Une formule exacte pour la validation croisée dans le cadre de la régression "pool-sample".

Tristan Mary-Huard[1,2], Julien Chiquet[1,3], Alain Célisse[4] & Mathias Fuchs[5]

[1] *AgroParisTech/INRA UMR 518, Paris*
[2] *UMR de Génétique Végétale, INRA/Univ. Paris-Sud/CNRS, 91190 Gif-sur-Yvette*
[3] *LaMME/Statistique et Génome, UMR 8071 CNRS/UEVE/USC INRA, Evry*
[4] Modal *Project-Team, UMR 8524 CNRS-Université Lille 1, F-59 655 Villeneuve d'Ascq Cedex*
[5] *Institut fur Medizinische Informationsverarbeitung Biometrie & Epidemiologie, Ludwig Maximilians Universitat, 81377 Munchen*

Résumé. En régression "pool sample", on dispose d'un échantillon de N individus pour lesquels les variables explicatives sont mesurées, tandis que la variable réponse n'est disponible que pour $n << N$ d'entre eux. Cette disymétrie entre information disponible sur les variables d'une part et la réponse d'autre part amène à modifier la forme des estimateurs classiques (OLS et Ridge) pour exploiter l'intégralité des données. Lorsque plusieurs modèles sont en compétition et doivent être comparés, cette modification doit être intégrée aux procédures de validation croisée. Nous proposons ici une approche fondée sur le rééchantillonnage des seules variables réponses pour la validation croisée. Nous montrons qu'une formule exacte et explicite peut alors être obtenue pour le critère de validation croisée proposé. La sélection de modèles peut être alors réalisée sur la base de ce critère sans en payer le coût algorithmique.

Mots-clés. Sélection de modèles, validation croisée, régression pool-sample

Abstract. In the "pool sample" regression framework, inference is based on N observations for which the explanatory variables are observed, while the response is only available for $n << N$ of them. Consequently, the classical (OLS, Ridge) estimators have to be modified to ensure the integration of all the information available. When model selection is at stake, i.e. several models have to be compared, one also needs to adapt the cross-validation criteria accordingly. We propose a new cross-validation approach where only the responses are resampled. We derive a closed-form expression for this new cross-validation criterion. Model selection can then be performed on the basis of the proposed criterion without suffering the computational burden of the resampling procedure.

Keywords. Model selection, cross-validation, pool-sample regression

1 Sélection de modèle en régression: cadre classique

On s'intéresse à la prédiction d'une variable réponse quantitative Y_0 d'un individu à partir de mesures x_0 réalisées sur cet individu. On suppose ici que x est un vecteur réel de taille d et que la réponse est une fonction linéaire des mesures:

$$Y_0 = x_0\beta + \varepsilon \ ,$$

où β est un vecteur de coefficients inconnus et ε une erreur de mesure supposée gaussienne, centrée et de variance σ^2 inconnue. Afin d'estimer β, on dispose d'un échantillon de n observations $(x_1, Y_1), ..., (x_n, Y_n)$ indépendantes et identiquement distribuées. On notera dans la suite X et Y respectivement la matrice (n, d) et le vecteur de taille n correspondant aux mesures et aux réponses des n observations.

Nous considérons ici les deux estimateurs suivants:

1. l'estimateur par moindres carrés (OLS):

$$\widehat{\beta} = (X^T X)^{-1} X^T Y$$

   qui peut être utilisé dès lors que le nombre de variables d est inférieur au nombre d'observations n (cas "petite dimension"),

2. l'estimateur ridge (Ridge):

$$\widehat{\beta} = (X^T X + \lambda I)^{-1} X^T Y,$$

   où λ est une constante de régularisation strictement positive; cet estimateur peut être en particulier utilisé lorsque que d est supérieur à n (cas "grande dimension").

Une fois les coefficients estimés, il est possible de prédire la variable réponse d'un nouvel individu en utilisant la fonction de prédiction

$$f_{\widehat{\beta}}(x) = x\widehat{\beta} \ .$$

Dans le cas "petite dimension", il est possible de considérer des estimateurs bâtis sur un sous-ensemble des d mesures initiales. Soit $m \in \mathcal{P}(\{1, ..., n\})$ un élément de l'ensemble des parties de $\{1, ..., n\}$. On note X_m la sous-matrice de X composée des colonnes du sous-ensemble m. L'inférence peut alors être réalisée dans chacun des sous-modèles

$$Y = X_m\beta_m + \varepsilon \ ,$$

et la sélection de modèles consiste à sélectionner le meilleur des estimateurs $\widehat{\beta}_m$, $m \in \mathcal{P}(1, ..., n)$ obtenus. Dans le cas grande dimension, la sélection de modèles consiste à choisir la valeur du paramètre λ aboutissant au meilleur estimateur. Dans les deux cas,

la performance d'un estimateur peut être évaluée via l'erreur de prédiction de la fonction de prédiction associée:

$$MSE(\widehat{\beta}) = \mathbb{E}_{X_0,Y_0}\left\{\left(Y_0 - f_{\widehat{\beta}}(X_0)\right)^2\right\} \ .$$

On définit l'estimateur optimal comme l'estimateur optimisant le critère précédent:

$$\begin{aligned} m^* &= \arg\min_m \ MSE(\widehat{\beta}_m) \quad \text{(OLS)} \\ m^* &= \arg\min_\lambda \ MSE(\widehat{\beta}_\lambda) \quad \text{(Ridge)} \end{aligned}$$

Toutefois, la distribution jointe du couple (X_0, Y_0) étant inconnue, l'erreur de prédiction doit être elle-même estimée. On s'intéresse ici à l'estimateur par validation croisée "Leave-p-Out" (LpO) défini comme suit:

$$R_{LpO}(\widehat{\beta}) = \binom{n}{p}^{-1} \sum_{e\in\mathcal{E}} \left(\frac{1}{p} \sum_{i\notin e} (f^e(x_i) - y_i)^2 \right) \quad , \tag{1}$$

où $\mathcal{E}$ est l'ensemble des sous-échantillons de taille $n-p$ de l'échantillon initial obtenus par tirage sans remise, et f^e désigne la fonction de prédiction $f_{\widehat{\beta}^e}$, où $\widehat{\beta}^e$ est l'estimateur de β obtenu à partir du sous-échantillon e. Le choix de l'estimateur se fait alors par minimisation du critère LpO [1]:

$$\begin{aligned} \widehat{m} &= \arg\min_m \ R_{LpO}(\widehat{\beta}_m) \quad \text{(OLS)} \\ \widehat{m} &= \arg\min_\lambda \ R_{LpO}(\widehat{\beta}_\lambda) \quad \text{(Ridge)} \end{aligned}$$

D'un point de vue algorithmique, la procédure LpO requiert l'ajustement d'un nombre exponentiel en p de modèles, ce qui rend l'utilisation du LpO rédhibitoire dès lors que $p > 1$. L'alternative est alors d'utiliser des procédures approchées de type V-fold, où seule une partie des rééchantillons sont visités. Ces alternatives diminuent sensiblement le temps de calcul, au prix d'une variabilité accrue.

2 Sélection de modèle en régression: cadre "pool-sample"

Bien que l'objectif de l'analyse reste identique (inférer le vecteur de paramètres β) à celui du paragraphe précédent, le cadre "pool-sample" se distingue du cadre classique par la composition de l'échantillon à disposition pour réaliser l'inférence. On suppose maintenant que l'on dispose de N observations pour lesquelles les variables explicatives

x ont été mesurées, mais que la variable réponse Y n'a été observée que pour les n premiers individus. On se place par ailleurs dans le cadre particulier où l'on suppose que $n << p << N$, ce qui représente un cadre intermédiaire entre les cas grande et petite dimension. Ce type de situation peut par exemple correspondre au problème prédiction de la structure 3D d'une protéine : il est aisé de récupérer une grande quantité d'information concernant la séquence de la protéine, mais obtenir sa structure (ou un score quantitatif associé à cette structure) nécessite un travail en laboratoire long et coûteux. On dispose ainsi de milliers de protéines pour lesquelles la séquence est renseignée mais seul un petit nombre parmi elles ont une structure 3D connue.
Le cadre "pool-sample" suggère de nouveaux estimateurs OLS et Ridge pour le vecteur des coefficients de regréssion, qui s'écrivent comme suit:

$$\widehat{\beta} \;=\; \tfrac{N}{n}(X_N^T X_N)^{-1} X_n^T Y \quad \text{et} \quad \widehat{\beta} \;=\; \tfrac{N}{n}(X_N^T X_N + \tfrac{N}{n}\lambda I)^{-1} X_n^T Y$$

où X_n et X_N sont les matrices composées respectivement des individus pour lesquels la réponse est observée, et de l'ensemble des individus. On utilise donc ici l'ensemble des données pour lesquelles l'information sur x est disponible pour estimer la matrice de variance-covariance de x.

On s'intéresse maintenant à l'adaptation de la procédure de validation croisée LpO pour les nouveaux estimateurs suggérés. On propose ici d'utiliser l'estimateur LpO adapté suivant :

$$\begin{aligned} R_{yCV}(f) &= \tbinom{n}{p}^{-1} \sum_{e\in\mathcal{E}} \left(\frac{1}{p} \sum_{i\notin e} \left(x_i \widehat{\beta}^e - y_i \right)^2 \right) \\ &= \tbinom{n}{p}^{-1} \sum_{e\in\mathcal{E}} \left(\frac{1}{p} \sum_{i\notin e} \left(x_i \frac{N}{n-p} M (X^e)^T y^e - y_i \right)^2 \right) \end{aligned}$$

où la matrice M est une matrice fixe ne dépendant pas de l'échantillon e. La matrice M correspond à $(X_N^T X_N)^{-1}$ pour l'estimateur OLS ou $M = (X_N' X_N + \frac{N}{n-p}\lambda I)^{-1}$ pour l'estimateur Ridge. Notons que l'estimateur LpO proposé ne correspond pas tout à fait à la déclinaison directe de la procédure LpO classique (1) dans le cadre pool-sample. En effet, si toutes les données à disposition sont utilisées pour estimer la matrice de variance covariance de x dans l'échantillon complet, alors seules $N - p$ observations devraient être employées pour cette même estimation lorsque des différents rééchantillonnages. On suppose donc ici que (i) soit seules les réponses de p observations sont retirées lors du rééchantillonnage mais que l'information sur x est conservée sur les N données, (ii) soit la matrice $(X_{N-p}^e)' X_{N-p}^e$ qui devrait apparaître a été approchée par la matrice complète $X_N' X_N$. Dans les deux cas, le rééchantillonnage ne portant que sur la partie de l'échantillon initiale composée des observations pour lesquelles la réponse y est disponible, la procédure de validation croisée proposée sera désignée par yLpO dans la suite.

3 Forme close pour la procédure yLpO

La principale contribution de ce travail consiste en une forme close et explicite pour la procédure yLpO. On considère la famille des quatres estimateurs présentés dans la section précédente, et dont la forme générique est

$$\widehat{\beta} = \frac{N}{n-p} M X^T Y \ .$$

Proposition 1. *Le critère yLpO a pour expression*

$$\begin{aligned} R_{yCV}(f) &= \frac{1}{n}\sum_{i=1}^{n} y_i^2 - \frac{2}{n-1} Y' X\widehat{\beta} + \frac{2N}{n(n-1)}\sum_{i=1}^{n} y_i^2 \phi_{ii} \\ &\quad + \frac{N^2}{n(n-1)(n-2)} \left\{ \frac{n-p-1}{n-p} \left(\frac{n^2}{N^2} ||X\widehat{\beta}||_2^2 - 2\sum_{i=1}^{n}\sum_{j=1}^{n} y_i y_j \phi_{ii}\phi_{ij} + \sum_{i=1}^{n} y_i^2 \phi_{ii}^2 \right) \right. \\ &\quad \left. + \frac{p-1}{n-p} \left(\sum_{i=1}^{n}\sum_{j=1}^{n} y_j^2 \phi_{ij}^2 - \sum_{i=1}^{n} y_i^2 \phi_{ii}^2 \right) \right\} \end{aligned}$$

où ϕ_{ij} est le terme générique de la matrice Φ définie par $\Phi = XMX'$.

Il est ainsi possible de calculer de manière exacte le critère yLpO sans en payer le coût algorithmique. Dans le cas de l'estimateur Ridge, il est de surcroît possible d'obtenir le chemin de régularisation associé à ce critère. Nous présenterons quelques applications du critère yLpO à des données simulées afin de décrire son comportement en temps que critère de sélection de modèles.

Bibliographie

[1] M. Stone. Cross-validatory choice and assessment of statistical predictions. J. Roy. Statist. Soc. Ser. B, 36 : 111 − 147, 1974.