Sélection d'estimateurs ridge en régression Gaussienne

Carole Binard [1]

[1] *Laboratoire de Mathématiques J.A. Dieudonné, Université de Nice Sophia Antipolis, 06108 Nice Cedex 02, France – binard@unice.fr*

Résumé. Dans le cadre de la régression Gaussienne à variance inconnue, Baraud *et al* (2012) ont développé une procédure permettant de sélectionner un estimateur de l'espérance d'un vecteur Gaussien Y, sélection opérée au sein d'une collection arbitraire d'estimateurs. Dans un premier temps, nous comparons les performances de cette procédure appliquée aux estimateurs Ridge à celles de la validation croisée. Dans un second temps, nous considérons des estimateurs Ridge à noyaux et comparons cette procédure à la validation croisée. Puis nous regardons, d'un point de vue théorique, la sélection d'estimateurs Ridge par morceaux qui consiste à sélectionner un "meilleur" paramètre de lissage sur chacun des morceaux d'une partition fixée de $[0, 1]$.

Mots-clés. régression Ridge, collection d'estimateurs, méthode de sélection, validation croisée, méthodes à noyaux, estimateurs par morceaux.

Abstract. In the Gaussian regression framework with unknown variance, Baraud *et al* (2012) developed a procedure that can select an estimator of the mean of a Gaussian vector Y. That procedure is based on an estimator selection within an arbitrary collection of estimators. In a first part, we consider a family of Ridge estimators and we compare the estimation performances of that procedure with the performances of the cross-validation. In a second part, we compare these two procedures by considering the selection of kernel Ridge estimators. Then, from a theoretical point of view, we study the selection of piecewise Ridge estimators to select a "best" regularization parameter on each piece of a fixed partition of $[0, 1]$.

Keywords. Ridge regression, collection of estimators, selection method, cross validation, kernel methods, piecewise estimators.

1 Méthode de sélection d'estimateurs développée par Baraud *et al* (2012) *versus* la validation croisée dans le cadre des estimateurs Ridge

Nous considérons le modèle de régression suivant

$$Y_i = f_i + \epsilon_i,\ i = 1, ..., n \tag{1}$$

avec $f = (f_1, \ldots, f_n)^T$ un vecteur de $\mathbb{R}^n$ à estimer et $\epsilon \sim \mathcal{N}(0, \sigma^2 I_n)$.
Nous supposons que $f = X\beta$ avec X une matrice réelle de taille $n \times p$ et β un vecteur inconnu de $\mathbb{R}^p$ et qu'un problème mal conditionné a été détecté. Afin de garantir un modèle de régression conservant toutes les variables, nous considérons un estimateur légèrement biaisé de la forme

$$\hat{f}_\lambda = \mathcal{K}(\mathcal{K} + \lambda I_n)^{-1} Y,\ \lambda > 0 \tag{2}$$

avec I_n la matrice identité de $\mathbb{R}^n$ et $\mathcal{K} = XX^T$ une matrice $n \times n$ semie-définie positive.

Nous montrons que le risque quadratique de l'estimateur Ridge $\hat{f}_\lambda$ s'écrit sous la forme

$$\mathbb{E}\left[\|f - \hat{f}_\lambda\|^2\right] = \sum_{i=1}^{n}\left[\left(\frac{\lambda}{s_i + \lambda}\right)^2 \langle f, v_i\rangle^2\right] + \sum_{i=1}^{n}\left[\left(\frac{s_i}{s_i + \lambda}\right)^2 \sigma^2\right] \tag{3}$$

avec, pour $i = 1, \ldots, n$, v_i le $i^{\text{ème}}$ vecteur propre de $\mathcal{K}$ associé à la valeur propre s_i telle que $s_1 \geq s_2 \geq \ldots \geq s_n \geq 0$.

L'objectif est alors d'identifier l'estimateur ou, de façon équivalente, la valeur du paramètre λ minimisant la quantité (3). Pour cela, différentes méthodes statistiques existent. Nous nous sommes concentrés sur deux techniques que sont la procédure développée par Baraud *et al* (2012) et la validation croisée, dont nous avons comparé les performances.

1.1 Procédure de sélection développée par Baraud *et al* (2012)

Développée essentiellement pour estimer l'espérance f d'un vecteur Y de n variables Gaussiennes indépendantes dont la variance est inconnue, la méthode proposée par Baraud *et al.* (2012) s'adapte à la sélection d'un paramètre de régularisation de méthodes de pénalisation faisant intervenir des estimateurs linéaires.
Dans ce cadre d'étude, la méthode procède à une sélection parmi une famille d'estimateurs de la forme $\mathbb{F} = \left\{\hat{f}_\lambda = A_\lambda Y, \lambda \in \Lambda\right\}$, où $\Lambda \subset \mathbb{R}_+$. Soient $\mathbb{S}$ une famille de sous-espaces de $\mathbb{R}^n$ satisfaisant, pour tout $S \in \mathbb{S}$, $\dim(S) \leq n - 2$ et, pour chaque $\hat{f}_\lambda \in \mathbb{F}$, un sous-espace

de $\mathbb{S}$ noté $\mathbb{S}_\lambda$ approximant $\hat{f}_\lambda$ tel que, dans le cas particulier où A_λ est symétrique, la collection $\mathbb{S}_\lambda$ (pour un $\lambda \in \Lambda$ donné) est engendrée par les vecteurs propres de A_λ

$$\mathbb{S}_\lambda = \left\{S_\lambda^1, ..., S_\lambda^{n/2}\right\} \tag{4}$$

où S_λ^k est l'espace engendré par les k premiers vecteurs propres de A_λ associés aux k plus grandes valeurs propres.
Soit un critère $crit_\alpha$ faisant intervenir l'erreur de modélisation de Y par $\Pi_S \hat{f}_\lambda$ (avec Π_S la projection orthogonale sur $S \in \mathbb{S}$), la qualité d'approximation de $\hat{f}_\lambda$ par $\Pi_S \hat{f}_\lambda$ et une pénalisation de S *via* une mesure de complexité $\Delta : \mathbb{S} \to \mathbb{R}^+$. Ce critère vise à estimer le risque (3) et peut se définir, pour tout $\lambda \in \Lambda = \left(s_{n/2}, +\infty\right)$, par

$$crit_\alpha\left(\hat{f}_\lambda\right) = \inf_{1 \leq d \leq \frac{n}{2}} \sum_{j \leq d} \left\langle Y, v_j \right\rangle^2 \left(\frac{\lambda}{\lambda + s_j}\right)^2 + \sum_{j > d} \left\langle Y, v_j \right\rangle^2 \left[1 + \frac{pen\left(S_\lambda^d\right)}{n-d} + \frac{1}{2}\left(\frac{s_j}{\lambda + s_j}\right)^2\right] \tag{5}$$

où v_j est le $j^{\text{ème}}$ vecteur propre de A_λ associé à la valeur propre $\frac{s_j}{s_j + \lambda}$ et, pour tout $S \in \mathbb{S}$, $pen\,(S)$ est une pénalisation de S liée à la mesure de complexité Δ.

Le paramètre de régularisation optimal $\hat{\lambda}$ est alors sélectionné en minimisant (5) et l'estimateur Ridge associé $\hat{f}_{\hat{\lambda}}$ est contrôlé par une borne non-asymptotique qui ressemble à une borne oracle.

1.2 Comparaison à la validation croisée

Nous considérons le cadre d'étude où $f = X\beta$ avec X une matrice réelle de taille $n \times p$ et β un vecteur de $\mathbb{R}^p$. Afin de comparer la procédure de Baraud *et al* (2012) avec la validation croisée, nous regardons le *ratio* $\frac{\mathbb{E}\left[\|f - \hat{f}_{\hat{\lambda}}^{pen_\Delta}\|^2\right]}{\mathbb{E}\left[\|f - \hat{f}_{\hat{\lambda}}^{cv}\|^2\right]}$ où $\mathbb{E}\left[\|f - \hat{f}_{\hat{\lambda}}^{pen_\Delta}\|^2\right]$ représente le risque de l'estimateur $\hat{f}_{\hat{\lambda}}^{pen_\Delta}$, obtenu par la procédure de Baraud *et al* (2012) et $\mathbb{E}\left[\|f - \hat{f}_{\hat{\lambda}}^{cv}\|^2\right]$ le risque de l'estimateur sélectionné par validation croisée à 10 parties (ou 10-fold).
Ce *ratio* a été évalué sur des exemples simulés extraits de l'article de Tibshirani (1996). Pour chaque exemple considéré, nous avons simulé 500 fois le modèle de régression suivant

$$Y = X\beta + \sigma\epsilon$$

où $\epsilon \sim \mathcal{N}(0, I_n)$,
et nous avons obtenu des *ratios* proches de 1.
Par exemple, quand $n = 100$, $p = 8$ et

- $\beta = (3, 1.5, 0, 0, 2, 0, 0, 0)$,

- $\sigma = 3$,
- $corr\left(X^{(i)}, X^{(j)}\right) = 0.5^{|i-j|}$

le *ratio* calculé $\frac{\mathbb{E}\left[\|f-\hat{f}^{pen_\Delta}_{\hat{\lambda}}\|^2\right]}{\mathbb{E}\left[\|f-\hat{f}^{cv}_{\hat{\lambda}}\|^2\right]}$ est égal à 1.080.
Sur cet exemple, ainsi que sur les autres exemples considérés, nous remarquons que les deux procédures semblent avoir des performances comparables. Cependant, si ces deux procédures semblent comparables au niveau de leurs performances, il est à noter que la sélection par validation croisée est plus coûteuse d'un point de vue computationnel et qu'elle souffre d'un manque de garanties théoriques contrairement à la procédure développée par Baraud *et al* (2012).

2 La sélection d'estimateurs Ridge pour des fonctions non linéaires et/ou par morceaux

Dans cette partie, f est un vecteur inconnu de $\mathbb{R}^n$ de la forme $\left(F\left(x_1\right), \ldots, F\left(x_n\right)\right)^T$ avec $F : \mathcal{X} \to \mathbb{R}$ inconnue et $x_1, \ldots, x_n \in \mathcal{X}$. L'objectif est d'estimer F quand, d'une part F est une fonction non linéaire et d'autre part F est une fonction définie par morceaux sur $[0, 1]$.

2.1 Sélection d'estimateurs Ridge à noyaux

Lorsque les fonctions de décision sont non linéaires, les méthodes linéaires d'analyse de données et d'apprentissage sont insuffisantes. Les méthodes à noyaux constituent alors un outil efficace pour, à la fois, tirer profit de la simplicité et des performances d'estimation des techniques linéaires mais aussi traiter de problèmes non linéaires. Il est dans ce cas possible d'utiliser une méthode linéaire pour résoudre un problème non linéaire *via* la transformation de l'espace de représentation des données d'entrées $\mathcal{X}$ en un espace de plus grande dimension dans lequel la méthode linéaire est utilisée.

Dans un Espace de Hilbert à Noyau Reproduisant (EHNR) $\mathcal{H}$ associé au noyau semi-défini positif $\kappa_{\mathcal{H}}$, nous étudions l'estimation de la fonction non linéaire F telle que $f = \left(F\left(x_1\right), \ldots, F\left(x_n\right)\right)^T$ pour $x_1, \ldots, x_n \in \mathcal{X}$. Le problème de minimisation en régression Ridge à noyaux est donné par

$$\hat{F} = \underset{F \in \mathcal{H}}{argmin}\ \mathcal{G}\left(F\right) \tag{6}$$

où $\mathcal{G}\left(F\right) = \sum_{i=1}^{n}\left(Y_i - F\left(x_i\right)\right)^2 + \lambda\|F\|^2_{\mathcal{H}}$ $(\lambda > 0)$, pour $x_1, x_2, \ldots, x_n \in \mathcal{X}$.

Par le théorème de représentation, la solution du problème (6) est de la forme

$$\hat{F}(x) = \sum_{i=1}^{n} a_i \kappa_{\mathcal{H}}(x_i, x)$$

et donc $f = \mathcal{K}a$ avec $\mathcal{K} = \begin{pmatrix} \kappa_{\mathcal{H}}(x_1,x_1) & \kappa_{\mathcal{H}}(x_2,x_1) & \dots & \kappa_{\mathcal{H}}(x_n,x_1) \\ \kappa_{\mathcal{H}}(x_1,x_2) & \kappa_{\mathcal{H}}(x_2,x_2) & \dots & \kappa_{\mathcal{H}}(x_n,x_2) \\ \vdots & \vdots & \ddots & \vdots \\ \kappa_{\mathcal{H}}(x_1,x_n) & \kappa_{\mathcal{H}}(x_2,x_n) & \dots & \kappa_{\mathcal{H}}(x_n,x_n) \end{pmatrix}$ et $a \in \mathbb{R}^n$.
Le problème (6) est alors équivalent à la minimisation en a de la fonction

$$\mathcal{L}(a) = \|Y - \mathcal{K}a\|^2 + \lambda a^T \mathcal{K} a \tag{7}$$

qui a pour solution $a = (\mathcal{K} + \lambda I_n)^{-1} Y$ et l'estimateur de f résultant est linéaire de la forme $\mathcal{K}(\mathcal{K} + \lambda I_n)^{-1} Y$.
Le problème est donc une nouvelle fois de trouver une valeur optimale pour le paramètre λ, optimalité définie par le critère (7).
Tout comme dans le cadre des fonctions F linéaires, nous avons comparé la procédure développée par Baraud *et al* (2012), adaptée à ce contexte, avec la validation croisée (leave-one-out et 10-fold). Nous avons procédé à des simulations en utilisant différents noyaux de l'ouvrage de Schölkopf *et al.* (2004) et différentes fonctions cibles F. Ce travail sur données simulées conduit aux mêmes conclusions que celles formulées précédemment.

2.2 Sélection d'estimateurs Ridge par morceaux

D'un point de vue applicatif, il peut être intéressant de chercher à adapter le paramètre λ à la régularité locale d'une fonction cible. C'est ce qui a conduit à considérer le cadre d'étude théorique suivant.
Soient F une fonction inconnue à valeurs réelles définie sur $[0,1]$ telle que $f = (F(x_1), \dots, F(x_n))^T$ avec f satisfaisant (1) et où $x_1, \dots, x_n$ sont n observations ordonnées de $[0,1]$.
Nous nous donnons une partition m de $[0,1]$ à D éléments. Pour $\lambda = (\lambda_1, \dots, \lambda_D)$, avec pour $i = 1, \dots, D, \lambda_i > 0$, nous définissons l'estimateur $\hat{f}_\lambda$ comme l'estimateur Ridge par morceaux de f qui, sur le $i^{\text{ème}}$ intervalle $\mathcal{I}_i$ de la partition m, coïncide avec l'estimateur Ridge de noyau $\mathcal{K}$ et de paramètre de lissage λ_i. En réécrivant f par $f = \sum_{i=1}^{D} f^{(i)}$ où, pour $i = 1, \dots, D$, $f^{(i)} = \left(F(x_1)\mathbb{1}_{x_1 \in \mathcal{I}_i}, \dots, F(x_n)\mathbb{1}_{x_n \in \mathcal{I}_i}\right)^T$, nous montrons qu'un estimateur Ridge par morceaux peut être sélectionné *via* la procédure de Baraud *et al.* (2012) au moyen de deux stratégies, l'une pas à pas et l'autre globale.

La stratégie pas à pas consiste à travailler sur chaque élément de la partition m séparément. Sur chacun des éléments de la partition un estimateur Ridge est sélectionné au moyen de

la procédure développée par Baraud *et al.* (2012) en utilisant le critère (5). Nous montrons alors que l'estimateur final, noté $\hat{f}_{\hat{\lambda}}$, satisfait une inégalité de type oracle de la forme

$$C\mathbb{E}\left[\|f-\hat{f}_{\hat{\lambda}}\|^2\right] \leq \inf_{\lambda} \mathbb{E}\left[\|f-\hat{f}_{\lambda}\|^2\right] + D\sigma^2 \tag{8}$$

La stratégie dite globale consiste à définir un critère de minimisation sur la partition globale et non plus sur chacun des éléments de m. En prouvant qu'un estimateur de f sur m est linéaire, nous pouvons réécrire le critère de sélection (5) et obtenir un estimateur, noté $\hat{f}_{\hat{\lambda}}$, qui satisfait une inégalité de type oracle de la forme

$$C\mathbb{E}\left[\|f-\hat{f}_{\hat{\lambda}}\|^2\right] \leq a\left(D\right) \inf_{\lambda} \mathbb{E}\left[\|f-\hat{f}_{\lambda}\|^2\right] + \sigma^2 \tag{9}$$

où $a\left(D\right)$ est une constante de l'ordre de $\log\left(D\right)$.

Si la stratégie globale conduit à une amélioration du terme de variance dans l'inégalité de type oracle, le terme de biais est, quant à lui, dégradé. Ceci nous a conduit à regarder les conditions pour lesquelles l'une des stratégies est préférable à l'autre.

Bibliographie

[1] Baraud, Y., Giraud, C. et Huet, S. (2012), Estimator selection in the gaussian setting, *Annales de l'Institut Henri-Poincaré - Probabilités et Statistiques*.
[2] Tibhsirani, R. (1996), Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society. Series B*
[3] Schölkopf,B., Tsuda,K. et Vert, J.-P. (2004), Kernel methods in computational biology, *Bradford Book, the MIT Press*.