l'empêche de généraliser correctement) et le sous-apprentissage (quand un prédicteur « lisse » trop les observations et devient incapable de reproduire les variations du prédicteur de Bayes). Il s'agit donc de trouver le meilleur compromis entre ces deux extrêmes. Dans de nombreux cas [3], ceci se formalise sous la forme d'un compromis biais-variance ou approximation-estimation.

Comment faire ? Comme pour la sélection de modèles, on considère habituellement des procédures de la forme :

$$\widehat{m} \in \operatorname*{argmin}_{m \in \mathcal{M}_n} \left\{ \operatorname{crit}(m; D_n) \right\} . \tag{1}$$

On peut les analyser avec le lemme fondamental de l'apprentissage (Arlot, 2018a, lemme 2). Deux stratégies principales sont possibles : choisir un critère $\operatorname{crit}(m; D_n)$ proche du risque $\mathcal{R}_P(\widehat{f}_m(D_n))$ simultanément pour tous les $m \in \mathcal{M}_n$, ou bien choisir un critère qui majore le risque. La validation croisée suit la première stratégie. Alors, au vu du raisonnement exposé par Arlot (2018a, section 3.9), il suffit [4] de démontrer que $\operatorname{crit}(m; D_n)$ est un bon estimateur du risque [5] de $\widehat{f}_m(D_n)$ pour en déduire que la procédure définie par (1) fonctionne bien.

C'est pourquoi, après avoir défini les procédures de validation croisée (section 2), nous commençons par étudier leurs propriétés pour l'estimation du risque d'une règle d'apprentissage fixée (section 3), avant d'aborder la sélection d'estimateurs (section 4). En guise de conclusion, la section 5 considère plusieurs questions pratiques importantes, dont celle du choix de la meilleure procédure de validation croisée.

2 Définition

Étant donné une règle d'apprentissage $\widehat{f}_m$, un échantillon D_n et une fonction de coût c, la validation croisée estime le risque $\mathcal{R}_P(\widehat{f}_m(D_n))$ en se fondant sur le principe

3. Par exemple, pour des estimateurs linéraires en régression, des règles par minimisation du risque empirique ou des règles par moyenne locale.

4. Il n'y a cependant pas équivalence, voir la section 4.

5. En toute rigueur, il faut parler d'estimation du risque *moyen*, le risque étant une quantité aléatoire. Par abus de langage, on parle dans ce texte d'estimation (et d'estimateurs) du risque de $\widehat{f}_m$, et du biais et de la variance de ces estimateurs. À chaque fois, il est sous-entendu que c'est le risque moyen qu'on estime, même si c'est le risque que l'on souhaite évaluer aussi précisément que possible.

suivant : on découpe l'échantillon D_n en deux sous-échantillons $D_n^{(e)}$ (l'échantillon d'entraînement) et $D_n^{(v)}$ (l'échantillon de validation), on utilise $D_n^{(e)}$ pour entraîner un prédicteur $\widehat{f}_m(D_n^{(e)})$, puis on mesure l'erreur commise par ce prédicteur sur les données restantes $D_n^{(v)}$. Alors, du fait de l'indépendance entre $D_n^{(e)}$ et $D_n^{(v)}$, on obtient une bonne évaluation [6] du risque de $\widehat{f}_m(D_n)$. En particulier, on évite l'optimisme excessif [7] du risque empirique $\widehat{\mathcal{R}}_n(\widehat{f}_m(D_n))$. Et l'on peut procéder à un ou plusieurs découpages du même échantillon, d'où un grand nombre de procédures de validation croisée possibles.

2.1 Cas général

Dans tout ce texte, $D_n = (X_i, Y_i)_{1 \leqslant i \leqslant n}$ désigne un échantillon de variables aléatoires indépendantes et de même loi P. On suppose qu'une fonction de coût $c : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$ est fixée et sert à définir le risque $\mathcal{R}_P$ et le risque empirique $\widehat{\mathcal{R}}_n$ sur D_n (une quantité définie par Arlot (2018a, section 3.1)).

Un sous-ensemble strict E de $\{1, \dots, n\}$, non-vide, est appelé « découpage » [8] de l'échantillon. Il correspond à la partition de D_n en deux sous-échantillons :

$$D_n^E := (X_i, Y_i)_{i \in E} \qquad \text{et} \qquad D_n^{E^c} := (X_i, Y_i)_{i \in \{1,\dots,n\} \setminus E} .$$

Pour tout découpage E de l'échantillon, on définit le risque empirique sur le sous-échantillon D_n^E par :

$$\widehat{\mathcal{R}}_n^E : f \in \mathcal{F} \mapsto \frac{1}{\mathrm{Card}(E)} \sum_{i \in E} c\big(f(X_i), Y_i\big) .$$

On peut maintenant définir formellement les estimateurs par validation croisée du risque.

Définition 1 (Validation croisée) Soit $\widehat{f}_m$ une règle d'apprentissage. L'estimateur par validation (simple) [9] du risque de $\widehat{f}_m$ pour l'échantillon D_n et le découpage E est défini par :

$$\begin{aligned} \widehat{\mathcal{R}}^{\mathrm{val}}(\widehat{f}_m; D_n; E) &= \widehat{\mathcal{R}}_n^{E^c}\big(\widehat{f}_m(D_n^E)\big) \\ &= \frac{1}{\mathrm{Card}(E^c)} \sum_{i \in E^c} c\big(\widehat{f}_m(D_n^E; X_i), Y_i\big) . \end{aligned}$$

On appelle D_n^E l'échantillon d'entraînement [10], tandis que $D_n^{E^c}$ est appelé échantillon de validation.

6. Comme expliqué en section 3, c'est en réalité le risque de $\widehat{f}_m(D_n^{(e)})$ que l'on évalue, d'où un léger biais (beaucoup moins problématique que celui du risque empirique).

7. Les raisons de cet optimisme sont détaillées par Arlot (2018a, section 3.9).

8. La terminologie « découpage de l'échantillon » n'est pas classique ; nous l'utilisons ici pour clarifier l'exposition. On devrait plutôt parler de sous-ensemble *propre* de $\{1, \dots, n\}$.

9. Le terme anglais pour la validation, ou validation simple, est « hold-out » : il s'agit de l'erreur sur des données « mises de côté » au moment de l'entraînement.

10. Le terme « échantillon d'apprentissage » est parfois utilisé pour désigner l'échantillon d'entraînement ; il arrive aussi qu'on l'utilise pour désigner la réunion de l'échantillon d'entraînement et de l'échantillon de validation, lorsqu'une partie des données est mise de côté dans un échantillon test.

L'estimateur par validation croisée[11] du risque de $\widehat{f}_m$ pour l'échantillon D_n et la suite de découpages $(E_j)_{1\leqslant j\leqslant V}$ est défini par :

$$\widehat{\mathcal{R}}^{\mathrm{vc}}\big(\widehat{f}_m; D_n; (E_j)_{1\leqslant j\leqslant V}\big) = \frac{1}{V}\sum_{j=1}^{V} \widehat{\mathcal{R}}^{\mathrm{val}}(\widehat{f}_m; D_n; E_j)\,.$$

Étant donné une famille de règles d'apprentissage $(\widehat{f}_m)_{m\in\mathcal{M}_n}$, la procédure de sélection d'estimateurs par validation croisée associée est définie par :

$$\widehat{m}^{\mathrm{vc}}\big(D_n; (E_j)_{1\leqslant j\leqslant V}\big) \in \operatorname*{argmin}_{m\in\mathcal{M}_n}\Big\{\widehat{\mathcal{R}}^{\mathrm{vc}}\big(\widehat{f}_m; D_n; (E_j)_{1\leqslant j\leqslant V}\big)\Big\}\,.$$

Une erreur courante mais grave est d'utiliser

$$\widehat{\mathcal{R}}^{\mathrm{vc}}\big(\widehat{f}_m; D_n; (E_j)_{1\leqslant j\leqslant V}\big)$$

pour estimer le risque de $\widehat{f}_m(D_n)$ lorsque la règle d'apprentissage $\widehat{f}_m$ dépend déjà elle-même des données. Par exemple, si $\widehat{f}_m$ est construite sur un sous-ensemble m des covariables disponibles, et si ce sous-ensemble a été choisi à l'aide d'une partie des données D_n (par une procédure automatisée ou simplement « à l'œil »), alors on obtient une estimation *fortement biaisée* du risque ! En général, cette estimation est très optimiste, conduisant à sous-estimer le risque de prévision réel.

Pour éviter ce biais, il faut prendre en compte la *totalité du processus* menant des données D_n au prédicteur $\widehat{f}_m(D_n)$ (c'est-à-dire, *tout* ce qu'on a fait à partir du moment où l'on a eu accès à au moins une observation). Formellement, il faut décrire comment m dépend des données, et le noter $\widehat{m}(D_n)$. On définit alors $\widetilde{f} : D_n \mapsto \widehat{f}_{\widehat{m}(D_n)}(D_n)$ puis on applique la validation croisée à $\widetilde{f}$ en calculant :

$$\widehat{\mathcal{R}}^{\mathrm{vc}}\big(\widetilde{f}; D_n; (E_j)_{1\leqslant j\leqslant V}\big)\,.$$

Le même problème se pose quand on veut estimer le risque de l'estimateur sélectionné par la validation croisée (ou par toute autre procédure de sélection d'estimateurs). Si l'on utilise la valeur (calculée en cours de procédure)

$$\widehat{\mathcal{R}}^{\mathrm{vc}}\big(\widehat{f}_m; D_n; (E_j)_{1\leqslant j\leqslant V}\big) \qquad \text{pour} \qquad m = \widehat{m}^{\mathrm{vc}}\big(D_n; (E_j)_{1\leqslant j\leqslant V}\big)\,,$$

qui est égale à

$$\min_{m\in\mathcal{M}_n} \widehat{\mathcal{R}}^{\mathrm{vc}}\big(\widehat{f}_m; D_n; (E_j)_{1\leqslant j\leqslant V}\big)\,,$$

alors on commet précisément l'erreur mentionnée ci-dessus, et l'on sous-estime fortement le risque. Il faut donc définir

$$\widetilde{f}^{\mathrm{vc}} : D_n \mapsto \widehat{f}_{\widehat{m}^{\mathrm{vc}}(D_n; (E_{j,n})_{1\leqslant j\leqslant V_n})}$$

11. En anglais : « cross-validation ».

(en spécifiant bien comment la suite de découpages $(E_{j,n})_{1\leqslant j\leqslant V_n}$ est choisie pour chaque entier $n \geqslant 1$) et lui appliquer la validation croisée en calculant

$$\widehat{\mathcal{R}}^{\mathrm{vc}}\big(\widetilde{f}^{\mathrm{vc}}; D_n; (E_j)_{1\leqslant j\leqslant V}\big)\,.$$

Dans le cas de la validation simple, ceci conduit à un découpage de l'échantillon en *trois sous-échantillons* : un échantillon d'entraînement D_n^E, un échantillon de validation D_n^V — pour choisir parmi les $\widehat{f}_m(D_n^E)$ —, et un échantillon test D_n^T — pour évaluer le risque de l'estimateur final $\widehat{f}_{\widehat{m}(D_n^E, D_n^V)}(D_n^E)$ —, où E, V et T forment une partition de $\{1, \ldots, n\}$.

Signalons toutefois que d'autres approches permettent d'éviter la nécessité de recourir à un découpage de l'échantillon en trois, en particulier le « reusable hold-out » récemment proposé par Dwork *et al.* (2015), qui repose sur l'idée de n'accéder à l'échantillon de validation que par l'intermédiaire d'un mécanisme de confidentialité différentielle [12].

Parenthèse 1 (Cadre plus général) On peut définir la validation croisée hors du cadre de prévision ou pour une fonction de risque plus générale que celle définie par Arlot (2018a). Il suffit en effet que le risque d'un élément f de l'ensemble $\mathcal{F}$ des sorties possibles d'une règle d'apprentissage vérifie :

$$\forall n \geqslant 1, \qquad \mathcal{R}_P(f) = \mathbb{E}\big[\mathcal{E}(f; D_n)\big] \tag{2}$$

où D_n est un échantillon de n variables aléatoires indépendantes et de loi P, et $\mathcal{E}$ est une fonction à valeurs réelles, prenant en entrée un élément de $\mathcal{F}$ et un échantillon de taille quelconque. La fonction $\mathcal{E}$ mesure l'« adéquation » entre f et l'échantillon D_n. L'estimateur par validation simple se définit alors par :

$$\widehat{\mathcal{R}}^{\mathrm{val}}(\widehat{f}_m; D_n; E) := \mathcal{E}\Big(\widehat{f}_m\big(D_n^E\big); D_n^{E^c}\Big)$$

et l'estimateur par validation croisée s'en déduit. Dans le cas de la prévision, (2) est vérifiée avec :

$$\mathcal{E}\big(f; (X_i, Y_i)_{1\leqslant i\leqslant n}\big) = \frac{1}{n}\sum_{i=1}^{n} c\big(f(X_i), Y_i\big) = \widehat{\mathcal{R}}_n(f)\,.$$

La relation (2) est également vérifiée dans d'autres cadres. Par exemple, en estimation de densité, il est classique de considérer le risque quadratique défini par (2) avec :

$$\mathcal{E}\big(f; (X_i)_{1\leqslant i\leqslant n}\big) = \|f\|_{L^2}^2 - \frac{2}{n}\sum_{i=1}^{n} f(X_i)\,.$$

L'excès de risque correspondant est $\|f - f^\star\|_{L^2}^2$ où $f^\star$ est la densité (inconnue) des observations (Arlot et Lerasle, 2016). On peut aussi obtenir la distance de Kullback-Leibler entre f et $f^\star$ comme excès de risque en définissant le risque par

12. Le terme anglais est « differential privacy ».

(2) avec :

$$\mathcal{E}\big(f;(X_i)_{1\leqslant i\leqslant n}\big) = -\frac{1}{n}\sum_{i=1}^{n}\ln\big(f(X_i)\big)\,,$$

qui est l'opposé de la log-vraisemblance de f au vu de l'échantillon $(X_i)_{1\leqslant i\leqslant n}$.

2.2 Exemples

Comme l'indique la définition 1, il y a autant de procédures de validation croisée que de suites de découpages $(E_j)_{1\leqslant j\leqslant V}$. Au sein de cette grande famille, certaines procédures sont plus classiques que d'autres.

La plupart des procédures utilisées vérifient les deux hypothèses suivantes :

$$(E_j)_{1\leqslant j\leqslant V} \text{ est indépendante de } D_n \qquad \textbf{(Ind)}$$

$$\mathrm{Card}(E_1) = \mathrm{Card}(E_2) = \cdots = \mathrm{Card}(E_V) = n_e \in \{1,\ldots,n-1\}\,. \qquad \textbf{(Reg)}$$

On suppose toujours dans ce texte que **(Ind)** et **(Reg)** sont vérifiées.

Parenthèse 2 (Sur l'hypothèse (Ind)) L'hypothèse **(Ind)** garantit que l'échantillon d'entraînement $D_n^{E_j}$ et l'échantillon de validation $D_n^{E_j^c}$ sont indépendants pour tout j, ce qui est crucial pour l'analyse menée en section 3.1 notamment. Cependant, il est parfois suggéré de choisir $(E_j)_{1\leqslant j\leqslant V}$ en utilisant l'échantillon D_n pour diverses raisons :

— pour que l'ensemble du support de P_X soit représenté dans chaque échantillon d'entraînement et chaque échantillon de validation.

— en classification, pour que toutes les classes soient représentées dans chaque échantillon d'entraînement et chaque échantillon de validation (en particulier lorsque les effectifs des classes sont très déséquilibrés, ou lorsque le nombre de classes est grand).

Nous ne connaissons cependant pas de résultat théorique justifiant l'intérêt d'un tel choix. En régression, dans les simulations de Breiman et Spector (1992, section 6.2), une telle stratégie n'a pas d'impact sur les performances.

Parenthèse 3 (Échantillon ordonné et hypothèse (Ind)) Souvent, en pratique, on dispose d'un échantillon « ordonné ». Par exemple, lorsque $\mathcal{X} = \mathbb{R}$, on a souvent $X_1 \leqslant X_2 \leqslant \cdots \leqslant X_n$ (ce qui signifie, si D_n contient bien les réalisations de n variables aléatoires indépendantes de loi P, que l'échantillon initial a été réordonné). Alors, si l'on veut utiliser une procédure de validation croisée vérifiant l'hypothèse **(Ind)**, il faut prendre soin d'appliquer au préalable une permutation aléatoire uniforme des indices $\{1,\ldots,n\}$. Sinon, le découpage $E = \{1,\ldots,n/2\}$ ne peut pas être considéré indépendant de D_n. Notons toutefois que ceci n'est pas nécessaire pour les procédures « leave-one-out » et « leave-p-out ».

Parenthèse 4 (Sur l'hypothèse (Reg)) Nous ne connaissons pas d'argument théorique en faveur des procédures de validation croisée vérifiant (**Reg**), si ce n'est qu'elles sont plus simples à analyser que les autres. Il n'empêche que l'hypothèse (**Reg**) est toujours vérifiée (au moins approximativement) dans les applications. Les résultats théoriques mentionnés dans ce texte restent valables (au premier ordre) lorsque (**Reg**) n'est vérifiée qu'approximativement, cas inévitable si l'on utilise la validation croisée « V-fold » avec n non-divisible par V.

Parmi les procédures vérifiant les hypothèses (**Ind**) et (**Reg**), on peut distinguer deux approches. Soit l'on considère la suite de *tous* les découpages E_j tels que $\mathrm{Card}(E_j) = n_e$. Lorsque $n_e = n - 1$, on obtient le leave-one-out [13] :

$$\begin{aligned}\widehat{\mathcal{R}}^{\mathrm{loo}}(\widehat{f}_m; D_n) :=& \widehat{\mathcal{R}}^{\mathrm{vc}}\Big(\widehat{f}_m; D_n; \big(\{j\}^c\big)_{1\leqslant j\leqslant n}\Big)\\ =& \frac{1}{n}\sum_{j=1}^{n} c\big(\widehat{f}_m(D_n^{(-j)}; X_j), Y_j\big)\end{aligned}$$

où $D_n^{(-j)} = (X_i, Y_i)_{1\leqslant i\leqslant n,\, i\neq j}$. Dans le cas général, en posant $p = n - n_e$, on obtient le leave-p-out [14] :

$$\widehat{\mathcal{R}}^{\mathrm{lpo}}(\widehat{f}_m; D_n; p) := \widehat{\mathcal{R}}^{\mathrm{vc}}\Big(\widehat{f}_m; D_n; \big\{E \subset \{1,\dots,n\} \,/\, \mathrm{Card}(E) = n - p\big\}\Big).$$

En pratique, il est souvent trop coûteux algorithmiquement (voire impossible) d'utiliser le leave-one-out ou le leave-p-out. Une deuxième approche est donc nécessaire : n'explorer que partiellement l'ensemble des $\binom{n}{n_e}$ découpages possibles avec un échantillon d'entraînement de taille n_e.

Considérer un seul découpage amène à la validation simple ou « hold-out » ; toutes ces procédures sont équivalentes car on a fait l'hypothèse (**Ind**). En revanche, dès que l'on considère un nombre de découpages $V \in \big[2, \binom{n}{n_e}\big[$, plusieurs procédures non-équivalentes sont possibles.

La plus classique est la validation croisée par blocs, appelée validation croisée « V-fold » ou « k-fold ». On se donne une partition $(B_j)_{1\leqslant j\leqslant V}$ de $\{1,\dots,n\}$ en V blocs de même taille [15], puis on procède à un « leave-one-out par blocs », c'est-à-dire, on utilise

13. En français, la procédure leave-one-out peut être nommée validation croisée « tous sauf un » : chaque découpage laisse exactement une observation hors de l'échantillon d'entraînement. En anglais, on trouve aussi les noms « delete-one cross-validation », « ordinary cross-validation », et même parfois simplement « cross-validation ». Dans le cas où $\widehat{f}_m$ est un estimateur des moindres carrés en régression linéaire (Arlot, 2018a, exemple 4 en section 3.2), le leave-one-out est parfois appelé PRESS (Prediction Sum of Squares), ou PRESS de Allen ; notons que ce terme désigne parfois directement le critère défini par la formule simplifiée (3) que l'on peut démontrer dans ce cadre.

14. En anglais, on trouve aussi les termes « delete-p cross-validation » et « delete-p multifold cross-validation ». En français, on peut l'appeler validation croisée « tous sauf p ».

15. Il faut prendre les B_j de même taille pour avoir l'hypothèse (**Reg**). Lorsque V ne divise pas n, il suffit de les prendre de tailles égales à un élément près ; les performances théoriques et pratiques sont alors similaires à ce que l'on a quand (**Reg**) est vérifiée exactement.

la suite de découpages $(B_j^c)_{1\leqslant j\leqslant V}$:

$$\begin{aligned}\widehat{\mathcal{R}}^{\mathrm{vf}}\big(\widehat{f}_m; D_n; (B_j)_{1\leqslant j\leqslant V}\big) :=& \widehat{\mathcal{R}}^{\mathrm{vc}}\big(\widehat{f}_m; D_n; (B_j^c)_{1\leqslant j\leqslant V}\big)\\ =& \frac{1}{V}\sum_{j=1}^{V}\widehat{\mathcal{R}}_n^{B_j}\left(\widehat{f}_m\left(D_n^{B_j^c}\right)\right).\end{aligned}$$

On peut également procéder à une validation croisée Monte-Carlo (ou validation croisée répétée), en choisissant $E_1, \ldots, E_V$ aléatoires, indépendants et de loi uniforme sur l'ensemble des parties de taille n_e de $\{1, \ldots, n\}$.

Remarque 5 (V-fold ou Monte-Carlo ?) On discute dans la suite les mérites de ces deux approches. Intuitivement, on peut déjà dire que la validation croisée V-fold présente l'avantage de faire un usage « équilibré » des données : chaque observation est utilisée exactement $V-1$ fois pour l'entraînement et une fois pour l'apprentissage. Ce n'est en rien garanti avec la validation croisée Monte-Carlo. En revanche, on peut s'interroger sur les inconvénients de toujours utiliser ensemble (soit pour l'entraînement, soit pour la validation) les observations d'un même bloc. L'approche « Monte-Carlo », par son caractère aléatoire, permet d'éviter d'éventuels biais induits par ce lien entre observations. Notons qu'il existe une manière d'éviter ces deux écueils : la validation croisée incomplète équilibrée [16] (Arlot et Celisse, 2010, section 4.3.2), qui présente l'inconvénient de n'être possible que pour d'assez grandes valeurs de V.

Signalons enfin que bien d'autres procédures de validation croisée « non-exhaustives » existent. Par exemple, avec la validation croisée V-fold répétée, on choisit plusieurs partitions $(B_j^\ell)_{1\leqslant j\leqslant V}$, $\ell \in \{1, \ldots, L\}$, et l'on explore l'ensemble des découpages

$$\left\{(B_j^\ell)^c \,/\, j \in \{1, \ldots, V\} \text{ et } \ell \in \{1, \ldots, L\}\right\}.$$

2.3 Astuces algorithmiques

La complexité algorithmique du calcul « naïf » de

$$\widehat{\mathcal{R}}^{\mathrm{vc}}\big(\widehat{f}_m; D_n; (E_j)_{1\leqslant j\leqslant V}\big) = \frac{1}{V}\sum_{j=1}^{V}\widehat{\mathcal{R}}_n^{E_j^c}\big(\widehat{f}_m(D_n^{E_j})\big)$$

est de l'ordre de V fois celle de l'entraînement de $\widehat{f}_m$ sur un échantillon de taille n_e (c'est en général le plus coûteux), plus l'évaluation de $\widehat{f}_m(D_n^{E_j})$ en $n - n_e$ points. Il est cependant parfois possible de faire beaucoup plus rapide.

Tout d'abord, on dispose dans certains cas de formules closes pour l'estimateur par validation croisée du risque de $\widehat{f}_m$, au moins pour le leave-one-out et le leave-p-out [17]

16. En anglais, on parle de « balanced-incomplete cross-validation », qui s'appuie sur la notion de « balanced-incomplete block-design ».

17. Au vu des résultats des sections 3 et 4, en particulier la proposition 2 en section 3.3 qui montre que la variance de la validation croisée est minimale pour le leave-p-out — à n_e fixée —, il semble inutile de considérer un autre type de validation croisée quand on sait calculer rapidement tous les estimateurs par leave-p-out.

(Arlot et Celisse, 2010, section 9). Par exemple, si $\widehat{f}_m$ est un estimateur des moindres carrés en régression linéaire, la formule de Woodbury (Press *et al.*, 1992, section 2.7) permet de démontrer (Golub *et al.*, 1979) :

$$\begin{aligned}\widehat{\mathcal{R}}^{\mathrm{loo}}\big(\widehat{f}_m; (x_i, y_i)_{1\leqslant i\leqslant n}\big) &= \frac{1}{n}\sum_{i=1}^{n} \frac{\Big[y_i - \widehat{f}_m\big((x_i, y_i)_{1\leqslant i\leqslant n}; x_i\big)\Big]^2}{1 - \mathbf{H}_{i,i}} \\ &= \frac{1}{n}\sum_{i=1}^{n} \frac{\big(y_i - (\mathbf{H}\mathbf{y})_i\big)^2}{1 - \mathbf{H}_{i,i}} \qquad (3) \\ \text{où} \qquad \mathbf{H} = \mathbf{X}\big(\mathbf{X}^\top\mathbf{X}\big)^{-1}\mathbf{X}^\top \qquad &\text{et} \qquad \mathbf{y} = (y_i)_{1\leqslant i\leqslant n} \in \mathbb{R}^n\,.\end{aligned}$$

Le calcul de l'estimateur leave-one-out avec (3) est aussi coûteux que celui d'*un seul* entraînement de $\widehat{f}_m$ sur D_n, via le calcul de la matrice $\mathbf{H}$.

Parenthèse 6 (Validation croisée généralisée) La formule close (3) obtenue pour le leave-one-out en régression linéaire a conduit à en définir une version « invariante par rotation » (Golub *et al.*, 1979), appelée validation croisée généralisée ou GCV (de l'anglais « generalized cross-validation »). Par rapport à la formule (3), les dénominateurs $1 - \mathbf{H}_{i,i}$ sont remplacés par $1 - n^{-1}\mathrm{trace}(\mathbf{H})$:

$$\mathrm{GCV}(\widehat{f}_m; D_n) = \frac{\|\mathbf{y} - \mathbf{H}\mathbf{y}\|^2}{n - \mathrm{trace}(\mathbf{H})}\,.$$

Ce critère s'applique, plus généralement, à tout estimateur « linéaire » en régression avec le coût quadratique, notamment les k plus proches voisins (Arlot, 2018a, section 5.4), les estimateurs par noyau (Arlot, 2018a, section 5.5) et les estimateurs ridge. Efron (1986) explique pourquoi, malgré son nom, GCV est beaucoup plus proche des critères C_p et C_L de Mallows que de la validation croisée proprement dite.

Par ailleurs, même en l'absence de formule close, on peut réduire la complexité algorithmique de la validation croisée en entraînant d'abord $\widehat{f}_m$ sur l'échantillon D_n tout entier (une fois pour toutes), puis, pour chaque découpage E_j, en « mettant à jour » $\widehat{f}_m(D_n)$ afin d'obtenir $\widehat{f}_m(D_n^{E_j})$. Lorsque cette mise à jour est efficace, le gain algorithmique est important. Cette idée s'applique dans plusieurs cadres, dont l'analyse discriminante (linéaire ou quadratique) et les k plus proches voisins (Arlot et Celisse, 2010, section 9). Considérons ici à nouveau l'estimateur des moindres carrés en régression linéaire. Son calcul nécessite d'inverser la matrice $\mathbf{X}^\top\mathbf{X}$ de taille p, ce qui a un coût de l'ordre de p^3. Lorsque $n - n_e$ est significativement plus petit que n, on peut utiliser la formule de Woodbury (Press *et al.*, 1992, section 2.7) pour déduire $(\mathbf{X}_{(E_j)}^\top\mathbf{X}_{(E_j)})^{-1}$ de $(\mathbf{X}^\top\mathbf{X})^{-1}$ à moindres frais,

$$\text{où} \qquad \mathbf{X}_{(E_j)} = (x_{i,k})_{1\leqslant i\leqslant n, k\in E_j}\,.$$

La formule de Woodbury est également utile en régression ridge, où l'essentiel du temps de calcul de l'estimateur est consacré à l'inversion de la même matrice $\mathbf{X}^\top\mathbf{X}$.

Enfin, dans un contexte de sélection d'estimateurs, il n'est pas toujours nécessaire de calculer

$$\widehat{\mathcal{R}}^{\mathrm{vc}}\big(\widehat{f}_m; D_n; (E_j)_{1\leqslant j\leqslant V}\big)$$

pour chaque $m \in \mathcal{M}_n$. Les valeurs de l'erreur de validation obtenues sur les premiers découpages E_1, E_2, E_3, ... peuvent suffire à éliminer certains $\widehat{f}_m$ (et donc de gagner en temps de calcul), sans trop perdre sur la qualité du prédicteur final (Krueger *et al.*, 2015).

2.4 Variantes

Pour le problème de sélection d'estimateurs, il y a plusieurs variantes de la validation croisée, qui ne suivent pas la définition 1 mais reposent sur le même principe d'entraînement et validation selon plusieurs découpages successifs.

Yang (2006, 2007) propose la « validation croisée par vote »[18], lorsque l'objectif est d'*identifier* la meilleure règle d'apprentissage (comme expliqué par Arlot (2018a, section 3.9) ; voir aussi la remarque 14 en section 5.1). Pour chacun des $j \in \{1, \ldots, V\}$, on sélectionne un estimateur par validation simple :

$$\widehat{m}_j \in \operatorname*{argmin}_{m\in\mathcal{M}_n} \big\{\widehat{\mathcal{R}}^{\mathrm{val}}(\widehat{f}_m; D_n; E_j)\big\}\,. \tag{4}$$

Ensuite, on réalise un vote majoritaire parmi les $\widehat{m}_j$ pour déterminer $\widehat{m}$. Clairement, ceci n'a de sens que lorsque $\mathcal{M}_n$ est discret. Yang propose cette variante dans un contexte où $\mathcal{M}_n$ est fini et de petite taille. Supposons par exemple que l'on veut choisir, pour un problème de classification, entre les plus proches voisins, la régression logistique et les forêts aléatoires. Si les paramètres de chaque méthode sont choisis par une boucle interne de validation croisée, on a $\mathrm{Card}(\mathcal{M}_n) = 3$. Cela fait donc sens d'effectuer un vote majoritaire parmi les $\widehat{m}_j$ obtenus sur $V \gg 3$ découpages différents.

La validation croisée agrégée[19] est une variante largement utilisée en pratique pour ses bonnes performances en prévision, mais peu mentionnée dans la littérature (Jung et Hu, 2015; Maillard *et al.*, 2017). L'idée est de ne pas *sélectionner* l'un des $\widehat{f}_m$ mais d'en *combiner* plusieurs pour obtenir un prédicteur encore plus performant (parfois meilleur que le choix oracle, d'après des résultats expérimentaux). Comme pour la validation croisée par vote, pour chaque découpage E_j, $j \in \{1, \ldots, V\}$, on sélectionne $\widehat{m}_j$ défini par (4). Ensuite, on construit un prédicteur $\widetilde{f}(D_n)$ en agrégeant les prédicteurs $\widehat{f}_{\widehat{m}_j}(D_n^{E_j})$ obtenus avec chaque découpage. En régression, on fait une moyenne :

$$\widetilde{f}(D_n) = \frac{1}{V}\sum_{j=1}^{V} \widehat{f}_{\widehat{m}_j}\big(D_n^{E_j}\big)\,.$$

18. Le terme anglais est « cross-validation with voting ». Par opposition, Yang nomme « cross-validation with averaging » la validation croisée habituelle, celle de la définition 1.

19. En anglais, on utilise les termes « CV bagging », « aggregated hold-out » ou « averaging cross-validation », pour un ensemble de méthodes similaires à celle qui est décrite ici.

En classification, on procède à un vote majoritaire :

$$\widetilde{f}(D_n;x) = \underset{y\in\mathcal{Y}}{\operatorname{argmax}}\Big\{\operatorname{Card}\big\{j\in\{1,\dots,V\} \,/\, \widehat{f}_{\widehat{m}_j}\big(D_n^{E_j};x\big)=y\big\}\Big\}.$$

Cette idée d'agrégation est à rapprocher du bagging, bien que la validation croisée agrégée ne coïncide pas exactement avec le bagging appliqué à la validation simple.

3 Estimation du risque : biais et variance

Si l'on utilise la validation croisée pour estimer le risque (moyen) d'une règle d'apprentissage $\widehat{f}_m$ fixée, il est naturel de s'intéresser à deux quantités : son biais et sa variance.

3.1 Biais

Sous les hypothèses (**Ind**) et (**Reg**), l'espérance d'un critère par validation croisée général se calcule aisément :

$$\mathbb{E}\Big[\widehat{\mathcal{R}}^{\text{vc}}\big(\widehat{f}_m;D_n;(E_j)_{1\leqslant j\leqslant V}\big)\Big] = \mathbb{E}\Big[\mathcal{R}_P\big(\widehat{f}_m(D_{n_e})\big)\Big] \tag{5}$$

où D_{n_e} désigne un échantillon de n_e variables indépendantes et de même loi P.

Démonstration Par définition de la validation croisée, on a :

$$\mathbb{E}\Big[\widehat{\mathcal{R}}^{\text{vc}}\big(\widehat{f}_m;D_n;(E_j)_{1\leqslant j\leqslant V}\big)\Big] = \frac{1}{V}\sum_{j=1}^{V}\mathbb{E}\Big[\widehat{\mathcal{R}}_n^{E_j^c}\Big(\widehat{f}_m\big(D_n^{E_j}\big)\Big)\Big].$$

D'après l'hypothèse (**Ind**), $D_n^{E_j}$ et $D_n^{E_j^c}$ sont deux échantillons indépendants de variables indépendantes de loi P, donc

$$\mathbb{E}\Big[\widehat{\mathcal{R}}^{\text{vc}}\big(\widehat{f}_m;D_n;(E_j)_{1\leqslant j\leqslant V}\big)\Big] = \frac{1}{V}\sum_{j=1}^{V}\mathbb{E}\Big[\mathcal{R}_P\Big(\widehat{f}_m\big(D_n^{E_j}\big)\Big)\Big].$$

Comme l'hypothèse (**Reg**) garantit que les E_j sont tous de même taille, on a

$$\mathbb{E}\Big[\mathcal{R}_P\Big(\widehat{f}_m\big(D_n^{E_j}\big)\Big)\Big] = \mathbb{E}\Big[\mathcal{R}_P\big(\widehat{f}_m(D_{n_e})\big)\Big]$$

pour tout j, d'où le résultat. □

En vue d'estimer le risque moyen

$$\mathbb{E}\Big[\mathcal{R}_P\big(\widehat{f}_m(D_n)\big)\Big],$$

d'après (5), le biais de la validation croisée s'écrit :

$$\mathbb{E}\Big[\mathcal{R}_P\big(\widehat{f}_m(D_{n_e})\big)\Big] - \mathbb{E}\Big[\mathcal{R}_P\big(\widehat{f}_m(D_n)\big)\Big]\,. \tag{6}$$

En particulier, il ne dépend pas du nombre V de découpages ! C'est seulement une fonction de la taille n_e de l'échantillon d'entraînement (en plus de n, P et $\widehat{f}_m$). Au vu de (6), la manière dont le risque moyen varie avec la taille n de l'échantillon joue donc un rôle clé dans l'analyse en espérance de la validation croisée.

Supposons tout d'abord que le risque moyen diminue quand on a plus d'observations :

$$\forall P \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y}), \qquad n \in \mathbb{N}\backslash\{0\} \mapsto \mathbb{E}\Big[\mathcal{R}_P\big(\widehat{f}_m(D_n)\big)\Big] \text{ est décroissante,} \tag{7}$$

en notant $\mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$ l'ensemble des mesures de probabilité sur $\mathcal{X} \times \mathcal{Y}$. Alors, le biais (6) est une fonction décroissante de la taille n_e de l'échantillon d'entraînement. En particulier, il est minimal lorsque $n_e = n - 1$ (par exemple, pour le leave-one-out).

Remarque 7 (Règles intelligentes) L'hypothèse (7), qui semble faible au premier abord, est la définition d'une règle d'apprentissage « intelligente » [20] (Devroye *et al.*, 1996, section 6.8). Mais attention ! Toutes les règles d'apprentissage classiques ne sont pas « intelligentes » : par exemple, la règle du plus proche voisin et certaines règles par noyau (avec un noyau et une fenêtre fixes) ne sont pas intelligentes en classification binaire (Devroye *et al.*, 1996, section 6.8 et problèmes 6.14–6.15) ! Une règle par partition sur une partition $\mathcal{A}$ indépendante de n n'est pas non plus intelligente, même si elle l'est « presque » (voir les exercices 1 et 2). Devroye *et al.* (1996, problème 6.16) conjecturent même qu'aucune règle universellement consistante n'est intelligente.

Pour quantifier plus précisément le biais, faisons une hypothèse plus forte, qui implique (7) :

$$\exists \alpha(m) \in \mathbb{R},\, \beta(m) > 0,\, \forall n \geqslant 1, \qquad \mathbb{E}\Big[\mathcal{R}_P\big(\widehat{f}_m(D_n)\big)\Big] = \alpha(m) + \frac{\beta(m)}{n}\,. \tag{8}$$

Par exemple, (8) est vérifiée pour toute loi P en estimation de densité par moindres carrés [21] ; $\alpha(m)$ est alors l'erreur d'approximation et $\beta(m)/n$ est l'erreur d'estimation (deux quantités définies par Arlot (2018a, section 3.4)). En régression avec le coût quadratique, (8) est approximativement vérifiée pour les règles par partition (Arlot, 2008). Alors, le biais (6) d'un critère par validation croisée s'écrit :

$$\beta(m)\left(\frac{1}{n_e} - \frac{1}{n}\right) = \left(\frac{n}{n_e} - 1\right)\frac{\beta(m)}{n}\,.$$

On peut distinguer trois situations :

20. En anglais, « smart rule ».

21. Ce cadre n'est pas un exemple de problème de prévision, mais on peut tout de même y définir la validation croisée, voir la parenthèse 1 en section 2.1.

— si $n_e \sim n$ (comme pour le leave-one-out), alors le biais est négligeable devant $\beta(m)/n$: au premier ordre, la validation croisée estime le risque moyen sans biais.
— si $n_e = \kappa n$ avec $\kappa \in]0,1[$, alors le biais vaut $(\kappa^{-1}-1)\beta(m)/n$: la validation croisée estime correctement l'erreur d'approximation $\alpha(m)$ mais surestime d'un facteur $\kappa^{-1} > 1$ l'erreur d'estimation $\beta(m)/n$. C'est notamment le cas de la validation croisée V-fold avec $\kappa^{-1} - 1 = 1/(V-1)$.
— si $n_e \ll n$, alors le biais est de l'ordre de

$$\frac{\beta(m)}{n_e} \gg \frac{\beta(m)}{n}\,.$$

La validation croisée surestime fortement l'erreur d'estimation, et donc aussi le risque (sauf si l'erreur d'approximation $\alpha(m)$ domine l'erreur d'estimation, auquel cas il peut n'y avoir quasiment pas de surestimation).

3.2 Correction du biais

Plutôt que de minimiser le biais en choisissant n_e proche de n (ce qui nécessite souvent de considérer un grand nombre V de découpages à cause de la variance), il est parfois possible de corriger le biais.

Définition 2 (Validation croisée corrigée) Soit $\widehat{f}_m$ une règle d'apprentissage. L'estimateur par validation croisée corrigée [22] du risque de $\widehat{f}_m$ pour l'échantillon D_n et la suite de découpages $(E_j)_{1\leqslant j\leqslant V}$ est défini par :

$$\widehat{\mathcal{R}}^{\mathrm{vc-cor}}\big(\widehat{f}_m; D_n; (E_j)_{1\leqslant j\leqslant V}\big) = \widehat{\mathcal{R}}^{\mathrm{vc}}\big(\widehat{f}_m; D_n; (E_j)_{1\leqslant j\leqslant V}\big) + \widehat{\mathcal{R}}_n\big(\widehat{f}_m(D_n)\big) - \frac{1}{V}\sum_{j=1}^{V} \widehat{\mathcal{R}}_n\big(\widehat{f}_m(D_n^{E_j})\big)\,.$$

La validation croisée corrigée a été proposée par Burman (1989), qui la justifie par des arguments asymptotiques, en supposant que $\widehat{f}_m$ est « régulière ». À la suite de Arlot (2008) et Arlot et Lerasle (2016), on peut démontrer qu'elle est *exactement* sans biais, pour tout $n \geqslant 1$, sous une hypothèse similaire à (8).

Proposition 1 *On suppose* (**Ind**) *vérifiée et qu'une constante* $\gamma(m) \in \mathbb{R}$ *existe telle que pour tout entier* $n \geqslant 1$ *:*

$$\mathbb{E}\Big[\mathcal{R}_P\big(\widehat{f}_m(D_n)\big) - \widehat{\mathcal{R}}_n\big(\widehat{f}_m(D_n)\big)\Big] = \frac{\gamma(m)}{n}\,. \tag{9}$$

Alors, pour tout $n \geqslant 1$ *et toute suite de découpages* $(E_j)_{1\leqslant j\leqslant V}$*, la validation croisée corrigée estime sans biais le risque moyen de* $\widehat{f}_m(D_n)$ *:*

$$\mathbb{E}\Big[\widehat{\mathcal{R}}^{\mathrm{vc-cor}}\big(\widehat{f}_m; D_n; (E_j)_{1\leqslant j\leqslant V}\big)\Big] = \mathbb{E}\Big[\mathcal{R}_P\big(\widehat{f}_m(D_n)\big)\Big]\,. \tag{10}$$

22. Le terme anglais est « bias-corrected cross-validation ».

Démonstration On part de la définition de la validation croisée corrigée, en remarquant que :

$$\widehat{\mathcal{R}}_n^{E_j^c} - \widehat{\mathcal{R}}_n = \frac{\mathrm{Card}(E_j)}{n}\big(\widehat{\mathcal{R}}_n^{E_j^c} - \widehat{\mathcal{R}}_n^{E_j}\big)\,.$$

Alors,

$$\begin{aligned}
&\widehat{\mathcal{R}}^{\mathrm{vc-cor}}\big(\widehat{f}_m; D_n; (E_j)_{1\leqslant j\leqslant V}\big)\\
&= \widehat{\mathcal{R}}_n\big(\widehat{f}_m(D_n)\big) + \frac{1}{V}\sum_{j=1}^{V}\Big[(\widehat{\mathcal{R}}_n^{E_j^c} - \widehat{\mathcal{R}}_n)\big(\widehat{f}_m(D_n^{E_j})\big)\Big]\\
&= \widehat{\mathcal{R}}_n\big(\widehat{f}_m(D_n)\big) + \frac{1}{V}\sum_{j=1}^{V}\frac{\mathrm{Card}(E_j)}{n}\Big[(\widehat{\mathcal{R}}_n^{E_j^c} - \widehat{\mathcal{R}}_n^{E_j})\big(\widehat{f}_m(D_n^{E_j})\big)\Big]\\
&= \mathcal{R}_P\big(\widehat{f}_m(D_n)\big) + (\widehat{\mathcal{R}}_n - \mathcal{R}_P)\big(\widehat{f}_m(D_n)\big)\\
&\quad + \frac{1}{V}\sum_{j=1}^{V}\frac{\mathrm{Card}(E_j)}{n}\Big[(\widehat{\mathcal{R}}_n^{E_j^c} - \mathcal{R}_P)\big(\widehat{f}_m(D_n^{E_j})\big)\Big]\\
&\quad + \frac{1}{V}\sum_{j=1}^{V}\frac{\mathrm{Card}(E_j)}{n}\Big[(\mathcal{R}_P - \widehat{\mathcal{R}}_n^{E_j})\big(\widehat{f}_m(D_n^{E_j})\big)\Big]\,.
\end{aligned}$$

Or, le troisième terme ci-dessus est d'espérance nulle car $D_n^{E_j}$ et $D_n^{E_j^c}$ sont indépendants d'après (**Ind**). En utilisant l'hypothèse (9), on en déduit que

$$\begin{aligned}
&\mathbb{E}\Big[\widehat{\mathcal{R}}^{\mathrm{vc-cor}}\big(\widehat{f}_m; D_n; (E_j)_{1\leqslant j\leqslant V}\big) - \mathcal{R}_P\big(\widehat{f}_m(D_n)\big)\Big]\\
&= \frac{-\gamma(m)}{n} + \frac{1}{V}\sum_{j=1}^{V}\frac{\mathrm{Card}(E_j)}{n}\frac{\gamma(m)}{\mathrm{Card}(E_j)} = 0\,.
\end{aligned}$$

□

Parenthèse 8 (Sur l'hypothèse (9)) L'hypothèse (9) porte sur l'espérance de la pénalité idéale (voir Arlot (2018a, section 3.9)). Elle est vérifiée en estimation de densité par moindres carrés (Arlot et Lerasle, 2016), et elle l'est (approximativement) pour les règles de régression par partition avec le coût quadratique (Arlot, 2008). On peut noter que dans les deux cas, les hypothèses (8) et (9) sont vérifiées avec $\gamma(m) = 2\beta(m)$, ce qui est lié à l'heuristique de pente (Birgé et Massart, 2007; Arlot et Massart, 2009; Lerasle, 2012). De manière plus générale, l'argument asymptotique de Burman (1989) peut se résumer à établir que (9) est approximativement valide sous des hypothèses de régularité sur $\widehat{f}_m$ et sur le coût c.

3.3 Variance

L'étude de la variance des estimateurs par validation croisée est plus délicate que celle de leur espérance. On peut toutefois établir quelques résultats généraux. En revanche, des résultats quantitatifs précis (et valables pour les procédures par blocs) ne sont actuellement connus que dans des cadres particuliers, comme celui considéré à la fin de cette section.

3.3.1 Inégalité générale

On a tout d'abord une inégalité générale entre les variances des estimateurs par validation simple, par validation croisée et par leave-p-out, à taille d'échantillon d'entraînement $n_e = n - p$ fixée.

Proposition 2 *Soit $n, V \geqslant 1$ des entiers, D_n un échantillon de n variables aléatoires indépendantes et de même loi, $n_e \in \{1, \dots, n-1\}$ et $E_0, E_1, \dots, E_V$ des parties de $\{1, \dots, n\}$ indépendantes de D_n et telles que $\mathrm{Card}(E_j) = n_e$ pour tout $j \in \{0, \dots, V\}$. On a alors :*

$$\begin{aligned}\mathrm{var}\big(\widehat{\mathcal{R}}^{\mathrm{val}}\big(\widehat{f}_m; D_n; E_0\big)\big) &\geqslant \mathrm{var}\Big(\widehat{\mathcal{R}}^{\mathrm{vc}}\big(\widehat{f}_m; D_n; (E_j)_{1\leqslant j\leqslant V}\big)\Big) \\ &\geqslant \mathrm{var}\big(\widehat{\mathcal{R}}^{\mathrm{lpo}}(\widehat{f}_m; D_n; n - n_e)\big)\,.\end{aligned}$$

Démonstration On commence par faire une remarque générale. Si $Z_1, \dots, Z_K$ sont des variables aléatoires réelles de même loi, alors :

$$\mathrm{var}(Z_1) \geqslant \mathrm{var}\left(\frac{1}{K}\sum_{i=1}^{K} Z_i\right). \tag{11}$$

En effet, par convexité, on a :

$$\frac{1}{K}\sum_{i=1}^{K} Z_i^2 \geqslant \left(\frac{1}{K}\sum_{i=1}^{K} Z_i\right)^2.$$

En intégrant cette inégalité, on obtient :

$$\mathbb{E}\big[Z_1^2\big] \geqslant \mathbb{E}\left[\left(\frac{1}{K}\sum_{i=1}^{K} Z_i\right)^2\right].$$

Le résultat s'en déduit en retranchant de chaque côté :

$$\Big(\mathbb{E}\big[Z_1\big]\Big)^2 = \left(\mathbb{E}\left[\frac{1}{K}\sum_{i=1}^{K} Z_i\right]\right)^2.$$

Or, la loi de $\widehat{\mathcal{R}}^{\mathrm{val}}\big(\widehat{f}_m; D_n; E\big)$ est la même quel que soit $E \subset \{1, \dots, n\}$ de taille n_e, y compris $E = E_j$ (par indépendance des E_j avec D_n). En effet, passer de E à E' de même taille équivaut à permuter les observations de D_n, ce qui ne

change pas la loi de D_n car les observations sont indépendantes et de même loi. L'inégalité (11) implique donc la première inégalité.

Pour obtenir la deuxième, notons Σ_n l'ensemble des permutations de $\{1,\dots,n\}$ et $D_n^\sigma = (X_{\sigma(i)}, Y_{\sigma(i)})_{1\leqslant i\leqslant n}$ pour toute permutation $\sigma \in \Sigma_n$. On a alors, pour tout $j \in \{1,\dots,V\}$,

$$\widehat{\mathcal{R}}^{\mathrm{lpo}}\big(\widehat{f}_m; D_n; n-n_e\big) = \frac{1}{n!}\sum_{\sigma\in\Sigma_n} \widehat{\mathcal{R}}^{\mathrm{val}}\big(\widehat{f}_m; D_n^\sigma; E_j\big)\,,$$

$$\text{et donc}\qquad \widehat{\mathcal{R}}^{\mathrm{lpo}}\big(\widehat{f}_m; D_n; n-n_e\big) = \frac{1}{n!}\sum_{\sigma\in\Sigma_n} \widehat{\mathcal{R}}^{\mathrm{vc}}\big(\widehat{f}_m; D_n^\sigma; (E_j)_{1\leqslant j\leqslant V}\big)\,.$$

Or, D_n^σ a même loi que D_n, puisque les observations sont indépendantes et de même loi, si bien que (11) s'applique et donne la deuxième inégalité. □

Parenthèse 9 (Variance conditionnelle et (11)) En utilisant la formule (12) ci-après et le fait qu'une variance est positive ou nulle, on obtient que

$$\mathrm{var}(Z) \geqslant \mathrm{var}\big(\mathbb{E}[Z \mid X]\big)\,,$$

ce qui permet de démontrer l'inégalité (11) d'une autre manière.

Parenthèse 10 (Proposition 2 pour des données dépendantes) Une lecture attentive de sa preuve démontre que la proposition 2 est encore valable lorsque D_n est échangeable, c'est-à-dire lorsque D_n^σ a même loi que D_n pour toute permutation (déterministe) σ.

Le résultat de la proposition 2 est naturel : à taille d'échantillon d'entraînement n_e fixée, il vaut mieux considérer plusieurs découpages qu'un seul, et il vaut encore mieux les considérer tous. Mais c'est un résultat limité : l'amélioration est-elle stricte ? Si oui, combien gagne-t-on ? La proposition 2 ne permet pas de savoir. Pire encore : on ne peut pas comparer ce qui se passe avec deux et avec trois découpages !

Intuitivement, plus le nombre V de découpages est grand, plus la variance est petite. La réalité est malheureusement un peu plus compliquée, et cela dépend de *comment* on découpe. Cinq découpages bien choisis peuvent être meilleurs que six mal choisis !

3.3.2 Validation croisée Monte-Carlo

Si l'on se limite à une famille particulière de découpages (la validation croisée Monte-Carlo), cette intuition peut être justifiée et quantifiée.

Proposition 3 *Soit $n > n_e \geqslant 1$ des entiers. Si $E_1,\dots,E_V \subset \{1,\dots,n\}$ sont indépendantes, de loi uniforme sur l'ensemble des parties de $\{1,\dots,n\}$ de taille n_e, et*

indépendantes de D_n, alors on a :

$$\begin{aligned}
&\operatorname{var}\Big(\widehat{\mathcal{R}}^{\mathrm{vc}}\big(\widehat{f}_m; D_n; (E_j)_{1\leqslant j\leqslant V}\big)\Big)\\
&= \operatorname{var}\big(\widehat{\mathcal{R}}^{\mathrm{lpo}}(\widehat{f}_m; D_n; n-n_e)\big) + \frac{1}{V}\mathbb{E}\left[\operatorname{var}\Big(\widehat{\mathcal{R}}_n^{E_1^c}\big(\widehat{f}_m\big(D_n^{E_1}\big)\big)\,\Big|\,D_n\Big)\right]\\
&= \operatorname{var}\big(\widehat{\mathcal{R}}^{\mathrm{lpo}}(\widehat{f}_m; D_n; n-n_e)\big)\\
&\quad + \frac{1}{V}\Big[\operatorname{var}\big(\widehat{\mathcal{R}}^{\mathrm{val}}(\widehat{f}_m; D_n; E_1)\big) - \operatorname{var}\big(\widehat{\mathcal{R}}^{\mathrm{lpo}}(\widehat{f}_m; D_n; n-n_e)\big)\Big]\,.
\end{aligned}$$

Démonstration Avant toutes choses, rappelons la définition de la variance conditionnelle et une propriété fondamentale. Si X et Z sont deux variables aléatoires, si Z est à valeurs réelles et si $\mathbb{E}[Z^2\,|\,X] < +\infty$, la variance conditionnelle de Z sachant X est définie par :

$$\operatorname{var}(Z\,|\,X) = \mathbb{E}\big[Z^2\,|\,X\big] - \mathbb{E}[Z\,|\,X]^2\,.$$

On a alors, en intégrant cette définition :

$$\operatorname{var}(Z) = \mathbb{E}\big[\operatorname{var}(Z\,|\,X)\big] + \operatorname{var}\big(\mathbb{E}[Z\,|\,X]\big)\,. \tag{12}$$

En vue d'appliquer la formule (12), on pose :

$$X = D_n \qquad \text{et} \qquad Z = \widehat{\mathcal{R}}^{\mathrm{vc}}\big(\widehat{f}_m; D_n; (E_j)_{1\leqslant j\leqslant V}\big)\,.$$

D'une part, puisque les E_j sont de loi uniforme sur l'ensemble des parties de $\{1,\dots,n\}$ de taille n_e, on obtient :

$$\mathbb{E}[Z\,|\,D_n] = \widehat{\mathcal{R}}^{\mathrm{lpo}}(\widehat{f}_m; D_n; n-n_e)\,.$$

D'autre part, conditionnellement à D_n, les E_j sont indépendantes et de même loi, si bien que l'on a :

$$\operatorname{var}(Z\,|\,D_n) = \frac{1}{V}\operatorname{var}\left(\widehat{\mathcal{R}}_n^{E_1^c}\Big(\widehat{f}_m\big(D_n^{E_1}\big)\Big)\,\Big|\,D_n\right).$$

Avec (12), on en déduit la première formule.

Lorsque $V = 1$, la validation croisée Monte-Carlo correspond à la validation simple, et donc :

$$\begin{aligned}
&\operatorname{var}\Big(\widehat{\mathcal{R}}^{\mathrm{val}}\big(\widehat{f}_m; D_n; E_1\big)\Big)\\
&= \operatorname{var}\Big(\widehat{\mathcal{R}}^{\mathrm{lpo}}\big(\widehat{f}_m; D_n; n-n_e\big)\Big) + \mathbb{E}\left[\operatorname{var}\Big(\widehat{\mathcal{R}}_n^{E_1^c}\big(\widehat{f}_m\big(D_n^{E_1}\big)\big)\,\Big|\,D_n\Big)\right].
\end{aligned}$$

La deuxième formule en découle. □

La proposition 3 démontre que la variance du critère par validation croisée Monte-Carlo est une fonction décroissante du nombre V de découpages, ce qui confirme l'intuition énoncée précédemment. Plus précisément, cette variance est une fonction affine

de $1/V$, comprise entre la variance du critère par validation simple ($V = 1$) et la variance du critère leave-p-out correspondant (obtenu quand $V \to +\infty$). En particulier, lorsque les variances des critères par validation simple et par leave-p-out ne diffèrent que par une constante multiplicative, l'essentiel de l'effet de V sur la décroissance de la variance est obtenu avec un nombre V de découpages petit (par exemple, $V = 20$ suffit à obtenir au moins 95% de la décroissance espérée).

Remarque 11 Les propositions 2 et 3 s'étendent à la validation croisée corrigée, et plus généralement à toute quantité de la forme

$$Z_V = \frac{1}{V} \sum_{j=1}^{V} F(D_n; E_j)$$

avec D_n, et $(E_1, \ldots, E_V)$ indépendantes telles que $\mathrm{Card}(E_j) = n_e$ presque sûrement pour tout j (et, pour la proposition 3, l'hypothèse supplémentaire que $E_1, \ldots, E_V$ sont indépendantes de même loi). Ainsi, en posant

$$Z_{n_e}^{\mathrm{lpo}} := \frac{1}{\binom{n}{n_e}} \sum_{E \,/\, \mathrm{Card}(E)=n_e} F(D_n; E)\,,$$

la proposition 2 devient

$$\mathrm{var}\big(F(D_n; E_1)\big) \geqslant \mathrm{var}(Z_V) \geqslant \mathrm{var}(Z_{n_e}^{\mathrm{lpo}})$$

et la proposition 3 devient

$$\begin{aligned} \mathrm{var}(Z_V) &= \mathrm{var}(Z_{n_e}^{\mathrm{lpo}}) + \frac{1}{V}\mathbb{E}\Big[\mathrm{var}\big(F(D_n; E_1) \,|\, D_n\big)\Big] \\ &= \mathrm{var}(Z_{n_e}^{\mathrm{lpo}}) + \frac{1}{V}\Big[\mathrm{var}\big(F(D_n; E_1)\big) - \mathrm{var}(Z_{n_e}^{\mathrm{lpo}})\Big]\,. \end{aligned}$$

On peut en particulier appliquer ces résultats à la validation croisée V-fold répétée, définie à la fin de la section 2.2 (voir l'exercice 3).

Il est malheureusement difficile d'énoncer un résultat plus précis en toute généralité. On peut seulement détailler comme suit la variance de l'estimateur par validation simple (la démonstration de ce résultat constitue l'exercice 4) :

$$\begin{aligned} \mathrm{var}\Big(\widehat{\mathcal{R}}_n^{E^c}\big(\widehat{f}_m(D_n^E)\big)\Big) &= \mathrm{var}\Big(\mathcal{R}_P\big(\widehat{f}_m(D_n^E)\big)\Big) \\ &\quad + \frac{1}{\mathrm{Card}(E^c)}\mathbb{E}\Big[\mathrm{var}\Big(c\big(\widehat{f}_m(D_n^E; X), Y\big) \,\Big|\, \widehat{f}_m(D_n^E)\Big)\Big]\,. \end{aligned} \tag{13}$$

La formule ci-dessus souligne le rôle de la taille de l'échantillon de validation, qui fait diminuer la variance de l'estimation du risque de $\widehat{f}_m(D_n^E)$, jusqu'à atteindre (virtuellement) sa valeur minimale possible (la variance du risque de $\widehat{f}_m(D_n^E)$) lorsque l'échantillon de validation est infini. Soulignons néanmoins l'intérêt limité de (13), qui ne permet pas (telle quelle) de comparer différentes valeurs de $n_e = \mathrm{Card}(E)$, notamment parce que n_e est lié à la taille de l'échantillon de validation $\mathrm{Card}(E^c) = n - n_e$.

3.3.3 Quantification précise dans un cadre particulier

En considérant une règle $\widehat{f}_m$ et un coût c particuliers, il est parfois possible d'énoncer des résultats plus précis, soit asymptotiquement (quand n tend vers l'infini, avec $n_e = \mathrm{Card}(E_j)$ éventuellement lié à n), soit pour une valeur de n fixée (Arlot et Celisse, 2010, section 5.2). À titre d'illustration, considérons le cas de l'estimation de densité avec le coût des moindres carrés et $\widehat{f}_m$ un estimateur par histogrammes réguliers de pas $h_m > 0$ (Arlot et Lerasle, 2016, énoncent un résultat valable pour des estimateurs par projection généraux). On a alors, pour la validation croisée V-fold :

$$\begin{aligned} &\mathrm{var}\Big(\widehat{\mathcal{R}}^{\mathrm{vf}}\big(\widehat{f}_m; D_n; (B_j)_{1\leqslant j\leqslant V}\big)\Big) \\ &\quad = \frac{1}{n^2} C_1^{\mathrm{vf}}(V,n)\mathcal{W}_1(h_m,P) + \frac{1}{n} C_2^{\mathrm{vf}}(V,n)\mathcal{W}_2(h_m,P) \end{aligned} \tag{14}$$

où les $\mathcal{W}_i(h_m,P)$ ne dépendent que de h_m et de la loi P des observations (Arlot et Lerasle, 2016, en donnent une formule close en section 5),

$$C_1^{\mathrm{vf}}(V,n) = 1 + \frac{4}{V-1} + \frac{4}{(V-1)^2} + \frac{1}{(V-1)^3} - \frac{V^2}{n(V-1)^2}$$

$$\text{et} \qquad C_2^{\mathrm{vf}}(V,n) = \left(1 + \frac{V}{n(V-1)}\right)^2 .$$

En particulier, la variance diminue avec V et le gain induit par l'augmentation de V est limité : pour n assez grand, le premier terme varie de 10 à 1 quand V augmente, tandis que le deuxième terme reste à peu près constant. Lorsque h_m est fixe par rapport à n, la diminution de variance liée à l'augmentation de V est extrêmement faible : elle ne se voit que dans le premier terme du membre de droite de (14), qui est un terme du deuxième ordre.

Une formule similaire peut être établie dans le même cadre pour la validation croisée Monte-Carlo (Arlot et Lerasle, 2016, section 8.1) :

$$\begin{aligned} &\mathrm{var}\Big(\widehat{\mathcal{R}}^{\mathrm{vc}}\big(\widehat{f}_m; D_n; (E_j)_{1\leqslant j\leqslant V}\big)\Big) \\ &\quad = \frac{1}{n^2} C_1^{\mathrm{MC}}(V,n,n_e)\mathcal{W}_1(h_m,P) + \frac{1}{n} C_2^{\mathrm{MC}}(V,n,n_e)\mathcal{W}_2(h_m,P) \end{aligned} \tag{15}$$

où les $\mathcal{W}_i$ sont les mêmes qu'à l'équation (14),

$$\begin{aligned} C_1^{\mathrm{MC}}(V,n,n_e) &= \frac{1}{V}\left(\frac{n^2}{n_e^2} + \frac{2n^2}{n_e(n-n_e)} - \frac{1}{n}\frac{n^3}{n_e^3}\right) \\ &\qquad + \left(1 - \frac{1}{V}\right)\left[1 + \frac{1}{n-1}\left(\frac{n}{n_e} + 1\right)^2 - \frac{1}{n}\frac{n^2}{n_e^2}\right] \end{aligned}$$

$$\text{et} \qquad C_2^{\mathrm{MC}}(V,n,n_e) = \frac{1}{V}\left(\frac{n}{n-n_e} + \frac{1}{n^2}\frac{n^3}{n_e^3}\right) + \left(1 - \frac{1}{V}\right)\left(1 + \frac{1}{n}\frac{n}{n_e}\right)^2 .$$

Ainsi, lorsque n tend vers l'infini et $n_e \sim \tau n$ avec $\tau \in]0,1[$ fixé, les valeurs extrêmes de la variance sont données par

$$C_1^{\mathrm{MC}}(1,n,n_e) \xrightarrow[n\to+\infty]{} \frac{1}{\tau^2} + \frac{2}{\tau(1-\tau)} > 11$$
$$\text{et} \qquad C_2^{\mathrm{MC}}(1,n,n_e) \xrightarrow[n\to+\infty]{} \frac{1}{1-\tau} > 1$$

d'une part (pour la validation simple), et par

$$C_1^{\mathrm{MC}}(\infty,n,n_e) \xrightarrow[n\to+\infty]{} 1 \qquad \text{et} \qquad C_2^{\mathrm{MC}}(\infty,n,n_e) \xrightarrow[n\to+\infty]{} 1$$

d'autre part (pour le leave-$(n-n_e)$-out). Le gain lié à l'augmentation du nombre de découpages est donc de l'ordre d'une constante (fonction de τ uniquement), et il n'est pas énorme (au plus un facteur 12 lorsque $\tau = 1/2$, par exemple).

Les formules (14) et (15) permettent également d'évaluer l'intérêt de considérer une suite « structurée » de découpages (avec le V-fold) par rapport à une suite aléatoire (avec la méthode Monte-Carlo). Il suffit pour cela de comparer les valeurs des constantes C_i pour une même taille d'échantillon d'entraînement $n_e = n(V_n-1)/V_n$ et un même nombre de découpages V_n (que l'on autorise à varier avec n). Si $V_n \in \{3,\dots,n\}$ pour tout $n \geqslant 1$, alors

$$\liminf_{n\to+\infty} \frac{C_1^{\mathrm{MC}}\left(V_n, n, \frac{n(V_n-1)}{V_n}\right)}{C_1^{\mathrm{VF}}(V_n,n)} > 1$$

et cette limite vaut 3 lorsque $\lim_{n\to+\infty} V_n = +\infty$, tandis que

$$\frac{C_2^{\mathrm{MC}}\left(V_n, n, \frac{n(V_n-1)}{V_n}\right)}{C_2^{\mathrm{VF}}(V_n,n)} \underset{n\to+\infty}{\sim} 2 - \frac{1}{V_n} \in \left[\frac{3}{2}, 2\right].$$

La validation croisée par blocs permet donc d'avoir une variance plus faible qu'avec la validation croisée Monte-Carlo, avec un gain de l'ordre d'une constante numérique, compris entre 3/2 et 3 environ [23].

Parenthèse 12 (Validation croisée corrigée) La variance du critère par validation croisée corrigée s'écrit d'une manière similaire à (14) (pour le V-fold) et (15) (pour la procédure Monte-Carlo), avec les mêmes conclusions sur l'influence de V.

Notons pour finir que l'on peut conjecturer que l'essentiel des conclusions obtenues ci-dessus pour les histogrammes en estimation de densité sont valables pour tout prédicteur $\widehat{f}_m$ « régulier », au moins asymptotiquement au vu des résultats de Burman (1989). Toutefois, des comportements expérimentaux différents sont rapportés dans

23. Le gain exact dépend de l'ordre de grandeur relatif de $\mathcal{W}_1(h_m,P)$ et $\mathcal{W}_2(h_m,P)$.

la littérature, notamment pour des prédicteurs « instables », par exemple un arbre CART, un réseau de neurones ou le résultat d'une procédure de sélection de variables par pénalisation « ℓ^0 » telle que AIC ou C_p (Breiman, 1996). Arlot et Celisse (2010, section 5.2) donnent d'autres références à ce sujet.

4 Propriétés pour la sélection d'estimateurs

Revenons au problème (décrit en section 1) de choisir parmi une famille de règles d'apprentissage $(\widehat{f}_m)_{m\in\mathcal{M}_n}$, avec un objectif de prévision : on veut minimiser le risque du prédicteur final $\widehat{f}_{\widehat{m}(D_n)}(D_n)$. Étant donné une suite de découpages $(E_j)_{1\leqslant j\leqslant V}$ — dont on suppose toujours qu'elle vérifie (**Ind**) et (**Reg**) — la procédure de validation croisée correspondante sélectionne :

$$\widehat{m}^{\mathrm{vc}}\big(D_n;(E_j)_{1\leqslant j\leqslant V}\big) \in \operatorname*{argmin}_{m\in\mathcal{M}_n}\Big\{\widehat{\mathcal{R}}^{\mathrm{vc}}\big(\widehat{f}_m;D_n;(E_j)_{1\leqslant j\leqslant V}\big)\Big\}\,.$$

4.1 Différence entre sélection d'estimateurs et estimation du risque

La section 3 détaille les performances de $\widehat{\mathcal{R}}^{\mathrm{vc}}(\widehat{f}_m)$ comme estimateur du risque de $\widehat{f}_m$. Intuitivement, ceci devrait permettre d'évaluer les performances de la validation croisée pour la sélection d'estimateurs. Cette intuition est en partie correcte, mais en partie seulement, attention ! On constate en effet empiriquement que la meilleure procédure de sélection d'estimateurs ne correspond pas toujours au meilleur estimateur du risque (Breiman et Spector, 1992). Il y a au moins deux raisons à ce phénomène.

4.1.1 Rôle des incréments

Un premier élément d'explication peut être obtenu en considérant l'exemple jouet suivant. Supposons que l'on veut comparer les procédures

$$\widehat{m}_1 \in \operatorname*{argmin}_{m\in\mathcal{M}_n}\big\{\mathcal{C}_1(m;D_n)\big\} \qquad \text{et} \qquad \widehat{m}_2 \in \operatorname*{argmin}_{m\in\mathcal{M}_n}\big\{\mathcal{C}_2(m;D_n)\big\}$$

où $\mathcal{C}_1$ est un estimateur sans biais du risque et, pour tout $m\in\mathcal{M}_n$,

$$\mathcal{C}_2(m;D_n) = \mathcal{C}_1(m;D_n) + Z$$

avec Z une variable aléatoire de loi normale $\mathcal{N}(1,1)$ indépendante de D_n et de m. Clairement, $\widehat{m}_1 = \widehat{m}_2$ et il n'y a aucune différence entre ces deux procédures pour la sélection d'estimateurs. En revanche, $\mathcal{C}_2$ est bien plus mauvais que $\mathcal{C}_1$ pour l'estimation du risque : $\mathcal{C}_1$ est un estimateur sans biais du risque tandis que $\mathcal{C}_2$ possède un biais strictement positif (égal à 1) et une variance strictement supérieure à celle de $\mathcal{C}_1$. Considérer seulement l'espérance et la variance de $\mathcal{C}_i(m)$ pour chaque $m\in\mathcal{M}_n$ peut donc être trompeur pour la sélection d'estimateurs.

Que faire pour corriger ce problème ? À la suite d'Arlot et Lerasle (2016, section 4), on peut remarquer que l'essentiel est de *bien classer* les règles $\widehat{f}_m$ les unes par rapport aux autres [24]. Autrement dit, une procédure de la forme

$$\widehat{m} \in \operatorname*{argmin}_{m \in \mathcal{M}_n} \big\{ \mathcal{C}(m; D_n) \big\}$$

est bonne lorsque, avec grande probabilité, pour tout $m, m' \in \mathcal{M}_n$,

$$\operatorname{signe}\big(\mathcal{C}(m; D_n) - \mathcal{C}(m'; D_n)\big) = \operatorname{signe}\big(\mathcal{R}_P(\widehat{f}_m(D_n)) - \mathcal{R}_P(\widehat{f}_{m'}(D_n)\big)\,,$$

sauf éventuellement lorsque $\widehat{f}_m$ et $\widehat{f}_{m'}$ ont des risques « proches » (une erreur entre m et m' étant alors de faible importance pour le risque du prédicteur final). Si l'on veut se focaliser sur une espérance et une variance (pour évaluer la qualité d'un critère $\mathcal{C}$ ou pour comparer deux critères $\mathcal{C}_1$ et $\mathcal{C}_2$), c'est donc

$$\mathbb{E}\big[\mathcal{C}(m; D_n) - \mathcal{C}(m'; D_n)\big] \quad \text{et} \quad \operatorname{var}\big(\mathcal{C}(m; D_n) - \mathcal{C}(m'; D_n)\big)\,, \qquad m, m' \in \mathcal{M}_n\,,$$

qu'il faut considérer ; les sections 4.2 et 4.3 le font pour le cas de la validation croisée.

4.1.2 Surpénalisation

Le paragraphe précédent explique pourquoi, entre deux critères dont les incréments ont la même espérance, il vaut mieux choisir celui dont la variance des incréments est la plus petite. Il est moins évident de comparer précisément des critères qui n'ont pas exactement la même espérance.

Pour ce faire, prenons le point de vue de la pénalisation (Arlot, 2018a, section 3.9), en posant [25] :

$$\mathcal{C}(m; D_n) = \widehat{\mathcal{R}}_n\big(\widehat{f}_m(D_n)\big) + \operatorname{pen}(m; D_n)\,.$$

L'idéal est alors que la pénalité $\operatorname{pen}(m; D_n)$ soit proche de

$$\operatorname{pen}_{\mathrm{id}}(m; D_n) := \mathcal{R}_P\big(\widehat{f}_m(D_n)\big) - \widehat{\mathcal{R}}_n\big(\widehat{f}_m(D_n)\big)$$

ou son espérance (qui est, dans de nombreux cas, proportionnelle à la « complexité » de la modélisation du lien entre X et Y réalisée par $\widehat{f}_m$ — par exemple, le nombre de paramètres que $\widehat{f}_m$ apprend). Considérons alors, pour tout $C > 0$, la procédure (théorique) qui sélectionne

$$m_C \in \operatorname*{argmin}_{m \in \mathcal{M}_n} \Big\{ \widehat{\mathcal{R}}_n\big(\widehat{f}_m(D_n)\big) + C\mathbb{E}\big[\operatorname{pen}_{\mathrm{id}}(m; D_n)\big] \Big\}\,.$$

---

24. À première vue, l'essentiel est de bien classer la meilleure règle parmi $(\widehat{f}_m)_{m \in \mathcal{M}_n}$ par rapport aux autres. Mais si l'on demande que ceci soit encore vrai pour toute sous-collection $(\widehat{f}_m)_{m \in \mathcal{M}'_n}$, $\mathcal{M}'_n \subset \mathcal{M}_n$, alors on demande à bien classer *toutes* les règles $\widehat{f}_m$ les unes par rapport aux autres.

25. En définissant $\operatorname{pen}(m; D_n) = \mathcal{C}(m; D_n) - \widehat{\mathcal{R}}_n(\widehat{f}_m(D_n))$, on peut écrire tout critère $\mathcal{C}$ comme un critère empirique pénalisé.

Lorsque $C = 1$, m_C suit le principe d'estimation sans biais du risque. Lorsque $C > 1$, m_C surpénalise, c'est-à-dire qu'elle choisit une règle d'apprentissage moins « complexe ». Lorsque $C < 1$, m_C sous-pénalise et choisit une règle plus « complexe ». La figure 1 représente, sur un exemple, la performance pour la sélection d'estimateurs de m_C en fonction de la constante de surpénalisation C. On peut observer que la perfor-

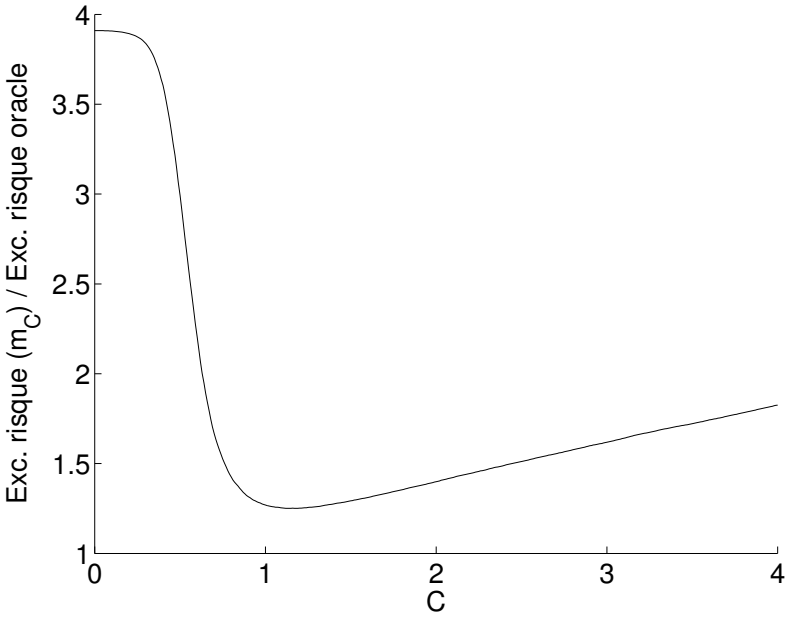

FIGURE 1 – *Performance de m_C pour la sélection de modèles (mesurée par l'espérance de $\ell(f^\star, \widehat{f}_{m_C})/\inf_{m\in\mathcal{M}_n}\{\ell(f^\star, \widehat{f}_m)\}$) en fonction de la constante de surpénalisation C, évaluée sur un jeu de données simulé, en régression linéaire avec des estimateurs des moindres carrés (même cadre que celui de (Arlot, 2018a, figure 4)). Ici, la valeur optimale de C est $1,12$.*

mance est bonne pour $C = 1$, et qu'elle est un peu meilleure lorsque C est légèrement plus grand. Surpénaliser (donc, utiliser un estimateur *biaisé* du risque) peut ainsi être bénéfique !

À notre connaissance, il n'existe pas de résultat théorique rendant pleinement compte de ce phénomène. On peut néanmoins proposer l'hypothèse suivante : en surpénalisant, on diminue la probabilité [26] de surapprendre fortement, au prix d'une diminution de la « complexité » moyenne de la règle sélectionnée. Lorsque le gain lié au premier point est plus fort que la perte liée au deuxième point, il est bénéfique de surpénaliser.

La règle du « 1 écart-type » [27] de Breiman *et al.* (1984) semble être, empiriquement,

26. Cette probabilité reste non-nulle, à cause de la variabilité du critère $\mathcal{C}(m; D_n)$.
27. En anglais, « 1 s.e. rule ».

un bon moyen de surpénaliser [28]. Une autre manière de surpénaliser a été proposée récemment par Saumard et Navarro (2018).

La section 4.4 discute les conséquences de ce phénomène sur les procédures de validation croisée.

4.2 Analyse au premier ordre : espérance

Au premier ordre, pour les procédures de validation croisée usuelles [29], les résultats théoriques connus à ce jour indiquent tous que la performance pour la sélection d'estimateurs est celle de la procédure (théorique) qui minimise l'espérance du critère correspondant.

Pour prendre un exemple, supposons que les hypothèses (**Ind**), (**Reg**) et (8) sont vérifiées. Alors, pour tout $m, m' \in \mathcal{M}_n$, on a :

$$\begin{aligned} &\mathbb{E}\Big[\widehat{\mathcal{R}}^{\mathrm{vc}}\big(\widehat{f}_m; D_n; (E_j)_{1\leqslant j\leqslant V}\big) - \widehat{\mathcal{R}}^{\mathrm{vc}}\big(\widehat{f}_{m'}; D_n; (E_j)_{1\leqslant j\leqslant V}\big)\Big] \\ &\qquad = \alpha(m) - \alpha(m') + \frac{\beta(m) - \beta(m')}{n_e} \end{aligned} \tag{16}$$

et

$$\begin{aligned} &\mathbb{E}\Big[\mathcal{R}_P\big(\widehat{f}_m(D_n)\big) - \mathcal{R}_P\big(\widehat{f}_{m'}(D_n)\big)\Big] \\ &\qquad = \alpha(m) - \alpha(m') + \frac{\beta(m) - \beta(m')}{n}\,. \end{aligned} \tag{17}$$

Comme en section 3.1, on voit que le biais de la validation croisée vient du fait que l'échantillon d'entraînement est de taille n_e au lieu de n. En particulier, ce biais ne dépend de la suite de découpages $(E_j)_{1\leqslant j\leqslant V}$ qu'à travers n_e.

Quel impact sur la sélection d'estimateurs ? En comparant (16) et (17), on observe que la différence des erreurs d'approximation $\alpha(m) - \alpha(m')$ est correctement estimée, tandis que la différence des erreurs d'estimation $(\beta(m) - \beta(m'))/n$ est estimée à un facteur multiplicatif $n/n_e > 1$ près. Puisque $\beta(m)$ mesure en général la « complexité » de modélisation de $\widehat{f}_m$ (par exemple, via la dimension du modèle sous-jacent), ceci signifie que la validation croisée tend à sélectionner une règle $\widehat{f}_{\widehat{m}}$ de complexité plus petite que celle de la meilleure règle (l'oracle). C'est naturel : la validation croisée « fait comme si » chaque règle d'apprentissage $\widehat{f}_m$ était entraînée avec n_e observations, ce qui la conduit à faire un choix plus conservateur que celui qu'il faut faire quand on dispose de $n > n_e$ observations.

Sur le plan quantitatif, l'impact de n_e est en général similaire à ce que l'on a décrit en section 3.1 pour l'estimation du risque. Shao (1997) énonce des résultats précis en régression linéaire par moindres carrés et Arlot et Celisse (2010, section 6) donnent les références de nombreux autres résultats, que l'on peut résumer ainsi. Pour la validation croisée, trois cas sont à distinguer :

28. Précisons toutefois que cette règle n'est pas directement formulée en ces termes.

29. Les procédures utilisant un échantillon de validation de taille $n - n_e \ll n$ *et* un nombre de découpages V petit font exception. Les garanties théoriques ne s'appliquent pas (Arlot et Celisse, 2010), et l'on constate en pratique que ces procédures sont extrêmement variables et se comportent mal, même lorsque leur espérance fournit une bonne procédure de sélection d'estimateurs.

— lorsque $n_e \sim n$, on obtient une performance optimale au premier ordre.
— lorsque $n_e \sim \kappa n$ avec $\kappa \in]0,1[$, on obtient une performance sous-optimale : le prédicteur final perd un facteur constant (fonction de κ notamment) par rapport au risque du meilleur prédicteur.
— lorsque $n_e \ll n$, la performance est également sous-optimale, mais la perte est d'un facteur qui tend vers l'infini lorsque n tend vers l'infini.

Pour la validation croisée corrigée, le biais étant nul, on obtient une performance optimale au premier ordre quel que soit n_e.

La principale limite de ces résultats au premier ordre est qu'ils ne font aucune différence entre des procédures utilisant des échantillons d'entraînement de même taille et dont les performances empiriques sont bien différentes (par exemple, la validation simple et le leave-p-out). Il semble donc bien qu'en pratique, les termes de deuxième ordre comptent !

4.3 Analyse au deuxième ordre : variance

L'étude de la variance (des incréments du critère, d'après la section 4.1) permet d'expliquer bon nombre de phénomènes observés empiriquement (en supposant l'heuristique de la section 4.1 correcte, ce que l'on fait tout au long de cette section).

Tout d'abord, les résultats généraux de la section 3.3, sur la variance d'un estimateur par validation croisée du risque d'une règle $\widehat{f}_m$, se généralisent à la variance des incréments

$$\widehat{\mathcal{R}}^{\mathrm{vc}}\big(\widehat{f}_m; D_n; (E_j)_{1\leqslant j\leqslant V}\big) - \widehat{\mathcal{R}}^{\mathrm{vc}}\big(\widehat{f}_{m'}; D_n; (E_j)_{1\leqslant j\leqslant V}\big)$$

quelles que soient $\widehat{f}_m$ et $\widehat{f}_{m'}$ deux règles d'apprentissage. Ceci confirme donc l'intuition selon laquelle, à taille d'échantillon d'entraînement n_e fixée, le leave-$(n-n_e)$-out est la meilleure procédure de sélection d'estimateurs, la validation simple la moins bonne, toutes les autres procédures de validation croisée ayant des performances entre les deux. Et pour la validation croisée Monte-Carlo, la performance en sélection d'estimateurs s'améliore quand V augmente. De plus, la variance des incréments étant une fonction affine croissante de $1/V$, si les valeurs de variance sont comparables pour validation simple et pour le leave-$(n-n_e)$-out, il n'est pas nécessaire de prendre V très grand pour avoir une performance très proche de l'optimum (à n_e fixée).

Pour pouvoir énoncer des résultats quantitatifs précis sur la variance des incréments, comme en section 3.3, considérons le cadre de l'estimation de densité avec le coût des moindres carrés et des estimateurs par histogrammes réguliers de pas $h_m > 0$ (Arlot et Lerasle, 2016). Alors, les résultats énoncés en section 3.3 se généralisent, seuls les termes $\mathcal{W}_i(h_m, P)$ devant être changés en $\mathcal{W}_i(h_m, h_{m'}, P)$. Toutes les comparaisons qualitatives entre méthodes énoncées en section 3.3 sont donc encore valables pour la sélection d'estimateurs.

La nouveauté vient des aspects quantitatifs. Si h_m et $h_{m'}$ sont de « bonnes » valeurs (et suffisamment différentes pour qu'il soit utile de savoir choisir entre les deux), alors $\mathcal{W}_2(h_m, h_{m'}, P) \ll \mathcal{W}_2(h_m, P)$: la variance des incréments est donc beaucoup plus

faible que la variance des critères, un phénomène déjà remarqué par Breiman et Spector (1992, sections 5.1 et 7.3) dans un autre cadre.

Par ailleurs, $n^{-1}\mathcal{W}_2(h_m, h_{m'}, P)$ est alors du même ordre de grandeur que $n^{-2}\mathcal{W}_1(h_m, h_{m'}, P)$. L'ordre de grandeur de la variance des incréments est donc aussi impacté par les valeurs de $C_1^{\mathrm{vf}}(V, n)$ ou $C_1^{\mathrm{MC}}(V, n, n_e)$. Ceci change la donne pour la validation croisée V-fold : lorsque V augmente, la variance des incréments diminue au premier ordre ! Cependant, cette diminution reste de l'ordre d'une constante numérique (au plus $C_1^{\mathrm{vf}}(2, n)$ qui tend vers 10 quand n tend vers l'infini). De plus, vu la formule donnant $C_1^{\mathrm{vf}}(V, n)$, il suffit de prendre $V = 5$ ou 10 pour que la variance soit très proche de son minimum. Nous renvoyons aux résultats de la fin de la section 3.3 pour discuter le cas Monte-Carlo et la comparaison avec le V-fold (à réinterpréter en tenant compte de la différence d'ordre de grandeur des $\mathcal{W}_i$ quand on considère les incréments). Les expériences numériques de Arlot et Lerasle (2016, section 6) permettent également de bien visualiser les ordres de grandeur des termes $\mathcal{W}_i$ dans des cas réalistes.

4.4 Analyse au deuxième ordre : surpénalisation

La deuxième partie de la section 4.1 met en évidence un deuxième facteur important à prendre en compte au deuxième ordre, en plus de la variance des incréments : le fait que « surpénaliser » un peu améliore souvent les performances.

Prenons le cas de l'estimation de densité par moindres carrés. Alors, en comparant la validation croisée V-fold et la validation croisée V-fold corrigée, Arlot et Lerasle (2016) montrent que la validation croisée V-fold surpénalise d'un facteur :

$$1 + \frac{1}{2(V-1)} \,.$$

Plus généralement, si l'on suppose (8) et (9) vérifiées avec $\gamma(m) = 2\beta(m)$ (comme c'est le cas en estimation de densité par moindres carrés), alors, la validation croisée avec un échantillon d'entraînement de taille n_e surpénalise d'un facteur :

$$\frac{1}{2}\left(1 + \frac{n}{n_e}\right) > 1 \,.$$

Lorsque surpénaliser est bénéfique, à variance constante, il est donc préférable de prendre $n_e \approx \kappa n$ avec $\kappa \in]0,1[$ fonction du facteur de surpénalisation optimal. Par exemple, si l'optimum est de surpénaliser d'un facteur $1,12$ (comme dans le cadre de la figure 1), on obtient que le mieux est d'avoir $n_e = n/1,24$, ce qui correspond au « 5-fold ». Si l'optimum est de surpénaliser d'un facteur $3/2$, le mieux est d'avoir $n_e = n/2$, ce qui correspond au « 2-fold ». Et lorsque le facteur de surpénalisation optimal est plus grand, il faut prendre n_e encore plus petit, ce qui est impossible avec une méthode « V-fold » !

Attention toutefois à ne pas oublier que ceci n'est correct que si l'on compare des critères de *même variance*. Si l'on augmente la variance pour surpénaliser (comme on le fait en diminuant V pour le V-fold), on gagne d'un côté mais on perd de l'autre. *In fine*, de nombreux cas de figure peuvent se produire quand on trace le risque du

prédicteur final obtenu par validation croisée V-fold en fonction de V : il peut décroître avec V, augmenter avec V, ou être minimal pour une valeur de V « intermédiaire », comme le montrent des simulations numériques (Arlot et Lerasle, 2016).

Si l'on veut y voir clair, il faut raisonner en changeant un seul facteur à la fois parmi le nombre V de découpages (qui influe sur la variance) et la taille n_e de l'échantillon d'entraînement (qui définit la surpénalisation, en influant assez peu sur la variance). Ceci peut se faire naturellement avec la validation croisée Monte-Carlo. Peut-on le faire avec une stratégie « V-fold », qui est un peu meilleure que la stratégie Monte-Carlo en termes de variance ? Oui, avec la pénalisation V-fold (Arlot, 2008; Arlot et Lerasle, 2016). Sans la définir précisément ici, décrivons-en le principe : il s'agit d'une méthode de pénalisation

$$\widehat{m}_C\big(D_n; (B_j)_{1\leqslant j\leqslant V}\big) \in \operatorname*{argmin}_{m\in\mathcal{M}_n}\Big\{\widehat{\mathcal{R}}_n\big(\widehat{f}_m(D_n)\big) + C\,\mathrm{pen}_{\mathrm{vf}}\big(m; D_n; (B_j)_{1\leqslant j\leqslant V}\big)\Big\}$$

définie pour une partition $(B_j)_{1\leqslant j\leqslant V}$ en V blocs de même taille, qui coïncide avec la validation croisée V-fold corrigée si $C = 1$. En estimation de densité par moindres carrés, elle coïncide avec la validation croisée V-fold si

$$C = 1 + \frac{1}{2(V-1)}\,.$$

Alors, on peut choisir pour C un « bon » facteur de surpénalisation (en supposant que l'on connaît une telle valeur), et ensuite on choisit V aussi grand que possible afin de minimiser la variance.

Remarque 13 (Pénalisation V-fold ou validation croisée Monte-Carlo ?) On a proposé deux stratégies pour découpler V et n_e : la validation croisée Monte-Carlo et la pénalisation V-fold. Laquelle choisir en pratique ? Au vu des calculs de variance d'Arlot et Lerasle (2016), la pénalisation V-fold semble préférable : pour une même valeur de V et de la constante de surpénalisation, on obtient le plus souvent une variance plus faible. Cependant, ceci est valable dans un cadre bien précis, et il n'est pas certain que cela soit toujours vrai hors de ce cadre. Pire encore, le fait que la pénalisation V-fold surpénalise bien d'un facteur C (en particulier, le fait qu'elle estime sans biais le risque lorsque $C = 1$) n'est vraisemblablement valable que si l'hypothèse (9) est vérifiée. Au final, il semble raisonnable d'utiliser la pénalisation V-fold pour la sélection parmi des prédicteurs « réguliers » — c'est-à-dire, susceptibles de vérifier (9), au moins approximativement. Dans les autres cas, mieux vaut utiliser la validation croisée Monte-Carlo (ou V-fold répétée, si c'est possible). Et quoi qu'il en soit, il ne faut pas oublier de bien choisir V et C ou n_e, selon le cas.

5 Conclusion

5.1 Choix d'une procédure de validation croisée

Comment choisir une procédure de validation croisée pour un problème donné ?

Tout d'abord, lorsque le temps de calcul disponible est limité, il faut prendre en compte la complexité algorithmique des procédures de validation croisée. Dans la plupart des cas, celle-ci est proportionnelle au nombre V de découpages (il faut mettre en œuvre V fois chaque règle $\widehat{f}_m$ considérée). Parfois, les critères de validation croisée peuvent être calculés plus efficacement, comme détaillé par Arlot et Celisse (2010, section 9).

Dans les cas « réguliers », comme l'estimation de densité par moindres carrés, les résultats décrits en section 4 suggèrent la stratégie suivante. Dans un premier temps, choisir un facteur de surpénalisation (1 si l'on veut estimer le risque ; souvent un peu plus pour un problème de sélection d'estimateurs). Si l'on utilise la validation croisée Monte-Carlo (ou le V-fold répété), ceci se fait en choisissant la taille n_e de l'échantillon d'entraînement. Si l'on utilise la pénalisation V-fold (définie en section 4.4, voir notamment la remarque 13), ceci se fait directement avec le paramètre C. Puis on choisit le nombre V de découpages, en le prenant aussi grand que possible pour optimiser la performance statistique, dans la limite des capacités de calcul. Les calculs de variance reportés en section 4.3 indiquent qu'il n'est pas nécessaire de prendre V très grand pour avoir une performance quasi optimale. Avec une précision importante : ceci n'est totalement vrai que si n_e est de l'ordre de κn avec $\kappa \in]0,1[$. Sinon (par exemple lorsque $n_e = n-1$), la variance de la validation simple peut être *beaucoup* plus grande que celle du leave-$(n-n_e)$-out, et alors il est nécessaire de prendre V très grand pour obtenir une variance du bon ordre de grandeur.

Si l'on s'impose d'utiliser la validation croisée V-fold, choisir V peut s'avérer plus délicat, car ce paramètre détermine simultanément le facteur de surpénalisation (via la taille $n(V-1)/V$ de l'échantillon d'entraînement) et le nombre de découpages. Pour la sélection d'estimateurs, en admettant que dans la plupart des cas il est bon de surpénaliser « un peu » (comme on le constate empiriquement), alors prendre V entre 5 et 10 est un très bon choix (voire optimal). En effet, on surpénalise alors d'un facteur entre $1,05$ et $1,12$ et l'on a une variance proche de sa valeur minimale possible. Cette fourchette de valeurs de V correspond d'ailleurs aux conseils classiques dans la littérature statistique[30]. Si l'objectif est d'estimer le risque d'un prédicteur, la situation est un peu différente car il faut choisir V le plus grand possible pour minimiser le biais (et la variance) : les formules donnant biais et variance en fonction de V permettent alors d'évaluer où se situe le meilleur compromis entre précision statistique et complexité algorithmique.

Dans le cas général, en particulier quand on s'intéresse à un ou plusieurs prédicteurs « instables », le comportement des procédures de validation croisée peut être différent, ce dont témoignent plusieurs résultats empiriques rapportés par Arlot et Celisse (2010, sections 5.2 et 8), en particulier ceux de Breiman (1996, section 7). Il semble alors préférable de réaliser des expériences numériques (par exemple, sur des données synthétiques) pour déterminer le comportement de la validation croisée en

30. C'est par exemple le conseil donné par Hastie *et al.* (2009, section 7.10.1). Les intuitions généralement proposées pour étayer ce conseil sont cependant différentes des arguments donnés dans ce texte, et parfois en contradiction avec les résultats théoriques rapportés dans ce texte. Par exemple, il est faux de dire que le 5-fold a *toujours* une variance plus faible que le leave-one-out.

fonction de n_e et V. Au vu des sections 3 et 4, trois quantités clé sont à étudier. D'une part, le risque moyen $\mathbb{E}[\mathcal{R}_P(\widehat{f}_m(D_n))]$ et sa dépendance en la taille n de l'échantillon permettent de comprendre le biais de la validation croisée (et son comportement au premier ordre pour la sélection d'estimateurs). D'autre part, pour prendre en compte la variance, la quantité à calculer dépend de l'objectif. Si l'on s'intéresse à l'estimation du risque, alors il faut calculer la variance de l'estimateur par validation croisée du risque d'une règle $\widehat{f}_m$. Si l'on s'intéresse à la sélection d'estimateurs, alors la quantité à calculer est

$$\mathrm{var}\Big(\widehat{\mathcal{R}}^{\mathrm{vc}}\big(\widehat{f}_m; D_n; (E_j)_{1\leqslant j\leqslant V}\big) - \widehat{\mathcal{R}}^{\mathrm{vc}}\big(\widehat{f}_{m^\circ}; D_n; (E_j)_{1\leqslant j\leqslant V}\big)\Big)$$
$$\text{où} \qquad m^\circ \in \operatorname*{argmin}_{m\in\mathcal{M}_n}\Big\{\mathbb{E}\Big[\mathcal{R}_P\big(\widehat{f}_m(D_n)\big)\Big]\Big\}$$

est le choix oracle et où m est « proche » de l'oracle. Notons que l'on peut s'épargner certains calculs de variance en utilisant la proposition 3 et la formule (13) en section 3.3, ainsi que leurs généralisations aux incréments d'estimateurs par validation croisée.

Remarque 14 (Objectif d'identification) On aboutit à des conclusions différentes quand on a pour objectif d'*identifier* le « vrai » modèle[31] ou la règle d'apprentissage dont l'excès de risque décroît le plus vite parmi $(\widehat{f}_m)_{m\in\mathcal{M}_n}$ (Arlot et Celisse, 2010, section 7). Par exemple, en régression linéaire par moindres carrés, la validation croisée choisit le vrai modèle avec probabilité 1 asymptotiquement si et seulement si $n_e \ll n$ (avec des hypothèses sur V, Shao, 1997). Ce phénomène se généralise à d'autres cadres et a été nommé « paradoxe de la validation croisée » par Yang (2006, 2007) : pour l'identification, plus on a d'observations à disposition, plus il faut en utiliser une fraction petite pour l'entraînement. On peut rapprocher ces résultats du fait qu'il est nécessaire de surpénaliser fortement pour une identification optimale, d'où le fait que BIC fonctionne, alors que AIC et C_p sont sous-optimales. Signalons enfin que pour l'identification, Yang conseille d'utiliser la variante « par vote majoritaire » de la validation croisée, définie en section 2.4.

5.2 Validation croisée ou procédure spécifique ?

Lorsqu'une autre procédure de sélection d'estimateurs est disponible — par exemple, C_p ou AIC —, faut-il la préférer à la validation croisée ?

Si l'on est dans le cadre spécifique pour lequel cette procédure a été construite (pour C_p, la régression linéaire homoscédastique avec le coût quadratique et des estimateurs des moindres carrés), alors c'est elle qu'il faut utiliser. Des expériences numériques montrent en effet que la validation croisée fonctionne alors souvent un peu moins bien que les procédures spécifiques. C'est le prix de l'« universalité » de la validation croisée.

En revanche, si l'on risque de sortir un peu du cadre où la procédure « spécifique » est connue pour être optimale (par exemple, pour C_p, si l'on soupçonne les données d'être hétéroscédastiques), alors il est plus sûr d'utiliser une procédure « universelle » comme la validation croisée (si les capacités de calcul disponibles le permettent).

31. Le vrai modèle est défini comme le plus petit modèle contenant $f^\star$, en supposant qu'il existe.

5.3 Limites de l'universalité

Il est bon de garder en tête que la validation croisée ne peut pas fonctionner parfaitement d'une manière totalement universelle, ne serait-ce qu'à cause des résultats « on n'a rien sans rien » (Arlot, 2018a, section 6). En particulier, la validation croisée suppose implicitement qu'il est possible d'évaluer le risque d'un prédicteur $f \in \mathcal{F}$ à partir de $n - n_e$ observations. Ce n'est clairement pas possible dans l'exemple construit pour démontrer le premier résultat « on n'a rien sans rien » de Arlot (2018a, théorème 3 en section 6).

Des problèmes se posent également lorsque les hypothèses explicitement faites par la validation croisée sont violées (données indépendantes et de même loi ; et pour la sélection d'estimateurs, la collection $(\widehat{f}_m)_{m \in \mathcal{M}_n}$ est supposée n'être « pas trop grande » (Arlot, 2018a, section 3.9)).

- Lorsque les données sont *dépendantes*, les échantillons d'entraînement et de validation ne sont plus nécessairement indépendants, ce qui peut induire un biais assez fort pour l'estimation du risque d'une règle d'apprentissage. Dans le cas de dépendance à courte portée, ce biais peut être évité en utilisant des échantillons d'entraînement D_n^E et de validation D_n^V tels que E et V sont suffisamment éloignés.
- Pour une série temporelle *non-stationnaire*, si l'on s'intéresse à la prévision du « futur » à partir du « passé », on ne peut pas utiliser la validation croisée telle quelle. En choisissant un échantillon d'entraînement dans le passé par rapport à l'échantillon de validation (et en les éloignant si besoin), on peut appliquer la validation simple, mais sans garantie en raison de la non-stationnarité. Si l'on a observé une série assez longue, on peut également utiliser une fenêtre glissante pour multiplier les couples entraînement/validation.
- En présence de *données aberrantes*, il est indispensable d'utiliser des prédicteurs robustes et/ou une fonction de coût robuste.
- Pour la sélection d'estimateurs parmi une *collection exponentielle* (Arlot, 2018a, section 3.9), le principe d'estimation sans biais du risque ne fonctionne plus et bon nombre de conclusions de ce texte sont erronées. Une idée est alors de « surpénaliser » fortement, en prenant un échantillon d'entraînement de petite taille (comme indiqué en section 4), ce qui permet d'avoir une procédure de la deuxième famille décrite par Arlot (2018a, section 3.9). Une deuxième stratégie — appelée « validation croisée en deux étapes » ou « double cross » (Stone, 1974) — est la suivante. D'abord, on forme un nombre polynomial de groupes de règles d'apprentissage. Puis, à l'intérieur de chaque groupe, on utilise la validation croisée pour sélectionner une règle ; on dispose donc d'une « méta-règle » associée à chaque groupe. Enfin, on sélectionne par validation croisée l'une de ces méta-règles (en suivant les conseils formulés à la fin de la section 2.1). Cette deuxième stratégie s'applique par exemple pour le problème de détection de ruptures.

Arlot et Celisse (2010, section 8) détaillent tous ces points et donnent des références bibliographiques.

6 Annexe : estimation de densité L^2

Cette section est une introduction rapide au cadre de l'estimation de densité L^2, suivie de quelques résultats en lien avec la validation croisée.
Référence principale : le livre de Tsybakov (2004) ou sa version anglaise (Tsybakov, 2009).
Voir aussi l'article d'Arlot et Lerasle (2016).

6.1 Cadre

Soit $(\mathcal{X}, \mathcal{A}, \mu)$ un ensemble mesuré. On observe $X_1, \dots, X_n \in \mathcal{X}$ des variables aléatoires indépendantes et de même loi P, et l'on cherche à estimer P. (Par rapport au cadre du cours, $\mathcal{Y} = \{1\}$ et $Y_i = 1$ ne contient aucune information.)

La mesure de référence μ étant fixée (par exemple, si $\mathcal{X} = \mathbb{R}^p$, le choix le plus courant est la mesure de Lebesgue), on suppose P absolument continue par rapport à μ, et l'on cherche à estimer la densité $s := \mathrm{d}P/\mathrm{d}\mu$, que l'on suppose appartenir à $L^2(\mu)$.

Dans ce chapitre, on note $\|\cdot\| = \|\cdot\|_{L^2(\mu)}$ la norme de $L^2(\mu)$, définie par

$$\forall f \in L^2(\mu), \qquad \|f\|^2 = \int f^2 \mathrm{d}\mu \,.$$

Comme ensemble des « prédicteurs » possibles, on pourrait prendre $\mathcal{F}_0$ l'ensemble des densités appartenant à $L^2(\mu)$. Dans la suite, on fait un autre choix, en considérant $\mathcal{F} = L^2(\mu)$. Le risque quadratique d'un élément f de $\mathcal{F}$ est défini par :

$$\mathcal{R}_P(f) = \mathbb{E}_{X\sim P}\left[\|f\|^2_{L^2(\mu)} - 2f(X)\right].$$

Proposition 4 *Si $s = \mathrm{d}P/\mathrm{d}\mu \in L^2(\mu)$, le risque $\mathcal{R}_P$ est minimal sur $\mathcal{F} = L^2(\mu)$ en $f^\star := s$. L'excès de risque de tout $f \in \mathcal{F}$ s'écrit :*

$$\mathcal{R}_P(f) - \mathcal{R}_P(f^\star) = \ell(f^\star, f) = \|f^\star - f\|^2_{L^2(\mu)} \,.$$

Démonstration Pour tout $f \in L^2(\mu)$,

$$\mathcal{R}_P(f) = \int (f^2 - 2fs)\mathrm{d}\mu = \|f\|^2_{L^2(\mu)} - 2\langle f, s\rangle_{L^2(\mu)} = \|f - s\|^2_{L^2(\mu)} - \|s\|^2_{L^2(\mu)}$$

dont on déduit les deux résultats. □

Remarques :
— La proposition 4 montre que $f^\star$ est l'unique minimiseur du risque sur $\mathcal{F}$.
— Pour toute densité f par rapport à μ, si $f \in L^\infty(\mu)$, alors on a que $f \in L^2(\mu)$ et $\|f\|^2_{L^2(\mu)} \leqslant \|f\|_\infty$.

— Si $f \in L^2(\mu)$, alors on a $\mathcal{R}_P(f_+) \leqslant \mathcal{R}_P(f)$, où f_+ désigne la partie positive de f : $f_+(x) = \max(f(x), 0)$ pour tout $x \in \mathcal{X}$. Donc, si l'on construit un estimateur f de $f^\star$ qui n'est pas positif, on fera toujours mieux en prenant sa partie positive. En revanche, on ne peut pas forcément se ramener à une densité : il n'est pas toujours vrai que renormaliser f_+ par $\int f_+ \mathrm{d}\mu$ fait diminuer le risque L^2 (et l'on pourrait même avoir $f_+ \notin L^1(\mu)$, lorsque $L^2(\mu) \not\subseteq L^1(\mu)$).

6.2 Estimateurs

L'exemple principal de ce cours est la famille des estimateurs des moindres carrés, définis comme suit. Soit $(\psi_\lambda)_{\lambda \in m}$ une famille orthonormée dans $L^2(\mu)$ et S_m l'espace vectoriel engendré. On définit le risque empirique (ou critère des moindres carrés) par :

$$\widehat{\mathcal{R}}_n(f) = \|f\|^2_{L^2(\mu)} - \frac{2}{n}\sum_{i=1}^n f(X_i) = \|f\|^2_{L^2(\mu)} - 2P_n(f)\,, \qquad \text{où} \qquad P_n := \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$$

désigne la mesure empirique. Notons que pour tout $f \in L^2(\mu)$,

$$\mathbb{E}\big[\widehat{\mathcal{R}}_n(f)\big] = \mathcal{R}_P(f)\,,$$

qui est la propriété que l'on attend pour un risque empirique. L'estimateur des moindres carrés associé au modèle S_m est alors défini par

$$\widehat{f}_m \in \operatorname*{argmin}_{f \in S_m} \widehat{\mathcal{R}}_n(f)\,.$$

Exemple 1 (Histogramme) Soit m une partition mesurable de $\mathcal{X}$ (ou d'un sous-ensemble de $\mathcal{X}$ de mesure pleine) telle que $\mu(\lambda) \in]0, +\infty[$ pour tout $\lambda \in m$. Si pour tout $\lambda \in m$ on pose $\psi_\lambda = \frac{1}{\mu(\lambda)^{1/2}}\mathbb{1}_\lambda$, alors on a une famille orthonormée qui engendre S_m, l'ensemble des fonctions de $L^2(\mu)$ qui sont constantes sur chaque élément de m. L'estimateur des moindres carrés associé est appelé *estimateur de la densité par histogramme*.

Exemple 2 (Histogramme régulier) Un estimateur par histogramme (exemple 1) est dit *régulier* lorsque $\mu(\lambda) = h > 0$ ne dépend pas de $\lambda \in m$.

Proposition 5 *Soit $(\psi_\lambda)_{\lambda \in m}$ une famille finie, orthonormée dans $L^2(\mu)$, et $\widehat{f}_m$ l'estimateur des moindres carrés associé. On a alors, pour μ presque tout $x \in \mathcal{X}$:*

$$\widehat{f}_m(x) = \sum_{\lambda \in m} \big(P_n(\psi_\lambda)\big) \times \psi_\lambda(x) = \frac{1}{n}\sum_{i=1}^n K_m(X_i, x)$$

$$\textit{avec} \qquad K_m(x,y) = \sum_{\lambda \in m} \psi_\lambda(x)\psi_\lambda(y)\,.$$

Démonstration Pour tout $f \in S_m = \mathrm{ev}\{\psi_\lambda \,/\, \lambda \in m\}$, on peut écrire

$$f = \sum_{\lambda \in m} f_\lambda \psi_\lambda$$

avec $f_\lambda \in \mathbb{R}$, et l'on a alors :

$$\|f\|^2_{L^2(\mu)} = \sum_{\lambda \in m} f_\lambda^2 \ .$$

Ainsi, le risque empirique

$$\begin{aligned}\widehat{\mathcal{R}}_n(f) &= \sum_{\lambda \in m} \big(f_\lambda^2 - 2P_n(\psi_\lambda) f_\lambda\big) \\ &= \sum_{\lambda \in m} \Big[\big(f_\lambda - P_n(\psi_\lambda)\big)^2 - \big(P_n(\psi_\lambda)\big)^2\Big]\end{aligned}$$

est bien minimal lorsque $f_\lambda = P_n(\psi_\lambda)$ pour tout $\lambda \in m$ (si ce choix est possible). L'hypothèse garantit que ce choix est possible : la somme étant finie, on obtient bien un élément de $L^2(\mu)$. On en déduit la première formule. La seconde en découle directement. □

6.3 Erreur d'approximation et erreur d'estimation

Soit

$$f_m^\star \in \operatorname*{argmin}_{f \in S_m} \mathcal{R}_P(f) = \operatorname*{argmin}_{f \in S_m} \|f - f^\star\|^2_{L^2(\mu)} \ .$$

La deuxième formule (qui découle de la formule de l'excès de risque, proposition 4) indique que $f_m^\star$ existe et est la projection orthogonale de $f^\star$ sur S_m dans $L^2(\mu)$ (unique dans $L^2(\mu)$, car on parle de fonctions définies à un ensemble de mesure nulle près).

On a alors, pour μ presque tout $x \in \mathcal{X}$:

$$f_m^\star(x) = \sum_{\lambda \in m} \big(P(\psi_\lambda)\big)\psi_\lambda(x)\,. \tag{18}$$

Démonstration Exercice (s'inspirer de la démonstration de la proposition 5). □

Remarquons que l'on a donc, au vu de la proposition 5,

$$f_m^\star(x) = \mathbb{E}\big[\widehat{f}_m(x)\big]$$

pour μ presque tout $x \in \mathcal{X}$.

6.3.1 Décomposition générale du risque

On peut utiliser $f_m^\star$ pour décomposer le risque de $\widehat{f}_m$ en erreur d'approximation et erreur d'estimation (voir le chapitre « fondamentaux »). En effet, $f_m^\star$ est définie comme la projection orthogonale de $f^\star$ sur S_m dans $L^2(\mu)$. Par conséquent, pour tout $f \in S_m$,

$$\begin{aligned}\ell(f^\star, f) &= \|f^\star - f\|_{L^2(\mu)}^2 \\ &= \|f^\star - f_m^\star\|_{L^2(\mu)}^2 + \|f_m^\star - f\|_{L^2(\mu)}^2\end{aligned}$$

puisque $f_m^\star$ est la projection orthogonale de $f^\star$ sur S_m dans $L^2(\mu)$.

En particulier, pour $f = \widehat{f}_m$, on obtient que

$$\ell(f^\star, \widehat{f}_m) = \underbrace{\|f^\star - f_m^\star\|_{L^2(\mu)}^2}_{=\text{ erreur d'approximation}=\ell(f^\star, f_m^\star)} + \underbrace{\|f_m^\star - \widehat{f}_m\|_{L^2(\mu)}^2}_{=\text{erreur d'estimation}}$$

6.3.2 Erreur d'estimation

Précisons ce que vaut l'erreur d'estimation (puis son espérance). D'après les formules précédemment obtenues pour $\widehat{f}_m$ et $f_m^\star$ (proposition 5 et équation (18)), on a :

$$\begin{aligned}\|f_m^\star - \widehat{f}_m\|_{L^2(\mu)}^2 &= \sum_{\lambda \in m} \big[P(\psi_\lambda) - P_n(\psi_\lambda)\big]^2 \\ &= \frac{1}{n^2} \sum_{\lambda \in m} \left[\sum_{i=1}^n \big(\psi_\lambda(X_i) - P(\psi_\lambda)\big)\right]^2 .\end{aligned}$$

On en déduit que

$$\mathbb{E}\Big[\|f_m^\star - \widehat{f}_m\|_{L^2(\mu)}^2\Big] = \frac{1}{n} \sum_{\lambda \in m} \operatorname{var}\big(\psi_\lambda(X_1)\big)$$

et donc

$$\mathbb{E}\Big[\ell(f^\star, \widehat{f}_m)\Big] = \|f^\star - f_m^\star\|_{L^2(\mu)}^2 + \frac{1}{n} \sum_{\lambda \in m} \operatorname{var}\big(\psi_\lambda(X_1)\big) . \tag{19}$$

6.3.3 Exemple

Exemple 3 (Histogramme régulier, suite) Dans le cas d'un histogramme régulier (exemple 2) avec $\mu(\lambda) = h$ pour tout $\lambda \in m$, on a d'une part,

$$\sum_{\lambda \in m} \mathbb{E}\big[\psi_\lambda^2(X_1)\big] = \mathbb{E}\left[\sum_{\lambda \in m} \psi_\lambda^2(X_1)\right] = \mathbb{E}\left[\frac{1}{h}\right] = \frac{1}{h} .$$

D'autre part,

$$\sum_{\lambda \in m} \mathbb{E}\big[\psi_\lambda(X_1)\big]^2 = \sum_{\lambda \in m} (P\psi_\lambda)^2 = \|f_m^\star\|^2 \leqslant \|f^\star\|^2 ,$$

où l'inégalité provient du fait que la norme d'une projection orthogonale est inférieure à la norme. On en déduit la formule

$$\mathbb{E}\Big[\|f_m^\star - \widehat{f}_m\|^2\Big] = \frac{1}{n}\left(\frac{1}{h} - \|f_m^\star\|^2\right)$$

puis l'encadrement

$$\frac{1}{n}\left(\frac{1}{h} - \|f^\star\|^2\right) \leqslant \mathbb{E}\Big[\|f_m^\star - \widehat{f}_m\|^2\Big] \leqslant \frac{1}{nh}\,.$$

Le plus souvent, on choisit $h = h_n \to 0$ lorsque $n \to +\infty$, auquel cas l'on obtient :

$$\mathbb{E}\Big[\|f_m^\star - \widehat{f}_m\|^2\Big] \sim_{n\to+\infty} \frac{1}{nh_n}\,.$$

Remarquons que si $\mu(\mathcal{X}) = 1$ — par exemple, si $\mathcal{X} = [0,1]^p$ et μ est la mesure de Lebesgue —, pour un histogramme régulier, on a $h\,\mathrm{Card}(m) = 1$ et donc $1/h = \mathrm{Card}(m)$ est égal au nombre de paramètres. Ainsi, l'espérance de l'erreur d'estimation est (approximativement) égale au nombre de paramètres divisé par le nombre d'observations.

6.3.4 Conséquence pour la validation croisée

L'équation (19) montre que l'espérance du risque d'un estimateur des moindres carrés $\widehat{f}_m$ entraîné avec n observations est de la forme $\alpha(m) + \beta(m)/n$. La règle d'apprentissage $\widehat{f}_m$ est donc une règle intelligente, et l'hypothèse (8) de la section 3.1 est vérifiée. On peut donc préciser la valeur du biais de la validation croisée (voir la section 3.1).

6.4 Pénalité idéale

Les calculs précédents permettent d'obtenir une formule simple pour la pénalité idéale

$$\mathrm{pen}_{\mathrm{id}}(m) := \mathcal{R}_P(\widehat{f}_m) - \widehat{\mathcal{R}}_n(\widehat{f}_m)$$

et pour son espérance, qui est évoquée à la parenthèse 8 en section 3.2, ainsi qu'au chapitre « fondamentaux » du cours (Arlot, 2018a, section 3.9). En effet,

$$\begin{aligned}
(\mathcal{R}_P - \widehat{\mathcal{R}}_n)\big(\widehat{f}_m(D_n)\big) &= 2(P_n - P)(\widehat{f}_m) \\
&= 2(P_n - P)(\widehat{f}_m - f_m^\star) + 2(P_n - P)(f_m^\star) \\
&= 2\sum_{\lambda\in m}\big((P_n - P)(\psi_\lambda)\big)^2 + 2(P_n - P)(f_m^\star) \\
&= 2\|\widehat{f}_m - f_m^\star\|^2 + \underbrace{2(P_n - P)(f_m^\star)}_{\text{centré}}
\end{aligned}$$

d'après la proposition 5 et l'équation (18). Par conséquent, au vu de la section 6.3.2

$$\mathbb{E}\Big[(\mathcal{R}_P - \widehat{\mathcal{R}}_n)\big(\widehat{f}_m(D_n)\big)\Big] = 2\mathbb{E}\Big[\|\widehat{f}_m - f_m^\star\|^2\Big] = \frac{2}{n}\sum_{\lambda\in m}\mathrm{var}\big(\psi_\lambda(X_1)\big)\,.$$

Ainsi, l'espérance de la pénalité idéale est de la forme $\gamma(m)/n$, et l'hypothèse (9) — énoncée en section 3.2 — est bien vérifiée.

7 Annexe : exercices

Exercice 1 On se place en classification 0–1. Soit $\widehat{f}$ la règle par partition associée à la partition triviale $\mathcal{A} = \{\mathcal{X}\}$; autrement dit, $\widehat{f}((x_i, y_i)_{1\leqslant i\leqslant n}; x)$ réalise un vote majoritaire parmi les y_i, sans tenir compte des x_i ni de x.
Démontrer que $\widehat{f}$ n'est pas intelligente. En déduire qu'aucune règle par partition (sur une partition $\mathcal{A}$ fixe quand n varie) n'est intelligente.

L'exercice 2 ci-dessous montre qu'on peut rendre les règles par partition intelligentes en les modifiant légèrement.

Exercice 2 On se place en classification 0–1. Soit $\widetilde{f}$ la règle par partition associée à la partition triviale $\mathcal{A} = \{\mathcal{X}\}$, avec une décision « randomisée » dans les cas d'égalité : s'il y a exactement $n/2$ des y_i qui sont égaux à 1, $\widetilde{f}\big((x_i, y_i)_{1\leqslant i\leqslant n}; x\big)$ vaut 1 avec probabilité 1/2 et vaut 0 avec probabilité 1/2. La notion de règle intelligente s'étend naturellement à $\widetilde{f}$ en prenant aussi l'espérance sur la randomisation interne $\widetilde{f}$ dans la définition de son risque moyen.
Démontrer que $\widetilde{f}$ est intelligente. En déduire que tout règle par partition (sur une partition $\mathcal{A}$ fixe quand n varie), randomisée de la même manière en cas d'égalité, est intelligente.

Exercice 3 Démontrer la remarque 11 en section 3.3. En particulier, si $(B_j^\ell)_{1\leqslant j\leqslant V}$, $\ell \in \{1, \dots, L\}$, est une suite de partitions de $\{1, \dots, n\}$, indépendantes et de même loi uniforme sur l'ensemble des partitions de $\{1, \dots, n\}$ en V blocs de même taille, indépendante de D_n, démontrer la formule suivante pour la variance de l'estimateur par validation croisée V-fold répétée du risque d'une règle d'apprentissage $\widehat{f}_m$:

$$\begin{aligned}
&\mathrm{var}\Big(\widehat{\mathcal{R}}^{\mathrm{vc}}\Big(\widehat{f}_m; D_n; \big((B_j^\ell)^c\big)_{1\leqslant j\leqslant V, 1\leqslant \ell\leqslant L}\Big)\Big) \\
&= \mathrm{var}\Big(\widehat{\mathcal{R}}^{\mathrm{lpo}}\Big(\widehat{f}_m; D_n; \frac{n}{V}\Big)\Big) \\
&\quad + \frac{1}{L}\Big[\underbrace{\mathrm{var}\Big(\widehat{\mathcal{R}}^{\mathrm{vf}}\big(\widehat{f}_m; D_n; (B_j^1)_{1\leqslant j\leqslant V}\big)\Big) - \mathrm{var}\Big(\widehat{\mathcal{R}}^{\mathrm{lpo}}\Big(\widehat{f}_m; D_n; \frac{n}{V}\Big)\Big)}_{\geqslant 0}\Big]\,.
\end{aligned}$$

Exercice 4 Démontrer la formule (13) donnant la variance du critère par validation simple.

Remerciements

Cet texte fait suite à un cours donné dans le cadre des Journées d'Études en Statistique 2016. Il s'agit d'une version préliminaire du chapitre 3 du livre *Apprentissage statistique et données massives*, édité par Frédéric Bertrand, Myriam Maumy-Bertrand, Gilbert Saporta et Christine Thomas-Agnan, paru aux éditions Technip (Arlot, 2018b).

Je remercie vivement tou(te)s les collègues avec qui j'ai travaillé sur ce sujet, en particulier mes coauteurs Alain Celisse, Matthieu Lerasle et Nelo Magalhães. Je remercie également Matthieu Lerasle pour avoir relu ce texte, ainsi que les étudiant(e)s qui ont suivi mon cours de master 2 « apprentissage statistique et rééchantillonnage » à l'Université Paris-Sud puis Paris-Saclay et les participant(e)s des JES 2016 pour leurs questions et commentaires.

Références

Sylvain Arlot : V-fold cross-validation improved : V-fold penalization, février 2008. arXiv :0802.0566v2.

Sylvain Arlot : Fondamentaux de l'apprentissage statistique. *In* Myriam Maumy-Bertrand, Gilbert Saporta et Christine Thomas-Agnan, éditeurs : *Apprentissage statistique et données massives*, pages 17–139. Éditions Technip, Paris, 2018a. Version préliminaire disponible à l'adresse `http://hal.archives-ouvertes.fr/hal-01485506`.

Sylvain Arlot : Validation croisée. *In* Myriam Maumy-Bertrand, Gilbert Saporta et Christine Thomas-Agnan, éditeurs : *Apprentissage statistique et données massives*, pages 141–174. Éditions Technip, Paris, 2018b. Version préliminaire disponible à l'adresse `http://hal.archives-ouvertes.fr/hal-01485508`.

Sylvain Arlot et Alain Celisse : A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4:40–79, 2010.

Sylvain Arlot et Matthieu Lerasle : Choice of V for V-fold cross-validation in least-squares density estimation. *Journal of Machine Learning Research (JMLR)*, 17(208):1–50, 2016.

Sylvain Arlot et Pascal Massart : Data-driven calibration of penalties for least-squares regression. *Journal of Machine Learning Research (JMLR)*, 10:245–279 (electronic), 2009.

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

Leo Breiman : Heuristics of instability and stabilization in model selection. *The Annals of Statistics*, 24(6):2350–2383, 1996.

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

Leo Breiman et Philip Spector : Submodel Selection and Evaluation in Regression. The X-Random Case. *International Statistical Review*, 60(3):291–319, 1992.

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.

Luc P. Devroye, László Györfi et Gábor Lugosi : *A Probabilistic Theory of Pattern Recognition*, volume 31 de *Applications of Mathematics (New York)*. Springer-Verlag, New York, 1996.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold et Aaron Roth : The reusable holdout : preserving validity in adaptive data analysis. *Science*, 349(6248):636–638, 2015.

Bradley Efron : How biased is the apparent error rate of a prediction rule ? *Journal of the American Statistical Association*, 81(394):461–470, 1986.

Gene H. Golub, Michael Heath et Grace Wahba : Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223, 1979.

Trevor Hastie, Robert Tibshirani et Jerome Friedman : *The Elements of Statistical Learning*. Springer Series in Statistics. Springer, New York, second édition, 2009. Data Mining, Inference, and Prediction.

Yoonsuh Jung et Jianhua Hu : A K-fold averaging cross-validation procedure. *Journal of Nonparametric Statistics*, 27(2):167–179, 2015.

Tammo Krueger, Danny Panknin et Mikio Braun : Fast cross-validation via sequential testing. *Journal of Machine Learning Research*, 16:1103–1155, 2015.

Matthieu Lerasle : Optimal model selection in density estimation. *Annales de l'Institut Henri Poincaré Probabilités et Statistiques*, 48(3):884–908, 2012.

Guillaume Maillard, Sylvain Arlot et Matthieu Lerasle : Cross-validation improved by aggregation : Agghoo, septembre 2017. arXiv :1709.03702.

W.H. Press, S.A. Teukolsky, W.T. Vetterling et B.P. Flannery : *Numerical Recipes in C*. Cambridge University Press, deuxième édition, 1992.

Adrien Saumard et Fabien Navarro : Model selection as a multiple testing procedure : Improving Akaike's information criterion, mars 2018. arXiv :1803.02078.

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

Mervyn Stone : Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 111–147, 1974.

Alexandre B. Tsybakov : *Introduction à l'estimation non-paramétrique*, volume 41 de *Mathématiques & Applications*. Springer-Verlag, Berlin, 2004.

Alexandre B. Tsybakov : *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York, 2009. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats.

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

Yuhong Yang : Consistency of cross validation for comparing regression procedures. *The Annals of Statistics*, 35(6):2450–2473, 2007.

Index