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

Premier cours: Apprentissage statistique et sélection d'estimateurs

SYLVAIN ARLOT
CNRS – ÉQUIPE SIERRA

Table des matières

Date: 10 Janvier 2011.

1. Le problème de l'apprentissage statistique

On utilise ici le cadre général décrit dans [3].

1.1. **Cadre général.** On observe des variables aléatoires $\xi_1, \ldots, \xi_n \in \Xi$ de loi commune P (les observations), supposées indépendantes. L'objectif est d'estimer à l'aide de l'échantillon $D_n = (\xi_i)_{1 \leq i \leq n}$ une quantité-cible $s^\star$ qui dépend de la distribution inconnue P, par exemple, la densité de P relativement à une mesure de référence μ, ou la fonction de régression associée à P.

Soit $\mathbb{S}$ l'ensemble des valeurs possibles pour $s^\star$. La qualité d'un candidat $t \in \mathbb{S}$ à estimer $s^\star$ est mesurée par sa *perte* $\mathcal{L}(t)$, où $\mathcal{L} : \mathbb{S} \mapsto \mathbb{R}$ est appelée *fonction de perte* (loss function). On suppose que la fonction de perte est minimale en $t = s^\star$.

Plusieurs fonctions de perte peuvent être considérées pour un problème statistique donné. Dans ce cours, on se limitera au cas important où la fonction de perte s'écrit

$$\mathcal{L}(t) = \mathcal{L}_P(t) = P\gamma(t) := \mathbb{E}_{\xi \sim P}[\gamma(t;\xi)] \ , \tag{1}$$

où $\gamma : \mathbb{S} \times \Xi \mapsto [0, \infty[$ est appelée *contraste*. Pour tout $t \in \mathbb{S}$, $P\gamma(t)$ mesure l'"incohérence" moyenne entre t et une nouvelle observation ξ de loi P.

Étant donnée la fonction de perte, on définit la *perte relative*

$$\ell(s^\star, t) := \mathcal{L}_P(t) - \mathcal{L}_P(s^\star) \geq 0 \ .$$

1.2. **Exemple : prédiction.** Le problème de la prédiction a pour objectif de «prédire» une quantité d'intérêt $Y \in \mathcal{Y}$ à partir d'une variable explicative $X \in \mathcal{X}$ et d'un échantillon $(X_1, Y_1), \ldots, (X_n, Y_n)$. Autrement dit, $\Xi = \mathcal{X} \times \mathcal{Y}$, $\mathbb{S}$ est l'ensemble des applications mesurables $\mathcal{X} \mapsto \mathcal{Y}$ et le contraste $\gamma(t;(x,y))$ mesurs une distance entre y et la valeur prédite $t(x)$. Deux cas particuliers classiques de ce problèmes sont la régression et la classification, détaillés ci-dessous.

1.3. **Exemple : régression.** On parle de régression lorsque $\mathcal{Y}$ est un ensemble continu, c'est-à-dire $\mathcal{Y} \subset \mathbb{R}$ (ou $\mathbb{R}^k$ en régression multivariée), avec le plus souvent $\mathcal{X} \subset \mathbb{R}^\ell$. Soit η la fonction de régression, c'est-à-dire, $\eta(x) = \mathbb{E}_{(X,Y)\sim P}[Y \mid X = x]$, de telle sorte que

$$\forall i \in \{1, \ldots, n\} \ , \quad Y_i = \eta(X_i) + \varepsilon_i \qquad \text{avec} \quad \mathbb{E}[\varepsilon_i \mid X_i] = 0 \ .$$

Un contraste souvent utilisé en régression est le *contraste des moindres carrés* $\gamma(t;(x,y)) = (t(x) - y)^2$, qui atteint son minimum sur $\mathbb{S}$ en $t = s^\star = \eta$, et la perte relative vaut

$$\ell(s^\star, t) = \mathbb{E}_{(X,Y)\sim P}\left[(s^\star(X) - t(X))^2\right] \ . \tag{2}$$

On peut remarquer que l'excès de risque de t est la carré de la distance $L^2(P)$ entre t et $s^\star$, si bien que *prédiction* et *estimation* sont ici des objectifs équivalents.

Preuve de (2). Soit $t \in \mathbb{S}$, (X, Y) une v.a. de loi P et $\varepsilon = Y - \eta(X)$. Alors, par définition de η, $\mathbb{E}[\varepsilon \mid X] = 0$ p.s., et donc

$$P\gamma(t) = \mathbb{E}\left[(\varepsilon + \eta(X) - t(X))^2\right] = \mathbb{E}\left[\varepsilon^2\right] + \mathbb{E}\left[(\eta(X) - t(X))^2\right] + 2\mathbb{E}\left[\varepsilon\,(\eta(X) - t(X))\right]$$

et le dernier terme est nul car, P-presque sûrement,

$$\mathbb{E}\left[\varepsilon\,(\eta(X) - t(X))\mid X\right] = (\eta(X) - t(X))\,\mathbb{E}\left[\varepsilon \mid X\right] = 0 \ .$$

La perte $P\gamma(t)$ est donc minimale pour $t = \eta = s^\star$, et la perte relative a bien l'expression donnée par (2). □

1.4. **Exemple alternatif : régression sur un plan d'expérience fixe.** On considère également souvent le cadre de la régression en supposant que le plan d'expérience (en anglais, *design*) $X_1, \ldots, X_n$ est déterministe. On le note alors souvent en lettres minuscules $x_1, \ldots, x_n$ par convention.

L'échantillon se résume alors à $Y = (Y_1, \ldots, Y_n) \in \mathbb{R}^n$ avec

$$\forall i \in \{1, \ldots, n\}\ , \quad Y_i = \eta(x_i) + \varepsilon_i$$

avec $\varepsilon_1, \ldots, \varepsilon_n$ des v.a. indépendantes, centrées. Lorsque les ε_i sont i.i.d., on parle de régression *homoscédastique* ; le cas contraire est appelé *hétéroscédastique*.

Si les Y_i ne sont plus de même loi, on peut tout de même considérer le risque et l'excès de risque correspondant à la loi $P = \mathcal{L}(X_{n+1}, Y_{n+1})$ définie par $X_{n+1} \sim \mathcal{U}(x_1, \ldots, x_n)$ et $\mathcal{L}(Y_{n+1} \mid X_{n+1} = x_i) = \mathcal{L}(Y_i)$. En notant $F = (F_i)_{1 \le i \le n} = (\eta(x_i))_{1 \le i \le n}$ et $\varepsilon = (\varepsilon_i)_{1 \le i \le n}$, on peut alors résumer le problème à

$$Y = F + \varepsilon \in \mathbb{R}^n$$

L'équivalent du contraste des moindres carrés est alors de chercher $t \in \mathbb{S} = \mathbb{R}^n$ tel que la perte quadratique

$$P\gamma(t) = \mathbb{E}_Y\left[\frac{1}{n}\|t - Y\|^2\right] = \mathbb{E}_Y\left[\frac{1}{n}\sum_{i=1}^{n}(t_i - Y_i)^2\right]$$

est minimale. Or,

$$\begin{aligned}
\mathbb{E}_Y\left[\frac{1}{n}\|t - Y\|^2\right] &= \mathbb{E}_Y\left[\frac{1}{n}\|t - F - \varepsilon\|^2\right] \\
&= \frac{1}{n}\mathbb{E}_Y\left[\|t - F\|^2 + \|\varepsilon\|^2 - 2\langle\varepsilon,\, t - F\rangle\right] \\
&= \frac{1}{n}\mathbb{E}_Y\left[\|t - F\|^2 + \|\varepsilon\|^2\right]
\end{aligned}$$

TAB. 1. Régression : design fixe vs. design aléatoire.

	Design aléatoire	Design fixe
Données	$(X_i, Y_i)_{1\le i\le n}$ i.i.d. $\sim P$	$Y = F + \varepsilon \in \mathbb{R}^n$
$\mathcal{D}(X_{n+1}, Y_{n+1})$	P	$\mathcal{D}(x_J, Y_J)$ avec $J \sim \mathcal{U}(1, \dots, n)$
$t \in \mathbb{S}$	$t : \mathcal{X} \to \mathbb{R}$	$t \in \mathbb{R}^n$
Perte de t	$P\gamma(t) = \mathbb{E}_{(X,Y)\sim P}\left[(Y - t(X))^2 \right]$	$\mathbb{E}_Y\left[\frac{1}{n}\|Y-t\|^2 \right] = \mathbb{E}_Y\left[\frac{1}{n}\sum_{i=1}^n (Y_i - t_i)^2 \right]$
Cible $s^\star$	$\eta : x \to \mathbb{E}[Y \mid X = x]$	$F = (\eta(x_1), \dots, \eta(x_n))$
Perte relative $\ell(s^\star, t)$	$\mathbb{E}_{(X,Y)\sim P}\left[(t(X) - \eta(X))^2 \right]$	$\frac{1}{n}\|F-t\|^2 = \frac{1}{n}\sum_{i=1}^n (\eta(x_i) - t_i)^2$
Risque empirique de t	$P_n\gamma(t) = \frac{1}{n}\sum_{i=1}^n (Y_i - t(X_i))^2$	$\|Y-t\|^2 = \frac{1}{n}\sum_{i=1}^n (Y_i - t_i)^2$

si bien que la perte quadratique est minimale pour $t = F = s^\star$ et la perte relative vaut

$$\ell(s^\star, t) = \frac{1}{n}\|t - F\|^2 = \frac{1}{n}\sum_{i=1}^n (t_i - F_i)^2 \quad .$$

Le Tableau 1 résume les correspondances entre quantités-clés entre les cadres de régression à désign déterministe et aléatoire.

La différence entre considérer les X_i aléatoires ou déterministes est souvent assez faible (même si les philosophies sous-jacentes et les objectifs sont assez différents), sauf si l'on pose des hypothèses trop fortes sur le design (e.g., ne considérer que $x_i = i/n$, qui est un cas trop éloigné de X_i i.i.d.). Voir par exemple [5, 6] où l'on montre un résultat à design fixe pour en déduire un à design aléatoire.

Du point de vue de l'analyse théorique, on choisit celui des deux qui simplifie le plus l'analyse. Ainsi, nous nous focaliserons surtout sur le cas du design fixe dans ce cours et dans le suivant, tandis que nous supposerons essentiellement le design aléatoire lors des deux derniers cours (ce qui permet de traiter plus simplement l'hétéroscédasticité et les méthodes de rééchantillonnage).

1.5. **Autres exemples.** Pour une présentation formelle d'autres exemples (estimation de densité, classification), nous renvoyons à la Section 1.2 de [3]. Voir aussi [12] pour un survol des résultats récents en classification.

2. ESTIMATEURS

2.1. **Définition générale.** On définit un *algorithme statistique* (ou *règle d'apprentissage*) $\mathcal{A}$ comme étant une application mesurable $\mathcal{A} : \bigcup_{n\in\mathbb{N}} \Xi^n \mapsto \mathbb{S}$. Étant donné un échantillon $D_n = (\xi_i)_{1\le i\le n} \in \Xi^n$, la sortie de $\mathcal{A}$, $\mathcal{A}(D_n) = \widehat{s}^{\mathcal{A}}(D_n) \in \mathbb{S}$, est un estimateur de $s^\star$. La qualité de $\mathcal{A}$ est alors mesurée par sa perte $\mathcal{L}_P\left(\widehat{s}^{\mathcal{A}}(D_n)\right)$, que l'on souhaite minimiser.

On définit alors le *risque* de $\mathcal{A}$ pour un échantillon de taille n et de loi P par

$$\mathbb{E}_{\xi\sim P^{\otimes n}}\left[P\gamma\left(\widehat{s}^{\mathcal{A}}(D_n)\right)\right] = R(\mathcal{A}, P, n) \quad .$$

Du point de vue statistique, on parle plutôt de l'estimateur $\widehat{s} = \widehat{s}(D_n) = \widehat{s}^{\mathcal{A}}(D_n) = \mathcal{A}(D_n) \in \mathbb{S}$, en oubliant sa relation fonctionnelle $\mathcal{A}$ avec l'échantillon D_n.

L'objectif est de construire $\widehat{s}(D_n) \in S$ tel que $\ell\,(\,s^\star, \widehat{s}(D_n)\,)$ est petit en espérance, ou avec grande probabilité.

2.2. **Consistance, No Free Lunch.** Nous renvoyons à [13] à propos des différentes notions de consistance et des résultats du type «No Free Lunch Theorems». Voir aussi [1].

2.3. **Exemples : Estimateurs par minimum de contraste.** La famille des *estimateurs par minimum de contraste* est classique en statistique. Étant donné un *modèle* $S \subset \mathbb{S}$, on dit que $\widehat{s}(D_n)$ est un estimateur par minimum de contraste sur S s'il minimise p.s. sur S le contraste empirique

$$t \mapsto P_n\gamma(t) = \frac{1}{n}\sum_{i=1}^{n} \gamma\,(t;\xi_i\,) \qquad \text{où} \quad P_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{\xi_i}\ .$$

L'heuristique sous-jacente est que l'espérance du contraste empirique $P_n\gamma(t)$ est $P\gamma(t)$, qui est minimale pour $t = s^\star$. On peut donc espérer que minimiser $P_n\gamma(t)$ sur $S \subset \mathbb{S}$ fournit un bon estimateur de $s^\star$, pourvu que S soit bien choisi.

Outre les estimateurs des moindres carrés (voire la sous-section suivante), un exemple classique est l'estimateur du maximum de vraisemblance en estimation de densité, qui correspond au contraste «log-vraisemblance» $\gamma(t;x) = -\ln(t(x))$.

Pour le problème de la prédiction, un estimateur minimisant le contraste empirique $P_n\gamma(t)$ est plus couramment appelé *minimiseur du risque empirique* [21].

2.4. **Exemple : Estimateurs des moindres carrés.** On prend $\gamma(t;(x,y)) = (t(x)-y)^2$ en régression (on pourrait aussi prendre le contraste des moindres carrés en estimation de densité), et le modèle S est un sous-espace vectoriel de $\mathbb{S}$. Par exemple, si S est l'espace des fonctions constantes par morceaux associé à une partition fixée de $\mathcal{X}$, alors on obtient un *régressogramme*. On peut aussi construire un modèle comme le sous-espace vectoriel généré par les premiers éléments d'une base telle que la base de Fourier ou une base d'ondelettes.

2.5. **Autres exemples d'estimateurs.** Pour les autres exemples d'estimateurs présentés à l'oral, nous renvoyons aux références suivantes :

- pour la régression ridge (à noyau) et les splines de lissage : [22, 18, 2],
- pour les SVM : [18, 11],
- pour le Lasso et autres méthodes de régularisation L^1 : [19, 25, 4],
- pour les k-plus proches voisins : [13].

3. Sélection d'estimateurs

Un grand nombre d'estimateurs peuvent être utilisés pour résoudre un problème statistique donné. Soit $(\mathcal{A}_m)_{m\in\mathcal{M}}$ une famille d'algorithmes statistiques et $(\widehat{s}_m(D_n))_{m\in\mathcal{M}}$ les estimateurs correspondants. Le problème de *sélection d'estimateur* (ou sélection d'algorithme statistique) est celui de choisir, à l'aide des données uniquement, l'un de ces algorithmes, c'est-à-dire choisir $\widehat{m}(D_n) \in \mathcal{M}$. L'estimateur final de $s^\star$ est alors $\widehat{s}_{\widehat{m}(D_n)}(D_n)$. La difficulté principale est alors que l'on ne dispose pas de données indépendantes de D_n pour évaluer la perte de chacun de $\widehat{s}_m(D_n)$: il faut utiliser D_n à nouveau.

Un exemple fondamental de ce problème est la sélection de modèles, où l'on dispose d'une famille $(S_m)_{m\in\mathcal{M}_n}$ de modèles, et pour tout $m \in \mathcal{M}_n$, $\widehat{s}_m$ est un estimateur par minimum de contraste qui lui est associé (par exemple, l'estimateur des moindres carrés en régression). Le problème de la sélection de modèles a fait l'objet de nombreux travaux, voir notamment [16].

Tout d'abord, précisons l'objectif visé par une procédure de choix d'estimateur : estimation ou identification (voir aussi à ce sujet les sections 2.2 à 2.4 de [3]).

3.1. Sélection d'estimateurs pour l'estimation.

On parle d'objectif d'*estimation* (ou de prédiction) lorsque l'on souhaite minimiser la perte relative $\ell\left(s^\star, \widehat{s}_{\widehat{m}}(D_n)\right)$ de l'estimateur final. Le choix optimal est alors appelé *oracle*, défini par

$$m^\star := m^\star(D_n) \in \arg\min_{m\in\mathcal{M}_n} \left\{\ell\left(s^\star, \widehat{s}_m(D_n)\right)\right\} \quad . \tag{3}$$

L'oracle $m^\star(D_n)$ dépendant de la distribution inconnue P des données, on ne peut espérer en général sélectionner exactement l'oracle avec grande probabilité. On se fixe donc un objectif plus atteignable, qui est de faire presque aussi bien que l'oracle en termes de perte relative. Selon le cadre, on peut le formaliser de différentes manières.

D'une part, dans un cadre asymptotique, on dit qu'une procédure de choix d'estimateur $\widehat{m}(\cdot)$ est *efficace* ou *asymptotiquement optimale* lorsque

$$\frac{\ell\left(s^\star, \widehat{s}_{\widehat{m}(D_n)}(D_n)\right)}{\inf_{m\in\mathcal{M}_n}\left\{\ell\left(s^\star, \widehat{s}_m(D_n)\right)\right\}} \xrightarrow[n\to\infty]{p.s.} 1 \quad .$$

D'autre part, dans un cadre non-asymptotique, on dit qu'une procédure de choix d'estimateur $\widehat{m}(\cdot)$ satisfait une *inégalité-oracle* avec constante C_n (et terme de reste R_n) lorsque

$$\ell\left(s^\star, \widehat{s}_{\widehat{m}(D_n)}(D_n)\right) \le C_n \inf_{m\in\mathcal{M}_n}\left\{\ell\left(s^\star, \widehat{s}_m(D_n)\right)\right\} + R_n \tag{4}$$

est vérifié en espérance, ou avec grande probabilité (c'est-à-dire, une probabilité au moins aussi grande que $1 - C'/n^2$, pour une constante $C' > 0$). Remarquons que si (4) est vérifiée pour n assez grand sur un événement de probabilité $1 - C'/n^2$, et si C_n tend vers 1 quand n tend vers l'infini, alors $\widehat{m}$ est asymptotiquement optimale.

L'oracle est parfois défini comme $\arg\min_{m\in\mathcal{M}_n} \{\mathbb{E}[\ell(s^\star, \widehat{s}_m(D_n))]\}$, conduisant à une forme affaiblie de l'inégalité-oracle (4) :

$$\mathbb{E}\left[\ell\left(s^\star, \widehat{s}_{\widehat{m}(D_n)}(D_n)\right)\right] \leq C_n \inf_{m\in\mathcal{M}_n} \{\mathbb{E}[\ell(s^\star, \widehat{s}_m(D_n))]\} + R_n \ . \qquad (5)$$

En effet, si (4) a lieu en espérance, alors (5) a lieu car

$$\mathbb{E}\left[\inf_{m\in\mathcal{M}_n} \{\ell(s^\star, \widehat{s}_m(D_n))\}\right] \leq \inf_{m\in\mathcal{M}_n} \{\mathbb{E}[\ell(s^\star, \widehat{s}_m(D_n))]\} \ .$$

Notons enfin que les procédures de sélection d'estimateurs pour l'estimation sont souvent utilisées pour construire des *estimateurs adaptatifs* (par exemple au sens du minimax) en choisissant une famille d'estimateurs adéquate [7].

3.2. **Sélection d'estimateurs pour l'identification.** On parle d'objectif d'*identification* lorsque l'on souhaite identifier la «meilleure» règle d'apprentissage (c'est-à-dire, celle qui minimise le risque) avec une probabilité maximale. C'est un objectif plus ambitieux que l'objectif d'estimation, dans la mesure où sélectionner l'oracle avec grande probabilité implique une inégalité-oracle (5) avec la même probabilité.

En retour, parler d'identification nécessite de faire une hypothèse supplémentaire, qui est que cette «meilleure règle» soit *identifiable*, c'est-à-dire bien définie de manière unique.

Considérons le cas de la sélection de modèles pour préciser les choses (voir par exemple [24] pour une description plus formelle du cas général).

L'objectif est alors de sélectionner le «vrai modèle» S_{m_0}, défini comme le plus petit modèle (au sens de l'inclusion) parmi $(S_m)_{m\in\mathcal{M}_n}$ qui contient $s^\star$. En particulier, il faut supposer que $s^\star \in \bigcup_{m\in\mathcal{M}_n} S_m$ (pour n assez grand), et que m_0 est effectivement identifiable (c'est-à-dire qu'il existe un m_0 tel que tout modèle S_m contenant $s^\star$ contient nécessairement S_{m_0}).

On mesure alors la qualité d'une procédure de choix de modèles pour l'identification par la probabilité de sélectionner m_0 (que l'on souhaite maximiser). L'équivalent de l'optimalité asymptotique est appelée *consistance* (souvent appelée «consistance en modèle» pour ne pas la confondre avec la consistance au sens de la perte), définie par

$$\mathbb{P}(\widehat{m}(D_n) = m_0) \xrightarrow[n\to\infty]{} 1 \ .$$

Comme nous l'avons remarqué, lorsqu'il existe un unique «vrai modèle» m_0, une procédure de choix de modèles consistante (en modèle) est nécessairement asymptotiquement optimale. On peut naturellement se demander s'il est possible de construire une procédure de choix de modèles atteignant simultanément l'objectif d'identification (lorsque $s^\star \in \bigcup_{m\in\mathcal{M}_n} S_m$) et l'objectif d'estimation (lorsque $s^\star \notin \bigcup_{m\in\mathcal{M}_n} S_m$).

Yang [23] a prouvé que c'est impossible en régression en général, en montrant qu'aucune procédure de choix de modèles ne peut être simultanément

adaptative au sens du minimax (une propriété qui découle de l'optimalité asymptotique lorsque la famille de modèles est bien choisie) et consistante en modèle.

En revanche, en ajoutant des hypothèses sur le problème considéré, il est parfois possible de posséder les deux propriétés simultanément (voir [20] et l'introduction de [23]).

3.3. **Décomposition biais-variance du risque : choix de modèles général.** À partir de cette section et jusqu'à la fin de ce premier cours, nous nous focaliserons sur le problème de la sélection de modèles, étant entendu que les principales idées développées ici pourraient être étendues au problème plus général de la sélection d'estimateurs. Nous reparlerons du problème général au cours des chapitres suivants.

Soit $(S_m)_{m\in\mathcal{M}_n}$ une famille de modèles, i.e., pour tout $m \in \mathcal{M}_n$, $S_m \subset \mathbb{S}$, et notons $\widehat{s}_m$ l'estimateur par minimum de contraste associé (γ étant un contraste fixé).

On peut décomposer l'excès de risque $\mathbb{E}[\ell(s^\star, \widehat{s}_m(D_n))]$ en deux termes. D'une part, comme $\widehat{s}_m(D_n) \in S_m$ p.s., l'excès de risque est minoré par l'*erreur d'approximation* (aussi appelée *biais*) définie par

$$\ell(s^\star, S_m) := \inf_{t\in S_m} \{\ell(s^\star, t)\} \quad .$$

Lorsque cet infimum est atteint en au moins un élément de S_m, on note $\ell(s^\star, s_m^\star) = \ell(s^\star, S_m)$. Nous ferons souvent cette hypothèse simplificatrice dans la suite, sachant que l'on peut également définir $s_m^\star$ comme un δ-minimiseur de $\ell(s^\star, t)$ sur S_m avec δ choisi assez petit (e.g., $\delta = n^{-2}$). Le biais quantifie la «distance» entre le modèle S_m et la cible $s^\star$. Il est d'autant plus petit que S_m est grand.

Le second terme, appelée *erreur d'estimation* (ou *variance*), est défini de manière générale par

$$\mathbb{E}[\ell(s^\star, \widehat{s}_m(D_n))] - \ell(s^\star, S_m) = \mathbb{E}[P\gamma(\widehat{s}_m(D_n))] - \inf_{t\in S_m} P\gamma(t) \quad .$$

Lorsque $s_m^\star$ est bien défini, on peut le réécrire

$$\mathbb{E}[P\gamma(\widehat{s}_m(D_n)) - P\gamma(s_m^\star)] \quad .$$

Ce terme quantifie la difficulté d'estimation au sein du modèle S_m. Il est en général d'autant plus grand que le modèle S_m est grand. Dans plusieurs cadres classiques, on peut prouver qu'il est (approximativement) proportionnel au nombre de paramètres libres du modèle S_m.

Au final, on a la décomposition de l'excès de risque

$$\begin{aligned}\mathbb{E}[\ell(s^\star, \widehat{s}_m(D_n))] &= \ell(s^\star, S_m) + \mathbb{E}[P\gamma(\widehat{s}_m(D_n)) - P\gamma(s_m^\star)] \qquad (6)\\ &= \text{Biais} + \text{Variance} \ .\end{aligned}$$

Un choix de modèle optimal en terme d'excès de risque nécessite donc de trouver un compromis entre ces deux termes, i.e., choisir S_m assez grand pour diminuer son biais, mais pas trop pour éviter le sur-apprentissage (lorsque le terme de variance devient prépondérant). On parle souvent de *compromis biais-variance*. Typiquement, pour l'oracle, les termes de biais et de variance sont très proches.

3.4. **Décomposition biais-variance du risque : régression homoscédastique sur un design fixe, moindres carrés.** Dans ce cadre simple, on peut calculer simplement le risque. Rappelons que l'on a $Y = F + \varepsilon$ et que l'on considère l'estimateur $\widehat{F}_m = A_m Y$ où A_m est la matrice de projection orthogonale sur l'espace vectoriel S_m. Alors,

$$\frac{1}{n}\left\|\widehat{F}_m - F\right\|^2 = \frac{1}{n}\left\|(A_m - I)F\right\|^2 + \frac{1}{n}\left\|A_m\varepsilon\right\|^2 \tag{7}$$

$$\begin{aligned}\text{d'où}\quad \mathbb{E}\left[\frac{1}{n}\left\|\widehat{F}_m - F\right\|^2\right] &= \frac{1}{n}\left\|(A_m - I)F\right\|^2 + \frac{\sigma^2\,\mathrm{tr}(A_m^\top A_m)}{n} \\ &= \frac{1}{n}\left\|(A_m - I)F\right\|^2 + \frac{\sigma^2\dim(S_m)}{n} \\ &= \text{Erreur d'approximation} + \text{Erreur d'estimation}\end{aligned}$$

si les ε_i sont indépendants, centrés et de variance σ^2, en utilisant que $A_m^\top A_m = A_m$ car A_m est une matrice de projection orthogonale, et que $\mathrm{tr}(A_m) = \dim(S_m)$. Dans cette preuve, on a utilisé le Lemme suivant.

Lemme 1. *Soient $M \in \mathcal{M}_n(R)$ et $\varepsilon_1, \ldots, \varepsilon_n$ des v.a. centrées et de même variance σ^2. Alors,*

$$\mathbb{E}\left[\,\langle \varepsilon,\, M\varepsilon\rangle\,\right] = \sigma^2\,\mathrm{tr}(M)\ .$$

Démonstration.

$$\begin{aligned}\mathbb{E}\left[\,\langle \varepsilon,\, M\varepsilon\rangle\,\right] &= \mathbb{E}\left[\sum_{i=1}^n \varepsilon_i (M\varepsilon)_i\right] = \mathbb{E}\left[\sum_{i=1}^n \varepsilon_i \sum_{j=1}^n M_{i,j}\varepsilon_j\right] \\ &= \sum_{i=1}^n \sum_{j=1}^n \left(\,M_{i,j}\mathbb{E}\left[\,\varepsilon_i\varepsilon_j\,\right]\right) = \sum_{i=1}^n \left(\,M_{i,i}\mathbb{E}\left[\,\varepsilon_i^2\,\right]\right) = \sigma^2\,\mathrm{tr}(M)\ .\end{aligned}$$

□

3.5. **Principe d'une sélection à l'aide des données.** Pour choisir $\widehat{m}(D_n) \in \mathcal{M}_n$, on utilise la stratégie générale suivante :

$$\widehat{m}(D_n) \in \arg\min_{m\in\mathcal{M}_n}\left\{\,\mathrm{crit}(m; D_n)\,\right\}\ ,$$

où $\mathrm{crit}(m; D_n)$ est un critère pouvant dépendre des données, idealement egal au risque $P\gamma(\widehat{s}_m(D_n))$ (à constante additive près).

3.5.1. *Principe de l'estimation sans biais du risque.* On prend

$$\mathrm{crit}(m; D_n) \approx \mathbb{E}\left[\,P\gamma(\widehat{s}_m(D_n))\,\right]$$

(à constante additive près). Cela fonctionne s'il n'y a pas trop de modèles (ou d'estimateurs), modulo une inégalité de concentration de $P\gamma(\widehat{s}_m(D_n))$ autour de son espérance (uniforme en $m \in \mathcal{M}$). Si $\mathcal{M}_n$ est trop riche, on ne peut pas obtenir cette concentration uniforme avec une simple borne d'union, si bien qu'il peut être nécessaire de faire differemment (i.e., utiliser la structure de la famille d'estimateurs, ou bien regrouper astucieusement les m par groupes).

3.5.2. *Pénalisation.* Le risque empirique est optimiste comme estimation du risque (exemple d'une famille de modèles emboîtés, sur-apprentissage).

Pour estimer sans biais le risque, il faut donc lui ajouter une quantité appelée pénalité. Une procédure de choix de modèle par pénalisation est une procédure de la forme

$$\widehat{m}(D_n) \in \arg\min_{m\in\mathcal{M}_n} \{ P_n\gamma(\widehat{s}_m(D_n)) + \mathrm{pen}(m; D_n) \} \quad . \tag{8}$$

Le choix idéal (à constante additive près) pour la pénalité est appelé *pénalité idéale*, définie par

$$\mathrm{pen}_{\mathrm{id}}(m; D_n) := (P - P_n)\gamma(\widehat{s}_m(D_n)) \quad .$$

En appliquant le principe d'estimation sans biais du risque, on aboutit à la pénalité déterministe idéale

$$\mathbb{E}[\mathrm{pen}_{\mathrm{id}}(m; D_n)] = \mathbb{E}[(P - P_n)\gamma(\widehat{s}_m(D_n))] \quad .$$

3.6. **Pénalité idéale : régression homoscédastique sur un design fixe, moindres carrés.** Reprenons l'exemple de la régression homoscédastique sur un design fixe, avec la perte quadratique. Le risque empirique vaut

$$\frac{1}{n}\left\|\widehat{F}_m - Y\right\|^2 = \frac{1}{n}\left\|\widehat{F}_m - F\right\|^2 - \frac{2}{n}\langle (A_m - I_n)F, \varepsilon\rangle - \frac{2}{n}\langle A_m\varepsilon, \varepsilon\rangle + \frac{1}{n}\|\varepsilon\|^2 \quad ,$$

d'où $\mathbb{E}\left[\frac{1}{n}\left\|\widehat{F}_m - Y\right\|^2\right] = \mathbb{E}\left[\frac{1}{n}\left\|\widehat{F}_m - F\right\|^2\right] - \frac{2\sigma^2\,\mathrm{tr}(A_m)}{n} + \sigma^2$

si les ε_i sont indépendants, centrés et de variance σ^2.

On en déduit l'expression suivante pour la pénalité idéale et son espérance (en lui ajoutant $n^{-1}\|\varepsilon\|^2$ qui ne dépend pas de m) :

$$\begin{aligned} \mathrm{pen}_{\mathrm{id}}(m) &= \frac{1}{n}\left\|\widehat{F}_m - F\right\|^2 - \frac{1}{n}\left\|\widehat{F}_m - Y\right\|^2 + \frac{1}{n}\|\varepsilon\|^2 \\ &= \frac{2}{n}\langle A_m\varepsilon, \varepsilon\rangle + \frac{2}{n}\langle (A_m - I_n)F, \varepsilon\rangle \qquad (9) \\ \mathbb{E}[\mathrm{pen}_{\mathrm{id}}(m)] &= \frac{2\sigma^2\,\mathrm{tr}(A_m)}{n} = \frac{2\sigma^2\dim(S_m)}{n} \end{aligned}$$

La pénalité $\frac{2\sigma^2\dim(S_m)}{n}$ est aussi appelée C_p de Mallows. Si l'on n'avait pas utilisé les propriétés de A_m (matrice de projection orthogonale), on aurait aboutit à la pénalité plus générale $\frac{2\sigma^2\,\mathrm{tr}(A_m)}{n}$, aussi appelée C_L de Mallows [15]. Pour les estimateurs linéaires, $\mathrm{tr}(A_m)$ est appelé nombre de degrés de

liberté généralisés associé à A_m, par analogie avec le cas des estimateurs par projection orthogonale.

3.7. **Autres procédures fondées sur l'estimation sans biais du risque.** Outre les exemples que nous verrons aux cours suivants, nous renvoyons à la Section 3 de [3] pour des références concernant d'autres procédures de choix de modèles (ou d'estimateurs) fondées sur l'estimation sans biais du risque.

4. Une inégalité-oracle pour la sélection de modèles

Le but de cette section est de prouver une inégalité-oracle à l'aide d'un schéma de preuve très général, qui sera utilisé dans les cours suivants, et qui est notamment pour l'essentiel celui utilisé par [7, 16].

4.1. **Début de preuve d'inégalité-oracle.**

Lemme 2. *Soit* pen : $\mathcal{M}_n \mapsto \mathbb{R}$ *une pénalité (pouvant dépendre des données). Soit* $(A(m))_{m\in\mathcal{M}_n}$, $(B(m))_{m\in\mathcal{M}_n} \in \mathbb{R}^{\mathcal{M}_n}$. *Sur l'événement* Ω *où pour tout* $m, m' \in \mathcal{M}_n$,

$$\big(\mathrm{pen}(m) - \mathrm{pen}_{\mathrm{id}}(m, D_n)\big) - \big(\mathrm{pen}(m') - \mathrm{pen}_{\mathrm{id}}(m')\big) \leq A(m) + B(m') \ , \tag{10}$$

$$\textit{on a} \quad \forall \widehat{m} \in \arg\min_{m\in\mathcal{M}_n} \big\{ P_n\gamma\big(\widehat{s}_m(D_n)\big) + \mathrm{pen}(m) \big\} \ , \tag{11}$$

$$\ell\big(s^\star, \widehat{s}_{\widehat{m}}(D_n)\big) - B(\widehat{m}) \leq \inf_{m\in\mathcal{M}_n} \big\{ \ell\big(s^\star, \widehat{s}_m(D_n)\big) + A(m) \big\} \ . \tag{12}$$

Preuve du Lemme 2. D'après la définition (11) de $\widehat{m}$, pour tout $m \in \mathcal{M}_n$,

$$P_n\gamma\big(\widehat{s}_{\widehat{m}}(D_n)\big) + \mathrm{pen}(\widehat{m}, D_n) \leq P_n\gamma\big(\widehat{s}_m(D_n)\big) + \mathrm{pen}(m)$$

ce qui peut se réécrire : $\forall m \in \mathcal{M}_n$,

$$\ell\big(s^\star, \widehat{s}_{\widehat{m}}(D_n)\big) + \mathrm{pen}(\widehat{m}) - \mathrm{pen}_{\mathrm{id}}(\widehat{m}, D_n) \leq \ell\big(s^\star, \widehat{s}_m(D_n)\big) + \mathrm{pen}(m) - \mathrm{pen}_{\mathrm{id}}(m, D_n) \ . \tag{13}$$

En utilisant l'équation (10) avec $m' = \widehat{m}$, (13) entraîne

$$\forall m \in \mathcal{M}_n, \quad \ell\big(s^\star, \widehat{s}_{\widehat{m}}(D_n)\big) \leq \ell\big(s^\star, \widehat{s}_m(D_n)\big) + A(m) + B(\widehat{m}) \ ,$$

d'où le résultat. □

4.2. **Une inégalité oracle pour la régression Gaussienne homoscédastique.** En s'appuyant sur les calculs précédents, le Lemme 2 et deux inégalités de concentration, on peut reprouver un résultat dû à Birgé et Massart [10, 9].

Théorème 1. *On se place dans le cadre de la régression à design fixe, avec la perte des moindres carrés. On suppose donnée une famille d'estimateurs des moindres carrés* $(\widehat{F}_m)_{m\in\mathcal{M}_n}$ *associée à une famille de modèles* $(S_m)_{m\in\mathcal{M}_n}$ *qui sont des s.e.v. de dimension finie de* $\mathbb{R}^n$. *On suppose que* $\mathrm{Card}(\mathcal{M}_n)$ *est fini.*

On suppose que les données sont de la forme $Y = F + \varepsilon$ *avec* $F \in \mathbb{R}^n$ *inconnu et* $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ *(cadre Gaussien homoscédastique). On note* A_m

la matrice de projection orthogonale sur S_m, de telle sorte que $\widehat{F}_m = A_m Y$ et $\operatorname{tr}(A_m) = \dim(S_m) = D_m$.

Soit $K > 1$ une constante. Pour tout $m \in \mathcal{M}_n$, on définit

$$\operatorname{pen}(m) = \frac{K\sigma^2 \dim(S_m)}{n} \ .$$

Alors, il existe $C(K) > 0$ telle que pour tout $x \geq 0$, avec probabilité au moins $1 - 4\operatorname{Card}(\mathcal{M}_n)e^{-x}$, on a pour tout

$$\widehat{m} \in \arg\min_{m \in \mathcal{M}_n} \left\{ \frac{1}{n} \left\| Y - \widehat{F}_m \right\|^2 + \operatorname{pen}(m) \right\}$$

l'inégalité-oracle : pour tout $\delta > 0$,

$$\frac{1}{n} \left\| \widehat{F}_{\widehat{m}} - F \right\|^2 \leq \left(\frac{1 + (K-2)_+}{1 - (2-K)_+} + \delta \right) \inf_{m \in \mathcal{M}_n} \left\{ \frac{1}{n} \left\| \widehat{F}_m - F \right\|^2 \right\} + \frac{C(K) x \sigma^2}{\delta n} \tag{14}$$

Notons que l'on a d'après la preuve du Théorème 1 la majoration

$$C(K) \leq L \max\left\{ (K-1)^{-3}, (K-1)^2 \right\} \ ,$$

avec $L > 0$ une constante numérique.

En particulier, en prenant $K = 2$ (pénalité C_p de Mallows), (14) s'écrit

$$\frac{1}{n} \left\| \widehat{F}_{\widehat{m}} - F \right\|^2 \leq (1 + \delta) \inf_{m \in \mathcal{M}_n} \left\{ \frac{1}{n} \left\| \widehat{F}_m - F \right\|^2 \right\} + \frac{L x \sigma^2}{\delta n} \ , \tag{15}$$

qui implique l'optimalité asymptotique de $\widehat{m}$ (en intégrant par rapport à x ou en prenant $x = 2\ln(n) + \ln(\operatorname{Card}(\mathcal{M}_n))$, et en faisant tendre δ vers 0 à une vitesse adéquate et en supposant que la vitesse d'apprentissage de l'oracle est non-paramétrique :

$$\inf_{m \in \mathcal{M}_n} \left\{ \frac{1}{n} \left\| \widehat{F}_m - F \right\|^2 \right\} \gg \frac{1}{n} \ .$$

Pour prouver le Théorème 1, on a besoin des deux inégalités de concentration suivantes [16].

Proposition 3. *Soit ξ un vecteur Gaussien standard de $\mathbb{R}^n$, $\alpha \in \mathbb{R}^n$, $M \in \mathcal{M}_n(\mathbb{R})$. On note $\|\alpha\|$ la norme euclidienne de α, $|\!|\!| M |\!|\!|$ la norme d'opérateur de M (sup des modules des valeurs propres). Alors, pour tout $x \geq 0$,*

$$\mathbb{P}\left(|\langle \xi, \alpha \rangle| \leq \sqrt{2x}\, \|\alpha\|_2 \right) \geq 1 - 2e^{-x}$$

$$\mathbb{P}\left(|\langle \xi, M\xi \rangle - \operatorname{tr}(M)| \leq 2\sqrt{x \operatorname{tr}(M^\top M)} + 2\, |\!|\!| M |\!|\!|\, x \right) \geq 1 - 2e^{-x} \ .$$

Preuve du Théorème 1. On découpe cette preuve en 5 étapes principales :

(1) Pour chaque $x \geq 0$ et $m \in \mathcal{M}_n$, on concentre $\langle A_m \varepsilon, \varepsilon \rangle$ autour de $\sigma^2 \operatorname{tr}(A_m)$ et $\langle (A_m - I_n)F, \varepsilon \rangle$ autour de zéro. Ceci définit un événement $\Omega_{m,x}$ de probabilité $1 - 4e^{-x}$, sur lequel pour tout $\theta > 0$

$$\left| \langle A_m \varepsilon, \varepsilon \rangle - \sigma^2 D_m \right| \leq \theta D_m \sigma^2 + \left(\theta^{-1} + 2 \right) x \sigma^2 \tag{16}$$

$$|\langle (A_m - I_n)F,\, \varepsilon\rangle| \leq \theta \left\|(A_m - I_n)F\right\|^2 + \frac{1}{2\theta} x\sigma^2 \tag{17}$$

en utilisant le fait que pour tout $a, b > 0$, $2\sqrt{ab} \leq \theta a + \theta^{-1} b$.

(2) Soit $\Omega_x = \bigcap_{m\in\mathcal{M}} \Omega_{m,x}$, pour lequel une borne d'union donne $\mathbb{P}(\Omega) \geq 1 - 4\,\mathrm{Card}(\mathcal{M}_n)e^{-x}$. Sur Ω_x, en utilisant (16), (17) et (9), pour tout $m \in \mathcal{M}_n$ et $\theta > 0$,

$$\begin{aligned}\left|\mathrm{pen}_{\mathrm{id}}(m) - \frac{2\sigma^2 D_m}{n}\right| &\leq \frac{2\theta\sigma^2 D_m}{n} + \frac{2\theta}{n}\left\|(A_m - I_n)F\right\|^2 + \left(4 + 3\theta^{-1}\right)\frac{x\sigma^2}{n} \\ &= 2\theta\mathbb{E}\left[\frac{\left\|F - \widehat{F}_m\right\|^2}{n}\right] + \left(4 + 3\theta^{-1}\right)\frac{x\sigma^2}{n}\ . \end{aligned} \tag{18}$$

(3) On majore ce terme de reste à l'aide de la perte relative $\left\|F - \widehat{F}_m\right\|^2$ directement (et non son espérance). En effet, en utilisant (7) et le fait que A_m est une matrice de projection orthogonale, on a

$$\begin{aligned}\left\|F - \widehat{F}_m\right\|^2 &= \left\|(A_m - I)F\right\|^2 + \left\|A_m\varepsilon\right\|^2 = \left\|(A_m - I)F\right\|^2 + \langle A_m\varepsilon,\, \varepsilon\rangle \\ &\geq \left\|(A_m - I)F\right\|^2 + \sigma^2 D_m - \theta\sigma^2 D_m - \left(\theta^{-1} + 2\right)x\sigma^2 \\ &\geq (1-\theta)\mathbb{E}\left[\left\|F - \widehat{F}_m\right\|^2\right] - \left(\theta^{-1} + 2\right)x\sigma^2\end{aligned}$$

sur Ω_x d'après (16). Donc, pour tout $\theta \in\,]0,1[$,

$$\mathbb{E}\left[\left\|F - \widehat{F}_m\right\|^2\right] \leq \frac{1}{1-\theta}\left\|F - \widehat{F}_m\right\|^2 + \frac{1}{1-\theta}\left(\theta^{-1} + 2\right)x\sigma^2\ . \tag{19}$$

En injectant (19) dans (18), on obtient ainsi pour tout $m \in \mathcal{M}_n$ et $\theta \in\,]0,1[$,

$$\left|\mathrm{pen}_{\mathrm{id}}(m) - \frac{2\sigma^2 D_m}{n}\right| \leq \frac{2\theta}{1-\theta}\left\|F - \widehat{F}_m\right\|^2 + \left(\frac{2\theta}{1-\theta}\left(\theta^{-1} + 2\right) + 4 + 3\theta^{-1}\right)x\sigma^2\ .$$

(4) D'après (5) et (19), l'hypothèse du Lemme 2 est donc vérifiée sur Ω_x avec

$$\begin{aligned} A(m) &= \frac{(K-2)_+ + 2\theta}{1-\theta}\left\|F - \widehat{F}_m\right\|^2 + C_1(\theta, K)\frac{x\sigma^2}{n} \\ B(m) &= \frac{(2-K)_+ + 2\theta}{1-\theta}\left\|F - \widehat{F}_m\right\|^2 + C_2(\theta, K)\frac{x\sigma^2}{n} \\ \text{avec}\quad C_1(\theta, K) &= 4 + 3\theta^{-1} + \frac{2\theta + (K-2)_+}{1-\theta}\left(\theta^{-1} + 2\right) \\ \text{et}\quad C_2(\theta, K) &= 4 + 3\theta^{-1} + \frac{2\theta + (2-K)_+}{1-\theta}\left(\theta^{-1} + 2\right)\ . \end{aligned}$$

Par le Lemme 2, on obtient donc que sur Ω_x, pour tout $m \in \mathcal{M}_n$ et $\theta \in\,]0,1[$,

$$\left(1 - \frac{(2-K)_+ + 2\theta}{1-\theta}\right)\frac{1}{n}\left\|F - \widehat{F}_{\widehat{m}}\right\|^2$$

$$\leq \left(1+\frac{(2-K)_+ + 2\theta}{1-\theta}\right) \inf_{m\in\mathcal{M}_n} \left\{ \frac{1}{n} \left\| F - \widehat{F}_m \right\|^2 \right\} + \left(C_1(\theta,K) + C_2(\theta,K) \right) \frac{x\sigma^2}{n}$$

(5) Pour en déduire le résultat annoncé, on suppose tout d'abord que $\theta \in \,]0,1[$ vérifie $\theta < (K-1)/3$, si bien que $\frac{(2-K)_+ + 2\theta}{1-\theta} < 1$. On obtient alors que

$$\frac{1}{n} \left\| F - \widehat{F}_{\widehat{m}} \right\|^2 \leq C_3(K,\theta) \inf_{m\in\mathcal{M}_n} \left\{ \frac{1}{n} \left\| F - \widehat{F}_m \right\|^2 \right\} + C_4(K,\theta) \frac{x\sigma^2}{n}$$

$$\text{avec} \quad C_3(K,\theta) = \left(1 - \frac{(2-K)_+ + 2\theta}{1-\theta}\right)^{-1} \left(1+\frac{(2-K)_+ + 2\theta}{1-\theta}\right)$$

$$\text{et} \quad C_4(K,\theta) = \left(1 - \frac{(2-K)_+ + 2\theta}{1-\theta}\right)^{-1} \left(C_1(\theta,K) + C_2(\theta,K) \right) \ .$$

Le résultat s'en déduit en choisissant un θ adéquat.

Plus précisément, en supposant que $0 < \theta < \min\{ L_1 \, , \, L_2(K-1) \}$ pour des constantes numériques $L_1, L_2 > 0$, des constantes numériques $L_3, L_4, L_5 > 0$ existent telles que si $K \in \,]1,2]$, alors

$$C_3(K,\theta) \leq \frac{1}{K-1} \left(1 + \frac{L_3\theta}{K-1} \right)$$

$$C_4(K,\theta) \leq \frac{L_4}{(K-1)\theta}$$

et si $K \geq 2$,

$$C_3(K,\theta) \leq (K-1)\left(1+L_5\theta\right)$$

$$C_4(K,\theta) \leq \frac{L_4(K-1)}{\theta} \ .$$

Il suffit donc de prendre

$$\theta = L_6\delta \min\left\{ (K-1)^2 \, , \, \frac{1}{K-1} \right\}$$

pour obtenir le résultat annoncé ($L_6 > 0$ étant une constante numérique), si bien que l'on a

$$C'(K,\delta) \leq C_4\left(K,\theta\right) \leq \frac{L_7}{\delta} \max\left\{ (K-1)^{-3} \, , \, (K-1)^2 \right\} \ ,$$

$L_7 > 0$ étant une constante numérique.

□

5. Références mentionnées à la fin de l'exposé

– à propos de programmation dynamique pour la détection de ruptures : [8, 17]
– à propos de la pénalité idéale pour le Lasso : [14, 26]

Références

[1] Sylvain Arlot. Classification supervisée : des algorithmes et leur calibration automatique, 2009. Notes d'un cours de troisième année à l'École Centrale Paris. `http://www.di.ens.fr/~arlot/enseign/2009Centrale/cours-classif.pdf.gz`.

[2] Sylvain Arlot and Francis Bach. Data-driven calibration of linear estimators with minimal penalties. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 46–54, 2009.

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

[4] Francis Bach. Tutorial on sparse methods for machine learning (theory and algorithms), 2010. Transparents d'un cours donné à ECML. `http://www.di.ens.fr/~fbach/`.

[5] Yannick Baraud. Model selection for regression on a fixed design. *Probab. Theory Related Fields*, 117(4) :467–493, 2000.

[6] Yannick Baraud. Model selection for regression on a random design. *ESAIM Probab. Statist.*, 6 :127–146 (electronic), 2002.

[7] Andrew Barron, Lucien Birgé, and Pascal Massart. Risk bounds for model selection via penalization. *Probab. Theory Related Fields*, 113(3) :301–413, 1999.

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

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

[10] Lucien Birgé and Pascal Massart. A generalized cp criterion for gaussian model selection. Technical report, Universités de Paris 6 et Paris 7, 2010. Prépublication 647, 39 pages.

[11] Gilles Blanchard, Olivier Bousquet, and Pascal Massart. Statistical performance of support vector machines. *Ann. Statist.*, 36(2) :489–531, 2008.

[12] Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. Theory of classification : a survey of some recent advances. *ESAIM Probab. Stat.*, 9 :323–375 (electronic), 2005.

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

[14] Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *Ann. Statist.*, 32(2) :407–499, 2004. With discussion, and a rejoinder by the authors.

[15] Colin L. Mallows. Some comments on C_p. *Technometrics*, 15 :661–675, 1973.

[16] Pascal Massart. *Concentration Inequalities and Model Selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.

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

[18] Bernhard Scholkopf and Alexander J. Smola. *Learning with Kernels : Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, USA, 2001.

[19] Robert Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1) :267–288, 1996.

[20] T. van Erven, P. D. Grünwald, and S. de Rooij. Catching Up Faster by Switching Sooner : A Prequential Solution to the AIC-BIC Dilemma, July 2008. arXiv :0807.1005.

[21] Vladimir Vapnik. *Estimation of Dependences Based on Empirical Data*. Springer Series in Statistics. Springer-Verlag, New York, 1982. Translated from the Russian by Samuel Kotz.

[22] Grace Wahba. *Spline Models for Observational Data*, volume 59 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1990.

[23] Yuhong Yang. Can the strengths of AIC and BIC be shared ? A conflict between model indentification and regression estimation. *Biometrika*, 92(4) :937–950, 2005.

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

[25] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of The Royal Statistical Society Series B*, 68(1) :49–67, 2006.

[26] Hui Zou, Trevor Hastie, and Robert Tibshirani. On the "degrees of freedom" of the lasso. *Ann. Statist.*, 35(5) :2173–2192, 2007.

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

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

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