REVUE DE STATISTIQUE APPLIQUÉE

JEAN-DANIEL ROLLE

Ordonner les modèles en régression linéaire à l'aide de la géométrie

Revue de statistique appliquée, tome 48, n° 2 (2000), p. 35-53

<http://www.numdam.org/item?id=RSA_2000__48_2_35_0>

Rev. Statistique Appliquée, 2000, **XLVIII** (2), 35-53

ORDONNER LES MODÈLES EN RÉGRESSION LINÉAIRE À L'AIDE DE LA GÉOMÉTRIE

Jean-Daniel Rolle

HEC-Management Studies,
Université de Genève et Haute Ecole de Gestion, Fribourg

RÉSUMÉ

Nous proposons une méthode géométrique de sélection de variables en régression linéaire basée sur la notion d'inertie. A chaque modèle est associé un couple $(\mathcal{N}_X, \mathcal{N}_Z)$ de nuages de données. Le nuage $\mathcal{N}_X$ (resp. $\mathcal{N}_Z$) est formé à partir de l'observation des prédicteurs (resp. de toutes les variables : prédicteurs et réponse) sur les individus. Un bon modèle sera celui pour lequel (i) la distance (inertie) de $\mathcal{N}_Z$ à un hyperplan est petite (auquel cas nous dirons que le modèle est *cohérent*) et (ii) la distance de $\mathcal{N}_X$ à un hyperplan est grande (auquel cas nous dirons que le modèle est *stable*). Un critère se propose d'allier (i) et (ii), afin de permettre un classement des modèles du meilleur au moindre en termes de conformité à ces deux aspects. Des exemples réels ainsi que des simulations montrent que la méthode est bien plus performante qu'une méthode aussi connue et utilisée que le C_p de Mallows.

Mots-clés : *Régression linéaire, sélection de variables, inertie de nuages par rapport à un hyperplan.*

ABSTRACT

We propose a geometrical method of variable selection in regression analysis. The method is based on inertia. A couple $(\mathcal{N}_X, \mathcal{N}_Z)$ of data clouds is associated with each model. The cloud $\mathcal{N}_X$ (resp. $\mathcal{N}_Z$) is built from the observation of predictors (resp. of all the variables : predictors and response) on individuals. A good model will be one for which (i) the distance (inertia) from $\mathcal{N}_Z$ to a hyperplane is small (in which case we say that the model is *coherent*) and (ii) the distance from $\mathcal{N}_X$ to a hyperplane is large (in which case we say that the model is *stable*). A criterion intends to mix (i) and (ii) to easily compare and order the models. Simulation and implementation on various real data sets show that the method tends to outperform the Mallows' C_p criterion.

Keywords : *Linear regression, variable selection, inertia of data clouds with respect to a hyperplane.*

1. Introduction et motivation

Les critères donnés plus bas se placent dans une logique de théorie de l'approximation et aucune hypothèse distributionnelle n'est nécessaire à leur niveau. Les notions de variance, covariance, ou de corrélation, utilisées dans cet article sont celles de l'échantillon. Le problème de la sélection de variables en régression linéaire revient pour l'analyste à découvrir, parmi un ensemble de variables potentiellement explicatives d'une variable appelée «réponse», un sous-ensemble adéquat. Au lieu de variable explicative, nous nous permettrons d'utiliser plutôt l'anglicisme «prédicteur». Outre le fait qu'il n'y a pas objectivement de «meilleur» modèle (par modèle nous entendons une réponse et un sous-ensemble des prédicteurs potentiels), la complexité du problème tient à un conflit d'objectif entre l'accroissement de la capacité prédictive ou informative d'un modèle par l'adjonction de prédicteurs supplémentaires et l'observation qu'un tel accroissement augmente l'incertitude de la prévision. Dans la terminologie utilisée ci-après, cela revient au conflit entre la cohérence d'un modèle (qui croît généralement avec l'adjonction de nouveaux prédicteurs) et sa stabilité (qui décroît avec elle). Il existe un nombre considérable de techniques de sélection de variables, notamment référencées dans Montgommery et Peck (1992); citons à titre d'exemple une méthode très utilisée en pratique – le C_p de Mallows – et sa version robuste, le RC_p de Ronchetti et Staudte (1994). La méthode du C_p repose sur la minimisation de l'erreur quadratique moyenne des réponses calculées (les $\hat{y}_i$ de la littérature). Les critères présentés ici s'inscrivent dans l'optique de l'investigation de tous les modèles possibles et se basent sur des caractéristiques propres au nuage des données. Une étude de simulation présentée dans la section 4 démontrera la nette supériorité d'un des critères sur le C_p et la méthode du R^2 ajusté. Dans son article portant sur la recherche de modèles de régression par l'analyse en composantes principales (ACP), Hawkins (1973) relève, dans une approche géométrique de la sélection de variables, l'utilité des composantes principales et introduit implicitement la notion d'inertie. L'idée de la méthode présentée ici repose de manière fondamentale sur la proximité à la *conjonction* des nuages $\mathcal{N}_X$ et $\mathcal{N}_Z$ définis ci-après. Par conjonction d'un nuage, nous entendons la disposition de ses points sur un hyperplan (voir la définition 1 ci-dessous). Pour préciser cette idée, considérons une matrice U d'observations de k variables sur n individus. Désignons par U_c la matrice recueillant les observations des variables centrées sur les individus. Soit L_i' la i-ème ligne de U, soit $\mathcal{N} = \{L_i \in \mathbb{R}^k; i = 1, \ldots, n\}$ le nuage des points L_i , et soit g son centre de gravité.

Définition 1 *On dira que* $\mathcal{N}$ *est en* conjonction *dans* $\mathbb{R}^k$ *si* $\mathrm{r}(U_c) = k - 1$, $\mathrm{r}(U_c)$ *désignant le rang de* U_c.

$\mathcal{N}$ est donc en conjonction dans $\mathbb{R}^k$ lorsque les vecteurs $L_1, \ldots, L_n$ $(n \geqslant k)$ génèrent un hyperplan affine

$$H = \mathcal{M}(U_c') + g,$$

où $\mathcal{M}(U_c')$ désigne l'espace-colonnes de la matrice U_c'. Cela signifie que les points de $\mathcal{N}$ sont disposés sur H. Si la notion de conjonction pour un nuage $\mathcal{N}$ est bien définie, il n'en va pas de même pour les notions de conjonction approximative, ou encore

d'éloignement par rapport à la conjonction. Pour illustrer ce point de vue, supposons que le sous-espace affine de dimension minimale contenant $\mathcal{N}$ soit $\mathbb{R}^k$ lui-même (et donc que $\mathrm{r}(U_c) = k$); dans ce cas, $\mathcal{N}$ n'est pas en conjonction (les points de $\mathcal{N}$ ne sont pas disposés sur un hyperplan). Si les points de $\mathcal{N}$ se trouvent approximativement sur un hyperplan h, un léger déplacement de tout ou partie de ces points suffira à les mettre en conjonction sur h. Il en ira tout autrement si les points de $\mathcal{N}$ sont fortement dispersés autour de h. La définition de h n'est évidemment pas unique. On pourrait par exemple s'aider de la méthode des moindres carrés en régressant une des variables sur les autres, ce qui reviendrait à effectuer un déplacement des points de $\mathcal{N}$ parallèlement à l'axe de la variable jouant le rôle de la réponse. On pourrait aussi utiliser les techniques de l'ACP sur les données centrées pour trouver un autre hyperplan. Un autre hyperplan encore serait fourni grâce à l'ACP sur les données centrées et réduites. D'autres méthodes fourniraient d'autres hyperplans. On s'aperçoit donc que le concept de proximité (ou d'éloignement) d'un nuage à la conjonction ne peut être défini de manière univoque. Des caractérisations utiles de la notion de conjonction seront données dans la section 3.

Désignons par $\mathcal{N}_Z$ un nuage de points $z_i = (x_i', y_i)'$, $i = 1, \ldots, n$, où les x_i' sont des vecteurs-ligne d'ordre p d'observations sur des prédicteurs, et les y_i des observations sur une réponse. Soit $\mathcal{N}_X$ le nuage des points prédicteurs x_i. Appellons V le sous-espace vectoriel de $\mathbb{R}^{p+1}$ de dimension p généré par les vecteurs $e_1, \ldots, e_p$ de la base standard de $\mathbb{R}^{p+1}$ (e_i est la $i^{\text{ème}}$ colonne de la matrice identité d'ordre $p+1$), et désignons par $\tilde{\mathcal{N}}_X$ le nuage des points $(x_i', 0)'$. Or $\tilde{\mathcal{N}}_X$ n'est autre que la projection orthogonale de $\mathcal{N}_Z$ sur V. Nous dirons que $\mathcal{N}_Z$ (ou le modèle associé) est *instable* si $\tilde{\mathcal{N}}_X$ est proche de la conjonction dans V ou, de manière équivalente, si $\mathcal{N}_X$ est proche de la conjonction dans $\mathbb{R}^p$. Dans le tableau 1, nous envisageons de manière informelle trois types de situations.

TABLEAU 1
Les trois types de modèles intervenant en sélection de variables

catégorie de modèle	$\mathcal{N}_X$ proche de la conjonction	$\mathcal{N}_Z$ proche de la conjonction
1	non	oui
2	non	non
3	oui	oui

Les bons modèles se trouvent dans la catégorie 1. Les modèles des autres catégories sont à exclure : imaginons qu'on ajuste à $\mathcal{N}_Z$ un hyperplan des moindres carrés. Dans le cas 2, la somme des carrés résiduelle sera (relativement) importante et l'ajustement mauvais. Les modèles de la catégorie 3 présenteront de la multicolinéarité (nous reviendrons sur ce point important dans la section 3) et sont dès lors également à proscrire. Le critère défini ne devra retenir que les modèles de la catégorie 1, à savoir ceux correspondant à un $\mathcal{N}_Z$ «proche» de la conjonction et qui ne soient pas instables. Première difficulté : comment mesurer le degré d'éloignement d'un nuage à la conjonction? Notre solution est bien naturelle : on ajuste au mieux un hyperplan au

nuage en minimisant l'inertie, laquelle dépend d'une norme et d'une projection (voir ci- dessous). On prend ensuite l'inertie relative (Q_Z) du nuage $\mathcal{N}_Z$ par rapport à cet hyperplan optimal comme mesure de la proximité du nuage à la conjonction. On fait de même pour $\mathcal{N}_X$ et on prend l'inertie relative (Q_X) comme mesure de l'instabilité de $\mathcal{N}_Z$. La deuxième difficulté est de combiner judicieusement Q_X et Q_Z de manière à obtenir un critère performant.

2. Un critère bâti sur l'inertie

La méthode est basée sur une propriété de distance à la conjonction qu'il est commode de quantifier à l'aide la notion d'inertie d'un nuage de points par rapport à un hyperplan optimal, c'est-à-dire passant au mieux (dans un sens à définir) auprès du nuage. Sans restreindre la généralité, nous supposons dans cette section que les variables ont été centrées (le centre de gravité de $\mathcal{N}_Z$ est donc l'origine). Nous allons utiliser ici deux types de normes : pour un vecteur $w = (w_1, \ldots, w_k)'$ de $\mathbb{R}^k$, nous définissons la norme *de la dernière coordonnée* comme $||w|| = |w_k|$ (qui est à vrai dire une semi-norme). L'autre norme est $||w|| = (w'T'Tw)^{1/2}$, où T est la matrice non singulière d'ordre k d'une transformation linéaire de $\mathbb{R}^k$. C'est la norme héritée du produit scalaire $< v, w >= v'T'Tw$ utilisé lors de la mise en œuvre de l'ACP, lorsque le vecteur des variables est transformé par une matrice T. Si C est la matrice de covariance des variables originales, celle des variables transformées sera TCT'. Lorsque les variables sont réduites, T est la matrice diagonale d'ordre k dont les éléments diagonaux sont les inverses des écarts-types des variables. Dans ce cas, TCT' est la matrice de corrélation des variables. Nous verrons ci-dessous que le choix définitif de la norme se fera en fonction de la performance du critère dans l'étude de simulation.

Par espace de travail, nous entendons un triplet $(\mathbb{R}^{p+1}, ||\cdot||, P_v)$, autrement dit $\mathbb{R}^{p+1}$ doté d'une norme et d'une projection P_v sur $\mathcal{M}(v)$, où $\mathcal{M}(v)$ est l'espace-colonnes d'une matrice v de dimension $(p+1) \times p$ et de rang p. Dans les trois premières colonnes du tableau 2, nous définissons deux types d'espaces de travail; le vecteur z dont il est question dans la deuxième colonne est défini par $z = (x', y)' \in \mathbb{R}^{p+1}$, avec $x \in \mathbb{R}^p$, I_p est la matrice identité d'ordre p et e_{p+1} est la dernière colonne de la matrice identité d'ordre $p+1$. La notation $[A|B]$ est utilisée pour une matrice formée par les blocs horizontaux A et B.

Soit un espace de travail $(\mathbb{R}^{p+1}, ||\cdot||, P_v)$ et soit l'hyperplan $\Delta = \mathcal{M}(v)$. L'inertie[1] de $\mathcal{N}_Z$ par rapport à l'origine et par rapport à Δ est définie respectivement par

$$I(\mathcal{N}_Z, 0) = \frac{1}{n-1} \sum_{i=1}^{n} ||z_i||^2 \qquad (1)$$

[1] Généralement l'inertie est définie dans le contexte moins général d'une norme héritée d'un produit scalaire, ce qui n'est pas nécessairement le cas ici.

TABLEAU 2

Récapitulation de normes et projections utilisées pour le critère

Espace de travail			Solution optimale	
Type	Norme	Projection sur $\mathcal{M}(v)$	Matrice v^* t.q. l'inertie de $\mathcal{N}_Z$ à $\mathcal{M}(v^*)$ soit min.	Qual. glob. de $\mathcal{M}(v^*)$ (i.e. Q_Z)
1	$\|\|z\|\|^2 = y^2$	$P_v = v[I_p\|0][v\|e_{p+1}]^{-1}$ (parallèle à e_{p+1})	$v^* = \begin{bmatrix} I_p \\ \hat{\beta}' \end{bmatrix}$	R^2
2	$\|\|z\|\|^2 = z'T'Tz$	$P_v = v(v'T'Tv)^{-1}v'T'T$ ($T'T-$orthogonale)	$v^* = T^{-1}[v_1^*, \ldots, v_p^*]$	$\frac{\lambda_1 + \cdots + \lambda_p}{\mathrm{tr}TC_ZT'}$

et

$$I(\mathcal{N}_Z, \Delta) = \frac{1}{n-1} \sum_{i=1}^{n} ||z_i - P_v z_i||^2. \tag{2}$$

Dans les espaces de travail 1 et 2, les hyperplans optimaux sont donnés par $\Delta^* = \mathcal{M}(v^*)$, où la matrice optimale v^* est à lire dans la quatrième colonne du tableau 2; $\hat{\beta}$ est le vecteur des coefficients estimés de la régression sur les données centrées. Les vecteurs $v_1^*, \ldots, v_p^*$ sont des vecteurs propres orthonormaux associés aux p plus grandes valeurs propres $\lambda_1 > \ldots > \lambda_p$ de TC_ZT', où T est une matrice non singulière de transformation d'ordre $p+1$ et C_Z la matrice de covariance des variables $X_1, \ldots, X_p, Y$.

Pour les espaces de travail de type 1 et 2, la norme $||\cdot||$ et la projection sont telles qu'on a la propriété de décomposition de l'inertie

$$I(\mathcal{N}_Z, 0) = I(\mathcal{N}_Z^*, 0) + I(\mathcal{N}_Z, \Delta^*), \tag{3}$$

où $\mathcal{N}_Z^*$ est le nuage $\mathcal{N}_Z$ projeté (par P_{v^*}) sur Δ^*. L'inertie totale est ainsi décomposée en l'inertie expliquée par l'hyperplan Δ^* (premier terme de droite) plus l'inertie résiduelle de $\mathcal{N}_Z$ autour de Δ^*. On mesure la qualité globale de l'hyperplan Δ^* par la part d'inertie expliquée

$$Q_Z = I(\mathcal{N}_Z^*, 0)/I(\mathcal{N}_Z, 0). \tag{4}$$

La qualité globale Q_Z correspondant aux ensembles de travail de type 1 et 2 figure dans la dernière colonne du tableau 2. Elle est obtenue à partir de l'ensemble des variables $\mathcal{Z} = \{X_1, \ldots, X_p, Y\}$, et est l'outil que nous allons utiliser pour mesurer le degré de conjonction de $\mathcal{N}_Z$. Nous pouvons faire de même et calculer une qualité globale Q_X obtenue à partir de l'ensemble $\mathcal{X} = \{X_1, \ldots, X_p\}$ des seuls prédicteurs. Ainsi, Q_X sera une mesure du degré de conjonction de $\mathcal{N}_X$ et donc une mesure du degré d'instabilité de $\mathcal{N}_Z$. Le critère donné plus bas dans l'équation (6) dépend de Q_X et de Q_Z, qui peuvent être définis de multiples manières en fonction de la norme et de la projection choisies. Même si le principe commun est le degré d'inertie relative,

tous les choix ne sont pas heureux pour la sélection de variables. De nombreuses possibilités ont été testées et rejetées sur la base d'une étude de simulation (décrite dans la section 4). Nous avons toutefois trouvé une façon très satisfaisante de définir Q_X et Q_Z (il s'agit du critère f_2 défini plus loin, et qui se révèle bien plus performant que le C_p). Cependant, il est possible que le choix d'autres normes et d'autres projections amène à des résultats encore supérieurs. Nous désignerons par

$$\mathcal{S} = [1, X_1, \ldots, X_p; Y] \tag{5}$$

un modèle de régression avec constante, où Y est la réponse et $X_1, \ldots, X_p$ sont des prédicteurs choisis dans l'ensemble des prédicteurs potentiels. Soit Q_Z (resp. Q_X) la qualité globale (4) pour $\mathcal{Z}$ (resp. $\mathcal{X}$), deux ensembles définis ci-dessus. Nous dirons que le modèle $\mathcal{S}$ de (5) est d'autant plus *cohérent* (resp. *instable*) que Q_Z (resp. Q_X) est proche de 1. Des caractérisations de l'instabilité et de la cohérence parfaite sont l'objet des propositions 2 et 3, de la section 3.

La stratégie développée dans cet article s'attache à classer les modèles du meilleur au moindre en fonction des deux propriétés de cohérence et de stabilité, en combinant de manière adéquate Q_X et Q_Z. En rapport avec la motivation géométrique résumée dans le tableau 1 de la section précédente, nous définissons un critère prenant des valeurs faibles sur les modèles de la catégorie 1 et des valeurs fortes sur ceux des autres catégories : soit $\mathcal{S}$ un modèle de régression, soit $\alpha \in [0\,,1]$ un nombre réel et, pour $Q_X \in [0\,,1[$, soit la fonction (convexe et non croissante) de α

$$f(\alpha; \mathcal{S}) = \frac{[1 - Q_Z]^{\alpha}}{[1 - Q_X]^{1-\alpha}}. \tag{6}$$

La fonction f sera appelée par la suite *fonction de choix*. Un modèle $\mathcal{S}_1$ est dit *meilleur* que $\mathcal{S}_2$ au niveau α si

$$f(\alpha; \mathcal{S}_1) < f(\alpha; \mathcal{S}_2).$$

Lorsque $Q_Z = 1$, cette fonction est identiquement zéro. Lorsque $Q_Z < 1$, elle est strictement décroissante sur l'intervalle $[0\,,1]$.

Le graphe de la fonction (6) est représenté dans la figure 1. À un $\mathcal{S}$ instable correspondra une forte valeur de $1/(1 - Q_X)$ (un nombre toujours plus grand que 1). À un $\mathcal{S}$ cohérent correspondra un Q_Z proche de 1. La stabilité n'est pas une fin en soi; ce qu'on recherche, c'est un modèle aussi cohérent que possible, apte à satisfaire à un besoin d'interprétation et de prévision, mais ayant une stabilité suffisante. C'est pourquoi nous avons fait le choix d'une fonction (6) convexe sur $[0, 1]$, pour laquelle nous nous intéressons plutôt aux valeurs sur la droite du graphique, disons dans l'intervalle $[0.5, 1]$. Ce procédé nous permet de valoriser Q_Z au détriment de Q_X, autrement dit de donner plus de poids à la cohérence qu'à la stabilité. Nous verrons dans les applications de la section 4 que ces choix pragmatiques, soucieux de la commodité de l'outil graphique, apparaissent judicieux. Nous sommes maintenant à même de voir comment le critère (6) conduit à l'élimination des modèles des catégories 2 et 3 du tableau 1. Si le modèle est dans la catégorie 2, Q_Z sera petit et $f(\alpha; \mathcal{S})$ aura de fortes valeurs sur la droite du graphique. Si le modèle est dans

la catégorie 3, Q_X sera grand et $f(\alpha; \mathcal{S})$ aura de fortes valeurs sur la gauche du graphique. Au contraire, si le modèle est dans la catégorie 1, $f(\alpha; \mathcal{S})$ aura de faibles valeurs aussi bien à gauche qu'à droite du graphique. Notons qu'un modèle $\mathcal{S}_1$ peut être meilleur (au sens de (6)) qu'un modèle alternatif $\mathcal{S}_2$ pour une valeur donnée de α et lui être inférieur pour une autre valeur. Le choix d'une valeur plus ou moins élevée pour α dépend de l'usage qu'on veut faire de notre analyse de régression (contrôle, prévision, etc.). Nous reparlerons du choix de α dans la section 4. Notons simplement qu'un α élevé (disons $\alpha = 0.9$) donne un poids important à la cohérence du modèle associé à un nuage $\mathcal{N}_Z$, alors qu'un α faible (disons $\alpha = 0.5$) donne un poids important à la stabilité. L'observation de la fonction de choix sur la figure 1 permet d'apprécier le comportement de $\mathcal{S}$ vis-à-vis d'autres modèles tant du point de vue de la cohérence (sur la droite du graphique) que de celui de la stabilité (sur la gauche). Cette visualisation simultanée est un réel avantage pratique.

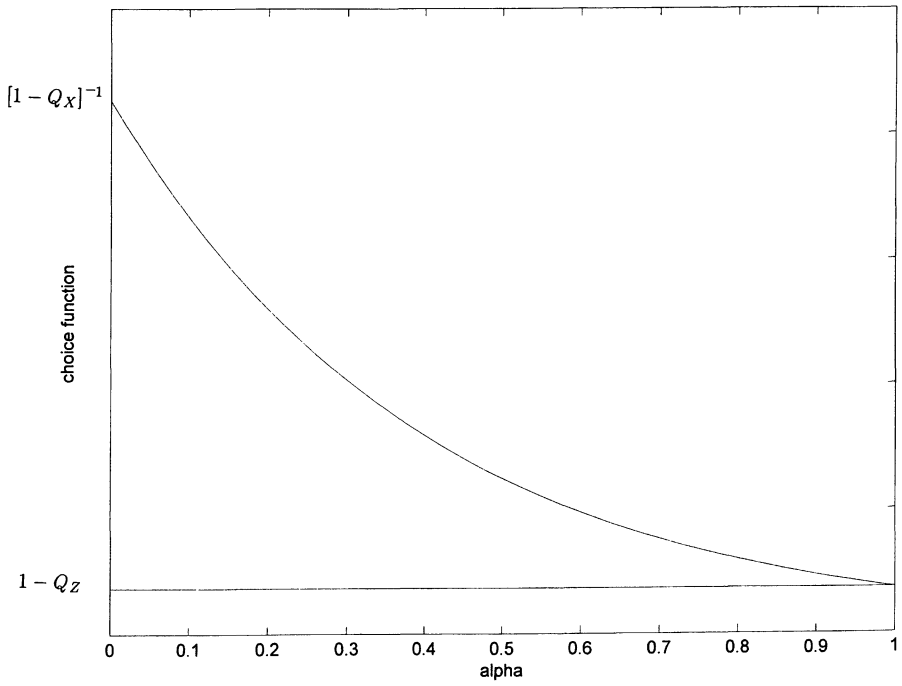

FIGURE 1
Graphe de la fonction de choix d'un modèle.
La fonction varie entre $1 - Qz$ *(minimum) et* $1/(1 - Qx)$ *(maximum).*
Ces deux bornes sont indépendantes de α.
En cas d'instabilité du modèle, $1/(1 - Qx)$ *sera grand.*
Au cas où le modèle est cohérent, $1 - Qz$ *sera petit.*

Si on utilise un espace de travail de type 2 (voir tableau 2) avec la matrice identité comme matrice de transformation, (6) devient

$$f_1(\alpha;\mathcal{S}) = \frac{[\lambda_{\min}(C_Z)/\mathrm{tr}C_Z]^{\alpha}}{[\lambda_{\min}(C_X)/\mathrm{tr}C_X]^{1-\alpha}} \quad , \tag{7}$$

où C_Z (resp. C_X) est la matrice de covariance des variables de $\mathcal{Z}$ (resp. $\mathcal{X}$), et où $\lambda_{\min}(\cdot)$ désigne la plus petite valeur propre. Défini ainsi, le critère n'est pas invariant pour un changement d'échelle dans la mesure des variables. A cet égard, soit U une matrice $n \times k$ d'observations de k variables $U_1, \ldots, U_k$ sur n individus et soit C la matrice de covariance de ces variables. Soit $s_1^2, \ldots, s_k^2$ les variances de ces variables. Du point de vue des applications de l'ACP à des données réelles, les inégalités suivantes (Magnus and Neudecker 1988, th. 14 p. 211) revêtent une importance particulière :

$$\lambda_{\min}(C) \leqslant \min\{s_1^2, \ldots, s_k^2\} =: m$$

et

$$\lambda_{\max}(C) \geqslant \max\{s_1^2, \ldots, s_k^2\} =: M.$$

On a donc

$$q = \frac{\lambda_{\min}(C)}{\mathrm{tr}C} \leqslant \frac{\lambda_{\min}(C)}{\lambda_{\max}(C)} \leqslant \frac{m}{M}. \tag{8}$$

En conséquence, q peut être rendu arbitrairement petit en fonction du choix de l'unité dans laquelle on mesure les variables (qui est celle de l'écart-type des variables). Cette situation est fréquente lorsque les données sont issues des sciences sociales, où les unités de mesure sont souvent arbitraires (Flury, 1988). Le critère (7) peut donner ici ou là des résultats satisfaisants lorsque les variances des variables sont du même ordre de grandeur, mais il souffre grandement de son absence d'invariance. L'idée de remplacer dans (7) les matrices de covariance par les matrices de corrélation, pour naturelle qu'elle soit, s'est révélée très décevante dans de nombreux cas de figure lors de simulations destinées à trouver un vrai modèle dans l'ensemble des modèles possibles : la méthode s'est souvent montrée très inférieure à celle du C_p. C'est notamment le cas dans l'étude de simulation résumée dans le tableau 3 de la section 4, où nous avons noté $\tilde{f}_1$ le critère (7) basé sur les matrices de corrélation. Il s'agissait dès lors de trouver un autre critère invariant à des changements d'échelle lors de la mesure des variables et capable de battre le C_p. Sa formulation est la suivante :

$$f_2(\alpha;\mathcal{S}) = \frac{[1-R^2]^{\alpha}}{[\lambda_{\min}(R_X)/p]^{1-\alpha}}, \tag{9}$$

où $\lambda_{\min}(R_X)$ est la plus petite valeur propre de la matrice de corrélation des prédicteurs. Le critère f_2 est hybride, en ce qu'il combine un espace de travail de type 1 pour Q_Z et de type 2 avec la matrice de réduction $T = \mathrm{diag}\,(s_1^{-1}, \ldots, s_p^{-1})$ pour Q_X. L'idée de définir un critère à partir de l'inertie est correcte, comme le

montrent les résultats étonnamment bons de l'étude de simulation. Mais il convenait de choisir une bonne norme et une bonne projection dans la définition de Q_Z et de Q_X. D'autres possibilités pour la mesure de l'inertie ont été explorées, sans succès pour l'instant, mais sont l'objet de nouvelles recherches. Nous verrons dans l'étude de simulation que f_2 de (9) surclasse aussi bien f_1 ou $\tilde{f}_1$ de (7), qui sont dès lors à oublier, que les méthodes du C_p et du R^2 ajusté.

3. Caractérisation de l'instabilité et de la cohérence

Cette section donne des précisions d'ordre technique; elle peut être esquivée par le lecteur s'intéressant aux seules applications du critère. Soit U une matrice $n \times k$ d'observations de k variables sur n individus et U_c la matrice recueillant les observations des variables centrées. La matrice U_c est fonction de U de la manière suivante : $U_c = M_1 U$, où $M_1 = I - \mathbf{1}\mathbf{1}'/n$ ($\mathbf{1}$ désignant le vecteur de uns d'ordre n). La matrice de covariance de ces variables est $C = U_c' U_c/(n-1)$. Nous supposons que le rang de U_C (et donc celui de C) est d'au moins $n-1$. Il est utile de rappeler quelques résultats classiques et d'en établir quelques autres. Notamment, nous soulignons clairement le lien entre la dépendance linéaire des colonnes d'une matrice $B = [\mathbf{1}|U]$ et la conjonction (c'est-à-dire la disposition sur un hyperplan) des points associés aux lignes de U. Soit L_i' la $i^{\text{ème}}$ ligne de U, et soit $\mathcal{N}$ le nuage des points L_i, et g son centre de gravité. Nous voulons préciser en termes de dépendance linéaire et de conjonction les notions de cohérence et de stabilité définies dans la section 2. $\mathcal{N}$ est en conjonction lorsque les vecteurs $L_1, \ldots, L_n$ génèrent un hyperplan affine $H = \mathcal{M}(U_c') + g$. Géométriquement, cela signifie que les points de $\mathcal{N}$ sont disposés sur H. Si nous désirons obtenir la représentation cartésienne de H, nous pouvons procéder ainsi : soit $\lambda_1 > \ldots > \lambda_{k-1} > \lambda_{\min}(C)$ les valeurs propres de C. Comme $\mathcal{N}$ est en conjonction dans $\mathbb{R}^k$, $\lambda_{\min}(C) = 0$ (cf proposition 1 ci-dessous). Soit c un vecteur propre associé à la valeur propre nulle. Alors l'équation de l'hyperplan affine passant au plus près de $\mathcal{N}$ au sens de l'ACP sur la matrice de covariance est donnée par

$$\tilde{H} = \{x \in \mathbb{R}^k; c'(x - g) = 0\}. \tag{10}$$

Mais comme $\mathcal{N}$ est en conjonction sur l'hyperplan H, on a forcément que $\tilde{H} = H$. La proposition suivante se révèle très utile.

Proposition 1 *Soit T une matrice non singulière d'ordre k, soit C et U comme ci-dessus, avec $\mathrm{r}(C) \geqslant k-1$, et considérons la matrice $B = [\mathbf{1}|U]$. Les affirmations suivantes sont équivalentes :*

1. $\lambda_{\min}(C) = 0$
2. $\lambda_{\min}(TCT') = 0$
3. *les colonnes de B sont linéairement dépendantes*
4. *$\mathcal{N}$ est en conjonction dans $\mathbb{R}^k$*
5. *le nuage $\mathcal{N}^T$ associé aux lignes de la matrice $U_T = U_C T'$ des données transformées par T est en conjonction dans $\mathbb{R}^k$.*

Soit $\mathcal{N}_Z$ le nuage défini plus haut et soit X (resp. Z) la matrice des observations des prédicteurs (resp. de l'ensemble des variables) sur les individus. Soit R_X (resp. R_Z) les matrices de corrélation correspondantes et soit Q_X (resp. Q_Z) les qualités globales d'hyperplans optimaux dans le cadre d'espaces de travail de type 1 et 2.

Définition 2 *Désignons par* $e_{p+1} = (0 \cdots 0\,1)'$ *le vecteur standard de* $\mathbb{R}^{p+1}$ *engendrant l'axe de la réponse et* h *le vecteur normal à l'hyperplan passant par le centre de gravité de* $\mathcal{N}_Z$ *et orthogonal au dernier axe factoriel dans l'ACP de* $\mathcal{N}_Z$ *sur matrice de covariance. Nous dirons que* $\mathcal{N}_Z$ *est vertical si* $h'e_{p+1} = 0$.

Nous donnons maintenant un ensemble de propriétés caractérisant l'instabilité parfaite d'un modèle.

Proposition 2 *Les propriétés suivantes sont équivalentes :*

1. $Q_X = 1$
2. $\mathcal{N}_X$ *est en conjonction dans* $\mathbb{R}^p$
3. $\lambda_{\min}(C_X) = 0$
4. $\lambda_{\min}(R_X) = 0$
5. *les colonnes de* $[\mathbf{1}|X]$ *sont linéairement dépendantes*
6. $\mathcal{N}_Z$ *est vertical et en conjonction dans* $\mathbb{R}^{p+1}$.

Les propriétés de la proposition 3 caractérisent la cohérence parfaite d'un modèle.

Proposition 3 *Les propriétés suivantes sont équivalentes :*

1. $Q_Z = 1$
2. $\mathcal{N}_Z$ *est en conjonction dans* $\mathbb{R}^{p+1}$
3. $\lambda_{\min}(C_Z) = 0$
4. $\lambda_{\min}(R_Z) = 0$
5. *les colonnes de* $[\mathbf{1}|Z]$ *sont linéairement dépendantes.*

Les énoncés de la proposition 2 impliquent ceux de la proposition 3; mais la réciproque n'est évidemment pas vraie.

4. Exemples et simulation

4.1. Étude de simulation

Nous nous concentrons ici sur la recherche d'un modèle optimal d'au moins 2 prédicteurs ($p \geqslant 2$) : si on a affaire à un modèle d'un seul prédicteur, le dénominateur dans (9) vaut l'unité : l'instabilité n'est alors pas prise en compte dans l'évaluation de ce modèle, et bien que le critère soit défini dans ce cas-là, il perd de sa substance. La non prise en compte des modèles avec $p = 1$ n'est pas un réel problème, car les méthodes de sélection de variables ne revêtent leur véritable utilité que lorsque le nombre de prédicteurs est au moins de trois ou quatre; lorsque ces derniers sont raisonnables, une méthode choisira rarement un modèle d'un seul prédicteur.

Afin d'évaluer la performance de notre procédure, nous avons mis en œuvre une étude de simulation la comparant à deux méthodes couramment utilisées : le C_p de Mallows et le R^2 ajusté (que nous noterons R_a^2). Une matrice de prédicteurs X et des valeurs pour les paramètres furent données. A chaque pas de la simulation, les erreurs furent générées selon une loi normale centrée réduite et des observations y_i furent calculées. Dans un premier temps, nous nous sommes intéressés à une simulation à partir d'une matrice X générée aléatoirement. Une autre simulation fut mise en œuvre à partir de la matrice X des données réelles de Hald, qu'on trouve dans Montgomery et Peck (1992). Les paramètres de simulation (coefficient de régression et écart-type) furent ceux estimés à partir de ces données. Dans l'étude, nous avons ainsi considéré quatre facteurs : la taille de l'échantillon (n), la matrice X, le vrai modèle, et le type de méthode.

Pour la simulation à partir de la matrice artificielle X, la taille de d'échantillon fut $n = 30$. Les colonnes de X (hors la constante) furent générées par une loi uniforme sur [0,1]. Pour la simulation, nous avons utilisé six variables plus la constante. Le premier vrai modèle possède deux variables non nulles (X_2 et X_4); le second vrai modèle en possède quatre (X_1,X_3,X_5 et X_6). Les valeurs des paramètres furent choisies de façon que les statistiques t associées aux coefficients de régression avoisinent 6 (sous la normalité); en ceci et pour la démarche générale de simulation, nous nous sommes inspirés de Ronchetti, Field et Blanchard (1997)[2]. Les résultats de notre étude apparaissent dans le tableau 3.

Pour toutes les simulations, nous avons pris $\alpha = 0.7$. En ce qui concerne la simulation à partir des (vraies) données de Hald, les écarts-types des prédicteurs sont 5.88, 15.56, 6.41, et 16.74. Afin d'égaliser grossièrement les variances des prédicteurs, nous avons choisi de multiplier X_1 par 3 et X_3 par 2, et d'utiliser un facteur d'échelle pour Y situant son écart-type autour de 15. Une telle modification des échelles permet au critère f_1 de se mieux comporter, mais illustre sa faiblesse face à f_2, qui est invariant aux changements d'échelle. Le vrai modèle choisi pour la simulation à partir des données de Hald, est le modèle à deux prédicteurs sur lequel la statistique C_p prend son minimum. Le tableau 3 parle de lui-même quant aux résultats du concours de simulation. Le rang moyen du vrai modèle dont il y est question est obtenu de la manière suivante : à la i-ème des 200 itérations, le vrai modèle se trouve placé en position r_i. Le rang moyen n'est autre que $\sum_{i=1}^{200} r_i/200$. Le nombre de modèles à deux prédicteurs au moins est de $2^6 - 7 = 57$ dans les cas (a) et (b) et de $2^4 - 5 = 11$ dans le cas (c). On observe ainsi qu'en termes de rang moyen, f_2 se révèle la meilleure méthode, alors que $\tilde{f}_1$ et R_a^2 sont inférieurs à des degrés divers dans les trois cas. En termes de placement en première position du vrai modèle, la statistique f_2 se révèle notablement supérieure au C_p. Nous observons que l'utilisation combinée du R^2 et d'une mesure de la multicolinéarité aident à découvrir efficacement le vrai modèle. Nous avons également mesuré la corrélation entre f_2 et C_p en nous y prenant de la manière suivante : examinons le cas (a) du tableau 3. A chaque itération i, le couple (f_2, C_p) est mesuré sur chacun des 57 modèles et on calcule sur cette base le coefficient de corrélation ρ_i entre f_2 et C_p. La corrélation $\overline{\rho}$ n'est autre que la moyenne des 200 ρ_i. On observe une corrélation

[2] Ces auteurs comparent dans une étude de simulation leur technique robuste de choix de modèle à la technique non robuste de «cross-validation» de Shao (1993).

TABLEAU 3

Résultat du concours de simulation établi sur 200 itérations. Les parties (a) *et* (b) *sont basées sur une matrice X dont les colonnes (à part la constante) furent générées par une loi uniforme. La partie* (c) *est établie à partir d'un jeu de données réelles où les paramètres intervenant dans la simulation sont les paramètres estimés à partir des données réelles. Lorsqu'une méthode fonctionne correctement, elle place le vrai modèle en position 1. Le nombre de fois où les méthodes ont placé le vrai modèle en position 1, 2 et 3 est donné dans le tableau, de même que la position moyenne sur les 200 itérations et une mesure $\overline{\rho}$ du degré de corrélation entre les méthodes C_p et f_2.*

	(*a*) vrai modèle : 24 $n = 30$; $X =$ unif[0,1] 57 modèles					(*b*) vrai modèle : 1356 $n = 30$; $X =$ unif[0,1] 57 modèles				
vrai modèle en	f_1	$\tilde{f}_1$	f_2	R_a^2	C_p	f_1	$\tilde{f}_1$	f_2	R_a^2	C_p
position 1	166	123	172	35	98	169	110	189	89	146
position 2	19	21	14	16	30	16	44	8	35	27
position 3	8	2	6	18	16	14	21	3	52	19
autres	7	54	8	131	56	1	25	0	24	8
rang moyen	1.46	3.57	1.36	6.77	3.44	1.20	1.89	1.07	2.05	1.45
$\overline{\rho}$			0.90					0.87		

	(*c*) vrai modèle : Hald $n = 13$; $X =$ Hald 11 modèles				
vrai modèle en	f_1	$\tilde{f}_1$	f_2	R_a^2	C_p
position 1	140	57	179	81	125
position 2	41	73	20	20	17
position 3	17	49	1	25	14
autres	2	21	0	74	44
rang moyen	1.42	2.23	1.15	2.82	1.93
$\overline{\rho}$			0.89		

relativement importante entre les deux méthodes. Incidemment, ce fait semble établir la forte composante géométrique inhérente à la méthode du C_p, ce qui ne constitue pas une surprise. Ces simulations nous indiquent que la méthode f_2 est établie sur des bases solides. Elle est largement préférable à f_1, car elle évite le recours à des changements d'échelle des variables lorsque celles-ci présentent des variances de

TABLEAU 4

Comparaison entre C_p et f_2 selon le rang moyen du vrai modèle, dans le cadre des trois cas (a), (b) et (c) dans le tableau 3.

	rang moyen					
	f_2			C_p		
α	(a)	(b)	(c)	(a)	(b)	(c)
.5	1.06	1.17	1.07	3.44	1.45	1.93
.6	1.14	1	1.12	3.44	1.45	1.93
.7	1.36	1.07	1.15	3.44	1.45	1.93
.8	2.29	1.30	1.18	3.44	1.45	1.93
.9	4.55	2.01	1.74	3.44	1.45	1.93

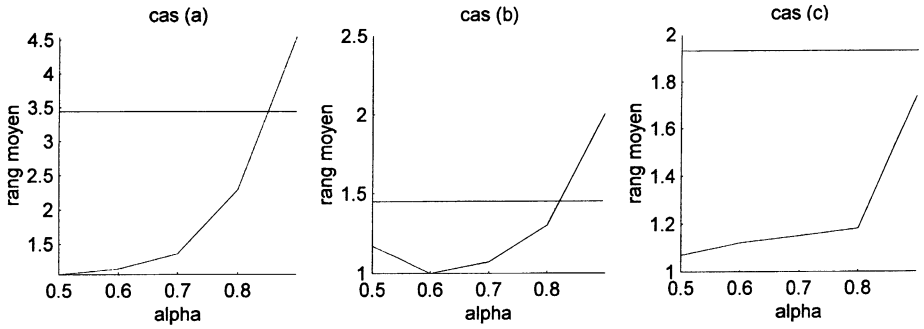

FIGURE 2

Illustration du comportement de f_2 en fonction de α pour les cas (a), (b) et (c) du tableau 4. Le rang moyen obtenu lors de 200 simulations est exprimé comme une fonction de α. Dans les trois cas l'intervalle contenant les valeurs de α pour lesquelles f_2 surpasse C_p est d'une étendue considérable.

différentes magnitudes. D'autres simulations montrent que même dans le cas idéal pour f_1 de variances approximativement égales pour les variables, f_2 lui est supérieur dans la capacité à retrouver le vrai modèle. Le critère invariant $\tilde{f}_1$ est décevant et doit être oublié lui aussi, puisqu'il se comporte moins bien que f_2. Concentrons-nous donc sur la performance de ce dernier par rapport au C_p. On pourrait se demander si la supériorité du premier critère sur le second demeure pour des choix différents de $\alpha = 0.7$.

Le tableau 4 et la figure 2 qui lui est associée visent à répondre à cette question.

Pour chacune des valeurs $\alpha = 0.5, 0.6, 0.7, 0.8$ et 0.9, la comparaison entre f_2 et C_p a été faite sur la base de 200 itérations. Le critère de comparaison entre f_2 et

C_p retenu dans ce tableau est le rang moyen du vrai modèle. La comparaison s'établit pour les trois cas (a), (b) et (c) du tableau 3.

Les valeurs du C_p, qui bien sûr ne fluctuent pas en fonction de α, sont de 3.44 pour (*a*), 1.45 pour (*b*) et 1.93 pour (*c*). Regardons par exemple le cas (*b*) de la figure 2. On observe logiquement que f_2 se comporte le mieux pour des valeurs intermédiaires de α, disons entre 0.55 et 0.75. La situation se dégrade pour de grandes valeurs de α (auquel cas le critère se rapproche de plus en plus du critère (insuffisant) du seul R^2; pour $\alpha = 1$, f_2 ne tient compte que de R^2). A l'inverse, pour de petites valeurs de α, f_2 donne un poids prépondérant à la stabilité dans le choix qu'il propose. Mais c'est dans la prise en compte équilibrée des deux éléments de cohérence et de stabilité (et donc l'utilisation de valeurs intermédiaires pour α) que f_2 donne sa pleine mesure pour trouver le vrai modèle. On note que f_2 est supérieur à C_p pour des α se situant dans une fourchette très vaste. Il est piquant de noter que dans le cas (*b*) et pour $\alpha = 0.6$ f_2 a trouvé à chaque fois le vrai modèle lors de la simulation. Si on regarde le cas (*a*), on observe une dégradation du f_2 pour des valeurs proches de 0.9. C'est dans le cas (*c*) que la fourchette des α assurant la supériorité de f_2 est le plus étendue.

4.2. Exemples à partir de données réelles

La méthode f_2 a été appliquée à une douzaine de jeux de données réelles à partir de modèles ayant de 3 à 12 prédicteurs potentiels. Pour illustrer son fonctionnement, nous allons utiliser un premier jeu de données intitulé «Air pollution in US cities» et fourni par Sokal et Rohlf (1981). Ce jeu de 41 observations met en relation une réponse (concentration de SO_2 dans l'air) avec six prédicteurs de type météorologique et socio-économique. Pour une valeur donnée de α (nous avons choisi ici, pour comparaison, $\alpha = 0.85$ et $\alpha = 0.95$), la fonction $f_2(\alpha; \mathcal{S})$ de (9) permet d'ordonner les modèles du meilleur (celui pour lequel la valeur de la fonction en α est la plus petite) au moindre (celui pour lequel cette valeur est la plus grande). Nous supposons pour l'exemple que le but de l'analyse de régression est une prévision sans forte extrapolation. Dans un tel cas de figure, la cohérence importe plus que la stabilité et des valeurs élevées de α sont raisonnables. Pour la mise en œuvre, nous avons construit un programme MATLAB effectuant le classement des modèles et traçant les graphes des fonctions de choix leur étant associées.

Dans le tableau 5 figurent, pour les données sur la pollution de l'air, les modèles ordonnés en fonction de (9), les niveaux de signification (p-values) associés aux coefficients de régression, le R^2 et les valeurs de f_2, C_p et R_a^2. Nous avons choisi de représenter les cinq modèles les mieux classés et les cinq plus mal classés par le critère (9), parmi les 57 ayant $p \geqslant 2$ prédicteurs. Comme le programme MATLAB ordonne les modèles, l'idée est de parcourir leur liste de haut en bas, en s'arrêtant sur celui que pour une raison ou une autre (statistique ou extra-statistique) on juge satisfaisant. Dans le cadre d'une prévision par exemple, le choix du modèle définitif peut dépendre de l'existence ou non de la valeur future d'un prédicteur. Il est prudent de considérer quelques modèles du haut de la liste et non uniquement le premier. Pour les données de la pollution de l'air, les graphes des fonctions de choix associées à chacun des modèles apparaissent dans la figure 3 sous (*a*). Le modèle en tête de liste est 1245 pour $\alpha = 0.85$, et 12345 pour $\alpha = 0.95$. La statistique C_p est minimale sur 12345. Or, pour une valeur aussi élevée de α, il est vraisemblable que ce modèle présente de la

TABLEAU 5

Exemple de la pollution de l'air : niveaux de signification, R^2, f_2, C_p et R_a^2 des modèles ordonnés grâce au critère f_2. Les niveaux de signification sont ceux de la constante β_0 et des coefficients de régression $\beta_1, \ldots, \beta_6$.

$\alpha = 0.95$											
Modèle	Niveaux de signification β_0	β_1	β_2	β_3	β_4	β_5	β_6	R^2	f_2	C_p	R_a^2
cinq meilleurs modèles											
12345	0	.01	0	.01	.09	.06		.67	.45	5.1	.62
123456	.02	.05	0	.01	.09	.17	.75	.67	.45	7	.61
1245	0	0	0		.06	.03		.6	.46	9.86	.56
12346	.07	.15	0	.01	.16		.2	.65	.47	6.99	.60
1235	.01	.04	0	.01		.11		.64	.48	6.07	.60
cinq plus mauvais modèles											
1456	.06	.04			.52	.31	.98	.25	.9	45.6	.17
46	.65				.82		.02	.14	.91	53.7	.09
56	.84					.32	.01	.16	.91	51.6	.11
456	.83				.9	.34	.02	.16	.93	53.5	.09
45	.69				.56	.73		.01	1.02	66.6	-.04
$\alpha = 0.85$											
Modèle	Niveaux de signification β_0	β_1	β_2	β_3	β_4	β_5	β_6	R^2	f_2	C_p	R_a^2
cinq meilleurs modèles											
1245	0	0	0		.06	.03		.6	.62	9.86	.56
12	0	.01	0					.52	.63	14.8	.49
26	.37		0				.02	.5	.63	16.6	.47
125	0	0	0			.06		.56	.63	12.0	.53
124	0	0	0		.13			.55	.64	13.8	.51
cinq plus mauvais modèles											
56	.84					.32	.01	.16	1.06	51.6	.11
45	.69				.56	.73		.01	1.10	66.6	-.04
456	.83				.9	.34	.02	.16	1.14	53.5	.09
156	.06	.05				.37	.89	.25	1.27	45.6	.19
1456	.06	.04			.52	.31	.98	.25	1.33	45.6	.17

multicolinéarité. Nous en concluons que la méthode du C_p peut proposer des modèles instables[3]. L'utilisation de la couleur sur ce graphique, qu'on ne peut reproduire ici, permet incidemment de découvrir des structures intéressantes pour la comparaison des modèles. Pour pallier l'absence de couleur, nous avons imprimé sur la figure 3 (*b*) les fonctions de choix des modèles 12345 et 1245, afin de les comparer. Le modèle 1245 apparaît nettement préférable, car il assure un ajustement aux données équivalent à celui de 12345, tout en présentant un degré de multicolinéarité bien moindre.

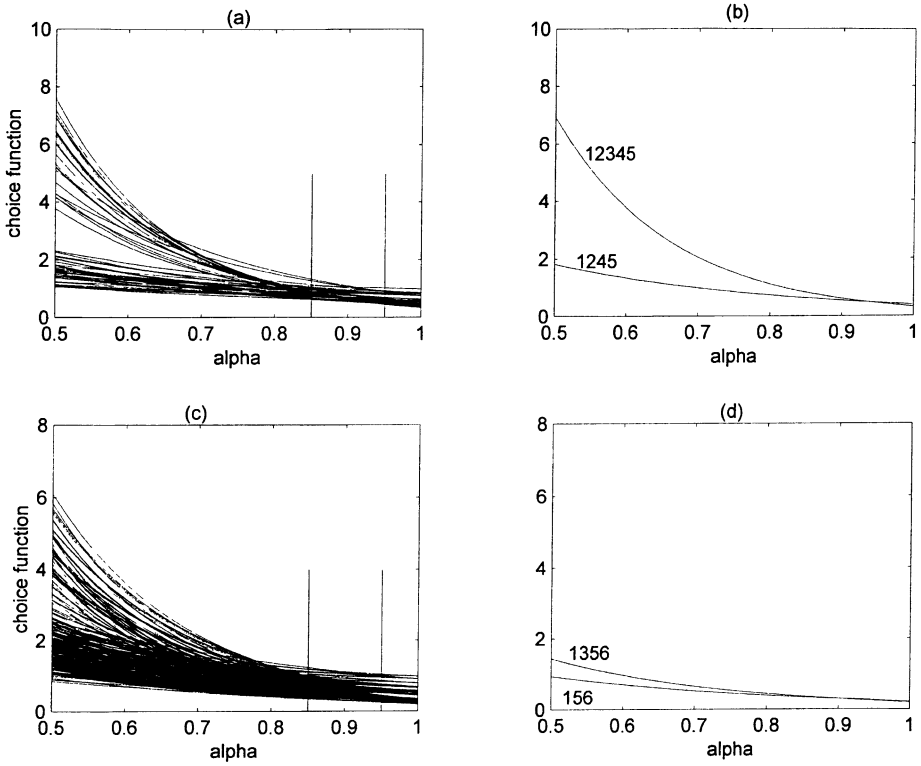

FIGURE 3

Graphes des fonctions de choix pour l'exemple de la pollution de l'air [(a) et (b)] et pour l'exemple des ventes [(c) et (d)]. Pour les données de la pollution de l'air, les modèles 12345 et 1245 sont les choix respectifs de f_2 pour $\alpha = 0.95$ et $\alpha = 0.85$; le second présente un degré de multicolinéarité moindre, tout en ayant un R^2 équivalent. Pour l'exemple des ventes, on observe en (d) qu'un modèle raisonnable est 156 : celui-ci est pratiquement aussi cohérent que le modèle voisin 1356 minimisant le C_p, mais plus stable, ainsi qu'en atteste le comportement de la fonction de choix sur la gauche du graphique. L'usage de la couleur sur l'écran de l'ordinateur, qui permet de faire ressortir un modèle donné de la masse des autres, augmente l'utilité du graphique.

[3] Montgommery et Peck (1992) relèvent que la méthode du C_p peut choisir des modèles présentant de forts facteurs d'inflation de variance et donc une forte multicolinéarité.

TABLEAU 6

Exemple des ventes : niveaux de signification, R^2, f_2, C_p et R_a^2 des modèles ordonnés grâce au critère f_2. Les niveaux de signification sont ceux de la constante β_0 et des coefficients de régression $\beta_1, \ldots, \beta_8$.

cinq meilleurs modèles pour $\alpha = 0.95$													
Modèle	Niveaux de signification β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	R^2	f_2	C_p	R_a^2
cinq meilleurs modèles													
1356	0	0		.05		.01	0			.80	.24	1.76	.78
12356	0	0	.58	.05		.01	0			.80	.24	3.48	.77
13568	0	0		.06		.01	0		.93	.80	.25	3.76	.77
123568	0	0	.59	.06		.01	0		.99	.80	.25	5.48	.76
13456	0	0		.13	.68	.01	0			.80	.25	3.6	.77
cinq meilleurs modèles pour $\alpha = 0.85$													
Modèle	Niveaux de signification β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	R^2	f_2	C_p	R_a^2
cinq meilleurs modèles													
156	0	0				0	0			.78	.35	3.44	.76
1456	0	0			.21	0	0			.79	.35	3.89	.76
12456	0	0	.69		.22	0	0			.79	.36	5.72	.75
1256	0	0	.72			0	0			.78	.36	5.31	.75
1356	0	0		.05		.01	0			.80	.36	1.76	.78

La même procédure a été appliquée à un jeu de données de 38 observations fourni par Weelwright et Makridakis (1974). Il s'agit de relier les ventes semestrielles d'un produit à 8 prédicteurs potentiels.

Dans le tableau 6 apparaissent, pour des valeurs de α de 0.85 et 0.95, les 5 modèles les mieux classés, parmi les 247 ayant $p \geqslant 2$ prédicteurs. Dans notre exemple, le modèle 156 (1356) est en tête de liste pour $\alpha = 0.85$ (0.95). Sur la figure 3 (*c*) apparaissent les graphes des 247 fonctions de choix de l'exemple. Le C_p prend sa valeur minimale (1.76) sur le modèle 1356. Le modèle 156 apparaît légèrement préférable, car il assure un ajustement aux données équivalent à celui de 1356, tout en présentant un degré de multicolinéarité un peu moindre.

Choix de α

Lors de la détermination du niveau de α, la règle est de choisir α de manière que les modèles retenus en tête de liste soient impérativement cohérents (leurs fonctions de choix ont de petites valeurs sur la droite du graphique) et présentent de bonnes

qualités de stabilité (leurs fonctions de choix n'ont pas de fortes valeurs sur la gauche du graphique). L'observation des graphes des fonctions de choix permet de trouver les valeurs de α réalisant ce double objectif. Pour ce qui est des deux exemples traités ci-dessus, la figure 3 indique que des valeurs de α de l'ordre de 0.7 à 0.8 permettent le choix de modèles à la fois stables et cohérents. Une approche sensée est de considérer plusieurs valeurs de α, par exemple 0.6, 0.7 et 0.8. On regarde à chaque fois les cinq meilleurs modèles et on choisit parmi eux celui ayant de bonnes caractéristiques statistiques (un bon R^2 et de bons niveaux de signification) ou extra-statistiques (information a priori, existence de données, etc.). Le choix d'un α trop petit (resp. grand) se traduit par la présence de modèles de faible R^2 (resp. de prédicteurs non significatifs) en tête de liste. Sur cette base, il se peut qu'on soit amené à changer la valeur de α.

Le choix définitif de α et donc du degré d'instabilité qu'on admet est aussi lié au but de notre analyse de régression. Si un modèle est destiné au contrôle ou à une extrapolation dans un cadre prédictif, toute multicolinéarité est à bannir et un α plutôt petit donnant la préférence à des modèles stables sera indiqué. Si en revanche le modèle est utilisé à des fins d'interpolation, une combinaison linéaire $\sum \beta_j x_{ij}$ peut être estimée correctement sur un modèle instable, en dépit du fait que les β_j puissent être pauvrement estimés. Dans une telle occurrence, une dose relativement importante d'instabilité et l'usage d'un α plus élevé sont admissibles.

5. Conclusion

Nous avons présenté dans cet article un critère empirique de nature géométrique permettant d'ordonner des modèles de régression en fonction du degré de conjonction de deux nuages de données ($\mathcal{N}_X$ et $\mathcal{N}_Z$). L'éloignement à la conjonction parfaite a été mesuré par l'inertie relative des nuages par rapport à un hyperplan optimal. Or l'inertie dépend du choix d'une norme et d'un type de projection. Il est tout à fait possible que l'usage d'autres normes et projections que celles utilisées ici pour mesurer la cohérence ou la stabilité d'un modèle amènent à des résultats encore supérieurs. Egalement, d'autres façons efficientes de combiner dans un critère cohérence et conjonction sont certainement possibles. La méthode f_2 a été testée sur de nombreux jeux de données réelles, ce qui a permis d'établir des règles pragmatiques pour son utilisation. Des simulations ont montré qu'elle est capable de trouver les vrais modèles et qu'elle se compare très favorablement au C_p de Mallows et à d'autres méthodes.

Références

P. BERTIER et J.-M. BOUROCHE. (1981), *Analyse des données multidimensionnelles*, PUF, coll. Systèmes-Décisions, Paris.

B. FLURY. (1988), *Common Principal Components and Related Multivariate Models*, New York : John Wiley & sons.

D.M. HAWKINS. (1973), «On the Investigation of Alternative Regressions by Principal Components Analysis», *Applied Statistics*, 22, 275–86.

J.R.MAGNUS and H. NEUDECKER. (1988), *Matrix Differential Calculus with Applications in Statistics and Econometrics*, New York : John Wiley & sons.

D.C. MONTGOMERY and E.A.PECK. (1992), *Introduction to Linear Regression Analysis*. John Wiley, New York.

E.RONCHETTI and R.G. STAUDTE. (1994), «A Robust Version of Mallows C_p»,*Journal of the American Statistical Association*, 89, 550–559.

E. RONCHETTI, C. FIELD and W.BLANCHARD. (1997), «Robust Linear Model Selection by Cross-Validation», *Journal of the American Statistical Association*, 92, 1017–1023.

J.SHAO. (1993), «Linear Model Selection by Cross-Validation», *Journal of the American Statistical Association*, 88, 486-494.

R.R. SOKAL and F.J. ROHLF. (1981), *Biometry*, Ed. W.H. Freeman, San Francisco.

S.C.WEELWRIGHT et S. MAKRIDAKIS. (1974), *Choix et Valeur des Méthodes de Prévision*, Ed. d'organisation, Paris.