Une méthode de sélection de variables adaptée à l'utilisation d'un modèle de régression en prévision

Elsa Freville

Pub. Inst. Stat. Univ. Paris
XXXXIV, fasc. 1, 2000, 47 à 63

Une méthode de sélection de variables adaptée à l'utilisation d'un modèle de régression en prévision

Elsa FREVILLE
INRETS-GRETIA, 2 av du Général Malleret-Joinville, 94114 Arcueil Cedex

Résumé : Une nouvelle méthode de sélection des variables de régression d'un problème de prévision est proposée. Celle-ci utilise un critère empirique qui est une transposition de celui de Mallows à la prévision. Ce dernier critère est pris comme référence pour les procédures de sélection de variables classiques, lorsque le modèle réel dépend d'une infinité de paramètres. Sous cette hypothèse, nous comparons notre critère avec celui de Mallows et nous utilisons l'analyse de la variance pour une application de la méthode à un problème de prévision du trafic routier.

AMS Classification : 62J05, 62J10.
Mots clefs : sélection de variables, prévision, risque quadratique minimum, régression multiple, analyse de la variance.

1. Introduction

Soit Y une variable aléatoire dépendante d'un vecteur x de paramètres explicatifs selon la fonction de régression,

$$Y = x'\beta + e, \qquad e \approx \aleph_1\left(0, \sigma^2\right)$$

x et β sont de dimension finie en régression classique, mais nous pouvons étendre cette notion au cas infini avec des vecteurs $x = (x_1, x_2, \ldots)'$ et $\beta = (b_1, b_2, \ldots)'$. $x'\beta$ désigne le produit scalaire des vecteurs x et β dans l_2, l'espace de Hilbert des suites de nombres réels de carrés intégrables (Shibata, 81). Dans le cadre empirique, ce modèle devient pour un nombre fini n d'observations indépendantes,

$$y = X\beta + \varepsilon, \qquad \varepsilon \approx \aleph_n\left(0, \sigma^2 I_n\right)$$

où $y = (y_1, y_2, \ldots, y_n)'$ désigne le vecteur des observations de Y, X la matrice $n \times \infty$ des n observations des paramètres explicatifs et $\varepsilon = (e_1, e_2, \ldots, e_n)'$ le vecteur des n résidus, supposé distribué selon une loi normale de dimension n, de moyenne nulle et de variance $\sigma^2 I_n$, avec I_n la matrice identité de dimension n.

En fait, dans ce cadre empirique, le concept du modèle de taille infinie est peu réaliste : d'une part car souvent seul un nombre $s_n(K)$ de paramètres

explicatifs est connu de façon précise, et d'autre part car, au plus, n-1 coefficients de régression peuvent être estimés par des techniques basées sur le principe des moindres carrés, ou du maximum de vraisemblance. De plus, lorsque le nombre de paramètres explicatifs augmente, il en est de même pour l'instabilité des estimations et des prévisions. Dans le contexte d'un modèle de taille infinie, le principe de **sélection de variables** prend alors toute sa signification, dans la mesure où il permet de limiter le nombre des paramètres explicatifs, pour améliorer la stabilité tout en conservant une bonne qualité d'ajustement.

Les procédures de sélection de variables recherchent alors un « modèle k », utilisant $s_n(k)$ variables explicatives parmi les $s_n(K)$ connues, de la forme

$$y = X_k\beta_k + \varepsilon_k \quad , \qquad \varepsilon_k \approx \aleph_n\left(m_k, \sigma^2 I_n\right)$$

avec $X_k = \left(x_{k(1)}, x_{k(2)}, \ldots, x_{k(s_n(k))}\right)$, la matrice des observations des k variables explicatives et $\beta_k = \left(b_{k(1)}, b_{k(2)}, \ldots, b_{k(s_n(k))}\right)'$ tel que $b_{k(m)} \neq 0$, pour $1 \leq m \leq s_n(k)$. ε_k est le vecteur des n réalisations d'une variable aléatoire de loi normale de variance σ^2, et de moyenne $m_k = \sum_{p \neq k(1),\ldots,k(s_n(k))} x_p b_p$, non nulle à cause de l'introduction, par la procédure de sélection, d'un biais pour l'estimation de y. Pour tout « modèle k », on notera $r_n(k) = \text{rang}(X_k' X_k)$ et $\hat{y}_k = X_k\hat{\beta}_k$ le vecteur des estimations des observations réelles y, où $\hat{\beta}_k$ désigne les estimateurs des moindres carrés pour β_k soit, si l'on suppose $r_n(k) = s_n(k)$, $\hat{\beta}_k = (X_k' X_k)^{-1} X_k' y$.

Ces notations sont également généralisées au « modèle K » qui utilise l'ensemble des $s_n(K)$ paramètres connus. Pour ce modèle, on suppose également la moyenne $m_K = \sum_{p \neq K(1),\ldots K(s_n(K))}^{\infty} x_p b_p \approx 0$, pour $s_n(K)$ suffisamment grand et les variables bien choisies.

Les techniques permettant de choisir le « modèle k » sont multiples et une présentation des plus courantes sera proposée dans la section 2. Toutes celles-ci utilisent une étude des estimations des n observations de Y et nous nous concentrerons sur celles basées sur un critère empirique de type fonction pénalisée. On retiendra comme critère théorique pour choisir le modèle optimal, appelé κ, celui du **risque quadratique** $R_n(k)$ **minimum** (Shibata 81) soit, $R_n(\kappa) = \underset{k}{\text{Min}}\, R_n(k) = \underset{k}{\text{Min}}\, E\left(\left\| \hat{y}_k - E(y)\right\|^2\right)$. L'utilisation des différents critères empiriques de sélection s'interprétera alors comme une procédure d'estimation de $R_n(k)$. L'analyse de ces critères permettra d'en montrer la

similitude et de retenir un critère équivalent à celui des Cp de Mallows (73) qui sera noté $\hat{R}_n(k)$. Pour estimer κ, nous utiliserons le modèle $\hat{\kappa}$, vérifiant la propriété $\hat{R}_n(\hat{\kappa}) = \underset{k}{\text{Min}}\, \hat{R}_n(k)$.

Mais l'originalité de cet article, résidera dans la présentation, dans la section 3, d'un critère de sélection du modèle k basé, non plus sur l'étude de la qualité des estimations, mais sur celle d'un nombre n^* de prévisions $\hat{y}_k^* = X_k^* \hat{\beta}_k$ de la variable Y. En effet, une telle démarche présente deux avantages : le premier est d'utiliser un critère calculé sur des données réduites, d'où des résultats plus rapides, le second est de prendre en compte la stabilité propre des prévisions qui peut être différente de celle, globale, des estimations de l'ensemble de l'historique. Comme en estimation, nous retiendrons la notion de **risque quadratique** $R_n^*(k)$ **minimal** comme critère de choix du modèle κ^* optimal avec, $R_n^*(\kappa^*) = \underset{k}{\text{Min}}\, R_n^*(k) = \underset{k}{\text{Min}}\, E\left[\left\| \hat{y}_k^* - E(y^*) \right\|^2 \right]$ et nous en déduirons une estimation $\hat{\kappa}^*$ de κ^* à l'aide d'un critère empirique $\hat{R}_n^*(k)$, et en résolvant $\hat{R}_n^*(\hat{\kappa}^*) = \underset{k}{\text{Min}}\, \hat{R}_n^*(k)$.

Les **sections 3.1** et **3.2** compareront respectivement les critères théoriques et empiriques pour l'estimation et la prévision, quant aux **sections 4 et 5**, ils présenteront une application des procédures de sélection au cas particulier des modèles d'analyse de la variance, dans le cadre théorique général (section 4) puis dans celui d'un modèle réel de prévision du trafic routier journalier (section 5).

2. Présentation des critères de sélection classiques

Il existe une littérature importante sur le problème de la sélection de variables, et ce pour différentes approches théoriques : basées sur des considérations asymptotiques, bayésiennes, issues de la théorie de l'information, du principe du maximum de vraisemblance ou des moindres carrés. Une bonne partie de ces travaux aboutit à la construction de critères empiriques ayant la structure d'une fonction d'ajustement pénalisée par la dimension du modèle qui se décompose donc en une somme de deux termes positifs:

- un premier caractéristique de l'**ajustement**, décroissant avec l'augmentation des paramètres explicatifs. Il s'agit soit du carré de la norme des résidus de l'estimation $\|y - \hat{y}_k\|^2$ (cas des critères FPE, FPE_α, RIC, ...) ,soit d'un terme dérivé du logarithme de la vraisemblance maximale $n \ln\left(n^{-1} \|y - \hat{y}_k\|^2 \right)$ comme c'est le cas pour les critères AIC, BICet ϕ,

- un second qui est une fonction de **pénalité** croissante avec la complexité du modèle. Celle-ci est proportionnelle à $r_n(k)$ le rang de la matrice des variables explicatives, soit sous la forme $\alpha\, r_n(k)$, soit sous celle $\alpha\, r_n(k)\hat{\sigma}^2$ qui est une estimation de la variance des estimations $\hat{y}_k$, où $\hat{\sigma}^2$ est un estimateur de la variance σ^2 des résidus du modèle.

En minimisant de tels critères, les procédures de sélection de modèle peuvent ainsi être interprétées comme la recherche d'un compromis entre ajustement et stabilité des estimations ce qui permet d'obtenir une solution au problème posé.

Parmi ces critères, on peut citer :

$\left\|y - X_k\hat{\beta}_k\right\|^2 + 2\, r_n(k)\ \hat{\sigma}_k{}^2$ critère FPE, Akaike (1970),

$\left\|y - X_k\hat{\beta}_k\right\|^2 + 2 \log K\ r_n(k)\ \hat{\sigma}_K{}^2$ critère RIC, Foster et George (1994),

$\left\|y - X_k\hat{\beta}_k\right\|^2 + \alpha\ r_n(k)\ \hat{\sigma}_K{}^2$ α>0 constant ou fonction croissante de n

critère FPE_α, Shibata (1984, 1986), Barron, Birgé, Massart (1999)

$n \log\left(n^{-1} \left\|y - X_k\hat{\beta}_k\right\|^2 \right) + 2\, r_n(k)$ critère AIC, Akaike (1973),

$n \log\left(n^{-1} \left\|y - X_k\hat{\beta}_k\right\|^2 \right) + \log(n)\ r_n(k)$ critère BIC, Schwarz (1978),

$n \log\left(n^{-1} \left\|y - X_k\hat{\beta}_k\right\|^2 \right) + c \log\log(n)\ r_n(k)$ c>2 constante réelle

critère Φ, Hannan et Quinn (1979)

$\|-\|$ désigne la norme euclidienne usuelle, $\hat{\sigma}_k{}^2$ et $\hat{\sigma}_K{}^2$ sont des estimateurs de la variance σ^2 obtenus, le premier avec le « modèle k », et le second avec celui utilisant la totalité des $s_n(K)$ variables explicatives où $\hat{\sigma}_k{}^2 = \|y - \hat{y}_k\|^2 / n$ ou $\|y - \hat{y}_k\|^2 / (n - r_n(k))$.

Remarquons que l'on a, pour n assez grand et α une constante positive,

$$n \ln\left(n^{-1} \left\|y - X_k\hat{\beta}_k\right\|^2 \right) + \alpha\ r_n(k) = n \ln\left(\sigma^2\right) + n \ln\left[1 + \left(\frac{\left\|y - \hat{y}_k\right\|^2}{n\ \sigma^2} - 1 \right) \right] + \alpha\, r_n(k)$$

$$= \frac{\|y - \hat{y}_k\|^2}{\sigma^2} + \alpha\, r_n(k) + n\left(\ln \sigma^2 - 1\right) + n\, o\left[\left(\frac{\|y - \hat{y}_k\|^2}{n\,\sigma^2} - 1\right)^2\right]$$

Les procédures minimisant des critères du type,

$$n \log\left(n^{-1} \left\|y - X_k \hat{\beta}_k\right\|^2\right) + \alpha\ r_n(k)$$

sont donc asymptotiquement équivalentes à celles utilisant les critères,

$$\left\|y - X_k \hat{\beta}_k\right\|^2 + \alpha\ r_n(k)\ \hat{\sigma}^2$$

et ce quel que soit l'estimateur $\hat{\sigma}^2$ de la variance σ^2 des résidus utilisé. Tous les critères présentés sont donc asymptotiquement équivalents à celui étudié par Shibata (1984, 1986) appelé FPE_α. Le problème consiste alors à choisir α pour obtenir le modèle optimal, ce qui dépend des hypothèses sur la structure du modèle réel.

Comme Shibata et d'autres, nous retiendrons le critère théorique du **risque quadratique** $R_n(k)$ **minimum** pour choisir le modèle κ optimal (Shibata 81, 84, 86, Barron, Birgé et Massart, 99) avec,

$$R_n(k) = E\left(\left\|X\beta - X_k \hat{\beta}_k\right\|^2\right) = \left\|X\beta - X_k \beta_k\right\|^2 + r_n(k)\,\sigma^2 \quad (2.1)$$

où $\beta_k = E\left(\hat{\beta}_k\right)$ et $R_n(\kappa) = \underset{k}{\text{Min}}\ R_n(k)$.

Grâce à la décomposition (2.1), on peut constater que cette notion est bien adaptée à la dualité du problème : $\left\|X\beta - X_k \beta_k\right\|^2$ est le **carré du biais** de l'estimation de y par le modèle k, quant à $r_n(k)\,\sigma^2$, il s'agit de la **variance** de l'estimation de y par le même modèle.

Comme κ ne peut être connu de façon exacte en raison de la présence de paramètres inconnus dans $R_n(k)$, Shibata (1981) introduit ainsi le critère $S_n(k)$ permettant d'estimer $R_n(k)$ et d'obtenir le modèle $\tilde{\kappa}$ vérifiant,

$$S_n(\tilde{\kappa}) = \underset{k}{\text{Min}}\ S_n(k)$$

avec,

$$S_n(k) = \left\|y - X_k \hat{\beta}_k\right\|^2 \left(1 + \frac{2 r_n(k)}{n}\right) = \left\|y - X_k \hat{\beta}_k\right\|^2 + 2\, r_n(k)\, \hat{\sigma}_k{}^2 \quad (2.2)$$

Shibata (81), sous certaines hypothèses reliées à la taille infinie du modèle réel, montre la convergence en probabilité de $\tilde{\kappa}$ vers κ.

Mais, l'hypothèse du modèle réel de taille infinie introduit un biais dans l'estimation de y par $X_k \hat{\beta}_k$,

$$m_k = \sum_{p \neq k(1),\ldots,k(s_n(k))} x_p b_p \neq 0$$

Nous préférons donc $\hat{\sigma}_K{}^2 = \|y - \hat{y}_K\|^2 / n$ à $\hat{\sigma}_k{}^2$ pour estimer σ^2 ce qui aboutit à utiliser le critère,

$$\hat{R}_n(k) = \left\|y - X_k \hat{\beta}_k\right\|^2 + 2\ r_n(k)\ \hat{\sigma}^2 K \qquad (2.3)$$

Ceci nous amène alors à retenir le modèle empirique $\hat{\kappa}$ vérifiant

$$\hat{R}_n(\hat{\kappa}) = \min_k \hat{R}_n(k)$$

Pour valider ce choix, étendons le résultat de Shibata (81) à notre critère par le théorème suivant,

Théorème :
Sous les hypothèses du théorème de Shibata,

i) $r_n(K) = o(n)$

ii) Pour tout $0 < \delta < 1$, $\lim_{n \to +\infty} \sum_k \delta^{R_n(k)} = 0 \quad (p)$

alors,

$$\lim_{n \to +\infty} \frac{R_n(\hat{\kappa})}{R_n(\kappa)} = 1 \quad (p)$$

La condition ii) d'application de ce théorème est liée à l'hypothèse de taille infinie du modèle réel qui entraîne $\lim_{n \to +\infty} R_n(k) = +\infty$. Quant à l'hypothèse i), elle permet également d'assurer une bonne qualité des estimations et des prévisions car les paramètres à estimer restent ainsi en nombre suffisamment restreint par rapport à la taille de l'historique.

Démonstration :
Nous n'indiquons que le changement d'ailleurs mineur par rapport à la démonstration de Shibata.

Soit $S^2(k) = \dfrac{\|y - X_k \beta_k\|^2}{n}$

Le critère empirique peut alors se décomposer sous la forme,

$$\hat{R}_n(k) = R_n(k) + \left[\sigma^2 r_n(k) - \left\| X_k \hat{\beta}_k - X_k \beta_k \right\|^2\right] + 2 r_n(k)\left[\hat{\sigma}_K{}^2 - \sigma^2\right]$$
$$+ n\left[S^2(k) - E\left(S^2(k)\right)\right] + n\sigma^2$$

d'où $\hat{R}_n(k) = R_n(k) + S_{1n}(k) + S_{2n}(k) + S_{3n}(k) + n\sigma^2$

Sous les hypothèses (i) et (ii) Shibata montre uniformément en k,

$$\lim_{n->\infty} S_{1n}(k) / R_n(k) = 0 \ (p) \text{ et } \lim_{n->\infty} [S_{3n}(k) - S_{3n}(\kappa)] / R_n(k) = 0 \ (p)$$

ainsi que $\lim_{n->\infty} 2 r_n(k) \left[\hat{\sigma}_k{}^2 - \sigma^2 \right] / R_n(k) = 0 \ (p)$.

En suivant pas à pas la démonstration de Shibata, cette dernière convergence peut facilement être étendue à $S_{2n}(k)$, ce qui donne uniformément en k, $\lim_{n->\infty} S_{2n}(k) / R_n(k) = 0 \ (p)$.

En rassemblant les différents résultats de convergence, nous obtenons, comme Shibata, la convergence uniforme en k,

$$\lim_{n->\infty} \left[\hat{R}_n(k) - \hat{R}_n(\kappa) \right] / R_n(k) = 1 - \lim_{n->\infty} R_n(\kappa) / R_n(k) \quad (p)$$

ce qui, écrit pour le modèle $\hat{\kappa}$, donne $\lim_{n\to+\infty} R_n(\hat{\kappa}) / R_n(\kappa) \leq 1 \quad (p)$

Comme $R_n(\hat{\kappa}) / R_n(\kappa) \geq 1$ pour tout n, le théorème est alors démontré.

Sous les hypothèses du modèle réel de taille infinie, $\hat{\kappa}$ converge donc vers le modèle optimal κ et l'on peut penser à étendre cette convergence vers κ aux modèles obtenus grâce aux critères AIC, FPE, FPE_2.

A noter, $\hat{R}_n(k)$ est équivalent au critère des Cp de Mallows (73) qui s'écrit, $C_p = \left\| y - X_k \hat{\beta}_k \right\|^2 / \hat{\sigma}_K{}^2 - (n - 2 r_n(k))$ et donne le même modèle $\hat{\kappa}$ lorsqu'il est minimisé pour un nombre n d'observations données.

Remarquons que les conclusions ne sont pas les mêmes si l'on suppose le **modèle réel de taille finie** ou de **structure différente**. Pour un modèle de taille finie **inconnu**, Zheng et Loh (95) et Rao et Wu (89) recommandent l'utilisation d'un α divergent fonction de la taille n de l'échantillon considéré et montrent la convergence en probabilité des modèles $\hat{k}_\alpha$ ainsi obtenus vers le modèle réel. C'est au tour des critères BIC, RIC et ϕ d'être optimaux. Une étude plus générale sur le choix du coefficient de pénalité α vient d'être récemment publiée par Barron, Birgé et Massart (99).

3. Un nouveau critère de sélection tenant compte de l'utilisation pour la prévision

Le critère présenté ici est une adaptation à l'utilisation en prévision du critère de Mallows retenu précédemment comme critère classique de référence. De même que le critère de risque quadratique minimum pour l'estimation a été retenu

comme critère d'optimalité pour le choix du modèle de régression, il en sera de même pour le critère en prévision.

Notons $\hat{y}_k^* = X_k^* \hat{\beta}_k$ les estimations des prévisions $y^* = X^* \beta$ des n^* valeurs futures de Y obtenues par le « modèle k », et $R_n^*(k)$ le risque quadratique de celles-ci. Le nouveau modèle optimal κ^* vérifie alors,

$$R_n^*(\kappa^*) = \min_k R_n^*(k) = \min_k E\left[\left\| \hat{y}_k^* - E(y^*) \right\|^2 \right] \qquad (3.1)$$

où $y^* = X^* \beta$ est donné par le modèle réel de taille infinie et, pour l'historique $y = X\beta + \varepsilon$, avec $\varepsilon \approx \aleph_n(0, \sigma^2 I_n)$.

Nous pouvons écrire,

$$R_n^*(k) = E\left(\left\| X^* \beta - X_k^* \hat{\beta}_k \right\|^2 \right) = E\left(\left\| X^* \beta - X_k^* \beta_k - X_k^* (X_k' X_k)^{-1} X_k' \varepsilon \right\|^2 \right) \qquad (3.2)$$

$$= \left\| X^* \beta - X_k^* \beta_k \right\|^2 + \sigma^2 \operatorname{trace}\left(X_k^* (X_k' X_k)^{-1} X_k'^* \right)$$

où $\beta_k = E(\hat{\beta}_k)$.

A présent, comme $R_n^*(k)$, de même que $R_n(k)$, ne peut être déterminé de façon exacte, nous sommes obligés de passer par une étape d'estimation. Pour cela, nous proposons le critère $\hat{R}_n^*(k)$ qui, minimisé, permet d'obtenir le modèle $\hat{\kappa}^*$ vérifiant $\hat{R}_n^*(\hat{\kappa}^*) = \min_k \hat{R}_n^*(k)$ avec,

$$\hat{R}_n^*(k) = \left\| X_K^* \hat{\beta}_K - X_k^* \hat{\beta}_k \right\|^2 + 2\hat{\sigma}^2 K \operatorname{trace}\left(X_k^* (X_k' X_k)^{-1} X_k'^* \right) \qquad (3.3)$$

où $\hat{\sigma}_K{}^2 = \left\| y - X_K \hat{\beta}_K \right\|^2 / n$ et où $X_K^* \hat{\beta}_K$ désigne la prévision obtenue en utilisant la totalité des $s_n(K)$ paramètres explicatifs connus.

Pour justifier de l'intérêt et de l'originalité de ce nouveau critère, nous nous attacherons dans la suite à comparer les modèles théoriques κ et κ^* puis les modèles empiriques $\hat{\kappa}$ et $\hat{\kappa}^*$.

3.1 Comparaison des critères théoriques

On se place sous l'hypothèse où X, la matrice des paramètres explicatifs du modèle de taille infinie a pour vecteurs colonnes des x_p, $p \geq 1$ orthogonaux deux à deux. Pour simplifier, supposons également X_k de plein rang avec $s_n(k)=k$ et de

vecteurs colonnes $x_p, 1 \le p \le k$. Pour les matrices des variables explicatives des prévisions, les $x_p^*, p \ge 1$ désignent les vecteurs colonnes de X^* et les x_p^*, $1 \le p \le k$, ceux de X_k^*.

Du fait de l'orthogonalité des x_p et de la taille infinie du modèle réel, on a

$$\|X\beta - X_k\beta_k\|^2 = \left\| \sum_{p=k+1}^{\infty} x_p\beta_p \right\|^2 = \sum_{p=k+1}^{\infty} \|x_p\|^2 \beta_p{}^2$$

de même,

$$\|X^*\beta - X_k^*\beta_k\|^2 = \left\| \sum_{p=k+1}^{\infty} x_p^*\beta_p \right\|^2 = \sum_{p=k+1}^{\infty} \|x_p^*\|^2 \beta_p{}^2 + \sum_{p=k+1}^{\infty} \sum_{\substack{q=k+1 \\ p\neq q}}^{\infty} \langle x_p^*, x_q^* \rangle \beta_p\beta_q$$

où $\langle x_p^*, x_q^* \rangle$ désigne le produit scalaire des vecteurs x_p^* et x_q^*, qui est a priori non nul pour $p \neq q$.

On peut également donner une expression explicite de $\text{trace}\left(X_k^*(X_k'X_k)^{-1}X_k'^*\right)$ grâce à l'orthogonalité des x_p car $(X_k'X_k)^{-1}$ est une matrice diagonale.

$$(X_k'X_k)^{-1} = \begin{pmatrix} 1/\|x_1\|^2 & & 0 \\ & \ddots & \\ 0 & & 1/\|x_k\|^2 \end{pmatrix} \text{ et } \text{trace}\left(X_k^*(X_k'X_k)^{-1}X_k'^*\right) = \sum_{p=1}^{k} \frac{\|x_p^*\|^2}{\|x_p\|^2}$$

Ceci permet de réécrire les expressions des risques théoriques de l'estimation et de la prévision sous la forme,

$$\begin{cases} R_n(k) = \displaystyle\sum_{p=k+1}^{\infty} \beta_p{}^2 \|x_p\|^2 + \sigma^2 \sum_{p=1}^{k} \frac{\|x_p\|^2}{\|x_p\|^2} \\ R_n^*(k) = \displaystyle\sum_{p=k+1}^{\infty} \beta_p{}^2 \|x_p^*\|^2 + \sum_{p=k+1}^{\infty} \sum_{\substack{q=k+1 \\ q\neq p}}^{\infty} \beta_p\beta_q \langle x_p^*, x_q^* \rangle + \sigma^2 \sum_{p=1}^{k} \frac{\|x_p^*\|^2}{\|x_p\|^2} \end{cases}$$

Même en supposant négligeable l'effet de la somme double présente dans $R_n^*(k)$, les termes des sommes sont pondérés par des coefficients différents. Les

deux critères théoriques $R_n(k)$ et $R_n^*(k)$ sont différents et, en général, ne donnent donc pas les mêmes modèles optimaux κ et κ^*.

3.2 Comparaison des critères empiriques.

En conservant les mêmes hypothèses sur les matrices X_k et X_k^*, nous pouvons également décomposer les critères empiriques $\hat{R}_n(k)$ et $\hat{R}_n^*(k)$. Pour commencer, remarquons que $\left(y - X_K\hat{\beta}_K\right)$ et $\left(X_K\hat{\beta}_K - X_k\hat{\beta}_k\right)$ sont orthogonaux car $X_K\hat{\beta}_K$ et $X_k\hat{\beta}_k$ sont respectivement les projections orthogonales de y sur les espaces vectoriels engendrés par X_K et X_k, noté $\mathrm{Vect}(X_K)$ et $\mathrm{Vect}(X_k)$, avec $\mathrm{Vect}(X_k) \subset \mathrm{Vect}(X_K)$.

En utilisant le théorème de Pythagore, $\hat{R}_n(k)$ se réécrit sous la forme,

$$\hat{R}_n(k) = \left\|y - X_K\hat{\beta}_K\right\|^2 + \left\|X_K\hat{\beta}_K - X_k\hat{\beta}_k\right\|^2 + 2\,\hat{\sigma}_K{}^2\,k$$

Notons pour tout p, $1 \le p \le K$, $\quad A_p = x_p\, y$

Grâce à un calcul similaire à celui effectué pour le cas théorique nous obtenons,

$$\left\|X_K\hat{\beta}_K - X_k\hat{\beta}_k\right\|^2 = \left\| \sum_{p=k+1}^{K} \left(\frac{x_p\, x_p'}{\|x_p\|^2} \right) y \right\|^2$$

$$= \sum_{p=k+1}^{K} \frac{A_p{}^2}{\|x_p\|^2} \frac{\|x_p\|^2}{\|x_p\|^2} + \sum_{p=k+1}^{K} \sum_{\substack{q=k+1 \\ p \neq q}}^{K} \frac{A_p A_q}{\|x_p\|^2 \|x_q\|^2} \langle x_p, x_q \rangle$$

$$= \sum_{p=k+1}^{K} \frac{A_p{}^2}{\|x_p\|^2} \frac{\|x_p\|^2}{\|x_p\|^2}$$

ce qui donne une nouvelle expression de $\hat{R}_n(k)$,

$$\hat{R}_n(k) = \left\|y - X_K\hat{\beta}_K\right\|^2 + \sum_{p=k+1}^{K} \frac{A_p{}^2}{\|x_p\|^2} \frac{\|x_p\|^2}{\|x_p\|^2} + \sum_{p=1}^{k} 2\,\hat{\sigma}_K{}^2 \frac{\|x_p\|^2}{\|x_p\|^2}$$

De la même façon, $\hat{R}_n^*(k)$ s'écrit,

$$\hat{R}_n^*(k) = \sum_{p=k+1}^{K} \frac{A_p^2}{\|x_p\|^2} \frac{\|x_p^*\|^2}{\|x_p\|^2} + \sum_{p=k+1}^{K} \sum_{\substack{q=k+1 \\ q \neq p}}^{K} \frac{A_p A_q}{\|x_p\|^2 \|x_q\|^2} \langle x_p^*, x_q^* \rangle$$

$$+ \sum_{p=1}^{k} 2\hat{\sigma}_K^2 \frac{\|x_p^*\|^2}{\|x_p\|^2}$$

Les critères empiriques se comparent donc de la même façon que ceux théoriques : la structure du critère en prévision est modifiée par une pondération différente des coefficients des sommes, uniquement fonction de la structure de X_k^*. On peut donc en déduire que les modèles $\hat{\kappa}$ et $\hat{\kappa}^*$ sont différents, sauf pour des structures de X_k et de X_k^* suffisamment voisines : par exemple si pour tout $1 \leq p \leq K$, $\|x_p\|^2 \approx \mu \|x_p^*\|^2$, où μ est une constante et si les $x_p^*, 1 \leq p \leq k$ sont orthogonaux deux à deux. Remarquons que pour certaines variables explicatives, $\|x_p^*\|^2$ peut être nulle, ce qui rend inutile l'utilisation de ces variables dans le modèle de régression pour obtenir des prévisions.

4. Application au cas du modèle d'analyse de la variance

On appelle ainsi tout modèle de régression dont les variables explicatives sont des facteurs qualitatifs. Ceux-ci sont codés sous forme disjonctive dans la matrice des variables explicatives. Pour un modèle à k variables qualitatives, cette matrice, appelée X_k, s'écrit sous la forme,

$$X_k = \left(\begin{array}{c|c|c|c|c} 1 \\ \vdots \\ 1 & \chi_1 & \chi_2 & \cdots & \chi_k \end{array} \right)$$

où les χ_p, $1 \leq p \leq k$ sont des matrices blocs correspondant aux tableaux disjonctifs $n \times m_p$ associés aux n observations de chacune des k variables qualitatives, m_p désignant le nombre de modalités ($m_p \geq 2$) de la $p^{ème}$ variable explicative. Chaque variable explicative qualitative p, à m_p modalités, est transformée en m_p variables quantitatives ayant chacune pour valeurs soit 1, si la modalité correspondante est vraie, soit 0 dans le cas contraire.

Le modèle prend alors la forme classique pour les observations,

$$y = X_k \beta_k + \varepsilon_k , \qquad \varepsilon_k \approx N_n\left(0, \sigma^2 I_n\right)$$

$s_n(k) = \sum_{p=1}^{k} m_p + 1$, et on notera X_k^* la matrice des variables explicatives pour les données à prédire.

Les particularités d'un tel modèle sont les suivantes:

- X_k n'est pas de plein rang et il n'y a unicité, ni des estimateurs $\hat{\beta}_k$, ni des prévisions, sauf si $\mathrm{Ker}(X_k) \subset \mathrm{Ker}\left(X_k^*\right)$, ce qui n'est pas le cas si la première réalisation d'une modalité qualitative a lieu pour une donnée à prédire et non au cours de l'historique.

- Soit K le nombre des paramètres qualitatifs connus pour expliquer la variable Y. Dans une procédure de sélection de variables, la structure particulière des X_k, matrices des variables explicatives des modèles possibles, implique une gestion de celles-ci par blocs : introduire (ou supprimer) une nouvelle variable p dans le modèle, revient à ajouter (ou retirer) à la matrice précédente celle χ_p des variables quantitatives associées à l'ensemble des modalités de la variable p.

Pour aller plus loin sur les particularités de ce type de modèle, on peut étudier le plan d'analyse de la variance à deux facteurs avec interaction,

$$Y_{ijr} = m + a_i + b_j + c_{ij} + e_{ijr} , \qquad e_{ijr} \approx \aleph_1\left(0, \sigma^2\right)$$

où Y_{ijr} désigne une variable aléatoire dépendant des deux caractères qualitatifs i et j, i ayant I modalités et j, J modalités, pour un ensemble de n_{ij} observations indépendantes de Y_{ijr}. On a $n = \sum_{1 \le i \le I,\ 1 \le j \le J} n_{ij}$.

Ce plan d'analyse de la variance peut encore s'écrire sous la forme du modèle de régression, $y = X_3 \beta + \varepsilon$, $\varepsilon \approx \aleph_n\left(0, \sigma^2 I_n\right)$, où $X_3 = \left(\begin{array}{c|c|c|c} 1 & & & \\ \vdots & \chi_1 & \chi_2 & \chi_3 \\ 1 & & & \end{array}\right)$

avec χ_1 le tableau disjonctif à I colonnes associé au caractère i

χ_2 le tableau disjonctif à J colonnes associé au caractère j

χ_3 le tableau disjonctif à I×J colonnes associé au croisement d'ordre 2 entre les caractères i et j.

Notons $\bar{y}_{ij.} = \frac{1}{n_{ij}} \sum_{r=1}^{n_{ij}} y_{ijr}$, $\bar{y}_{i..} = \frac{1}{J} \sum_{j=1}^{J} \bar{y}_{ij.}$, $\bar{y}_{.j.} = \frac{1}{I} \sum_{i=1}^{I} \bar{y}_{ij.}$ et $\bar{y}_{...} = \frac{1}{IJ} \sum_{j=1}^{J} \sum_{i=1}^{I} \bar{y}_{ij}$.

Rappelons que $\text{Rang}(X_3) = 1 + (I-1) + (J-1) + (IJ - I - J + 1) = IJ$ et que les estimateurs du maximum de vraisemblance sont donnés par : $\hat{m} = \bar{y}_{...}$, $\hat{a}_i = \bar{y}_{i..} - \bar{y}_{...}$ pour $1 \le i \le I$, $\hat{b}_j = \bar{y}_{.j.} - \bar{y}_{...}$ pour $1 \le j \le J$, et $\hat{c}_{ij} = \bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...}$ pour $1 \le i \le I$ et $1 \le j \le J$.

Il en résulte l'estimateur de Y_{ijr} : $\hat{y}_{ijr} = \hat{m} + \hat{a}_i + \hat{b}_j + \hat{c}_{ij} = \bar{y}_{ij.}$

Cet exemple permet de remarquer que la suppression des paramètres a_i et/ou b_j ne change pas le calcul de l'estimation $\hat{y}_{ijr}$: la matrice des variables explicatives est toujours de même rang et les estimateurs des paramètres, modifiés en conséquence, donnent toujours les mêmes estimations et les mêmes prévisions. Ceci illustre le caractère singulier des modèles de type analyse de la variance : pour ceux-ci, la présence d'une variable simple est facultative si elle est représentée par une variable résultant de son interaction avec d'autres. Pour en revenir aux procédures de sélection, ceci entraîne la possibilité d'obtenir des critères $\hat{R}_n(k)$ et $\hat{R}_n^*(k)$ stationnaires pour des modèles de tailles différentes. Il conviendra alors d'être prudent sur le choix des variables dans les procédures automatiques de sélection de variables.

5. Exemple d'application

On souhaite trouver le modèle de régression optimal pour un problème de prévision de trafic routier qui consiste à prévoir, un an à l'avance, le débit journalier en un point donné du réseau. C'est ce type de prévisions qui est actuellement utilisé par le CNIR (Centre National d'Information Routière) pour gérer la mise en place du dispositif Bison Futé (Ziani et Danech-Pajouh, 98). Les informations connues à aussi long terme étant limitées, on ne peut tenir compte ni des notions d'écoulement de trafic, ni des prévisions météorologiques ou concernant d'éventuels travaux et seuls les renseignements fournis par le calendrier seront utilisables comme paramètres explicatifs : jour de la semaine, mois, départs ou retours en vacances, ponts... Ces données sont appelées les caractéristiques calendaires. En fait, pour tenir compte de la tendance annuelle du trafic journalier, le modèle utilise plutôt la variable aléatoire du débit journalier relatif qui est le rapport du débit journalier sur sa tendance annuelle (ou Trafic Moyen Journalier Annuel: TMJA). Les TMJA, quant à eux, sont traités séparément par une méthode différente plus adaptée.

Notons q_r le débit relatif journalier, le modèle recherché est du type,

$$q_r = X_k \beta_k + \varepsilon_k , \qquad \varepsilon_k \approx N_n\left(m_k, \sigma^2 I_n\right)$$

où n désigne le nombre d'observations de l'historique et où k est le nombre de variables explicatives du modèle, $1 \le k \le K$. K est le nombre total des variables explicatives possibles. Du fait de la nature qualitative des paramètres explicatifs, nous sommes dans le cadre du **plan d'analyse de la variance** développé dans la section précédente.

Pour évaluer nos deux procédures de sélection de variables, nous avons choisi de présenter les résultats obtenus pour la station de mesures de St Arnoult, dans le sens Paris province. Les calculs ont été faits dans le cadre des prévisions de l'année 1998, avec un historique prenant en compte les débits journaliers du 1er Janvier 1987 au 31 Décembre 1997, à l'exception de ceux de 1992, soit n=3652. Nous avons pris comme modèle de départ, celui de taille finie utilisé actuellement par le logiciel Bison Futé. Le nombre des paramètres explicatifs est de 34, tous sont issus des caractéristiques calendaires, 13 sont des caractéristiques simples (jour de la semaine, mois, présence d'un jour férié...) et 21 sont issues de croisements d'ordre 2 entre ces variables simples (jour de la semaine croisé avec la variable présence d'un jour férié par exemple...). Nous pouvons également raisonnablement supposer ce modèle de taille infinie car il revient à utiliser un modèle comportant 917 paramètres et $r_n(K) = 517$.

A cause de la présence d'un tel nombre de variables explicatives qualitatives, cela rend peu réaliste la recherche exhaustive du modèle minimisant l'un des critères empiriques, c'est à dire sur l'ensemble des $2^{34} - 1$ modèles possibles. Nous avons donc choisi d'utiliser une procédure de recherche approchée de type « pas à pas descendant » : elle élimine les variables une à une, en choisissant à chaque étape celle permettant d'obtenir le critère le plus petit.

Nous avons également conservé toutes les variables issues de caractéristiques calendaires simples avant d'utiliser notre procédure de sélection, malgré la présence facultative de certaines (voir section 4). En effet, ceci a permis d'éviter de perdre trop d'information dans la régression quand une variable caractéristique d'un croisement de paramètres simples était supprimée.

Pour trouver le modèle optimal, on a cherché :

- pour la méthode traditionnelle, le modèle $\hat{\kappa}$ minimisant le critère empirique,

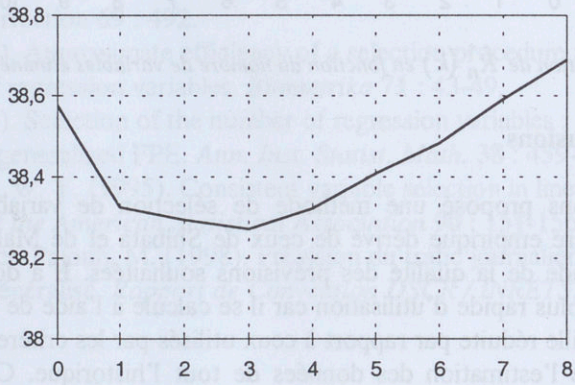
$$\hat{R}_n(k) = \left\| y - X_k \hat{\beta}_k \right\|^2 + 2\, r_n(k)\, \hat{\sigma}^2_K$$

- pour celle utilisant le critère en prévision, le modèle $\hat{\kappa}^*$ minimisant,

$$\hat{R}^*_n(k) = \left\| X^*_K \hat{\beta}_K - X^*_k \hat{\beta}_k \right\|^2 + 2\, \hat{\sigma}^2_K \, \text{trace}\left(X^*_k \left(X'_k X_k\right)^{-1} X'^*_k\right)$$

avec $\hat{\sigma}^2_K = \left\| y - X_K \hat{\beta}_K \right\|^2 / n$ un estimateur de σ^2 obtenu avec le modèle complet.

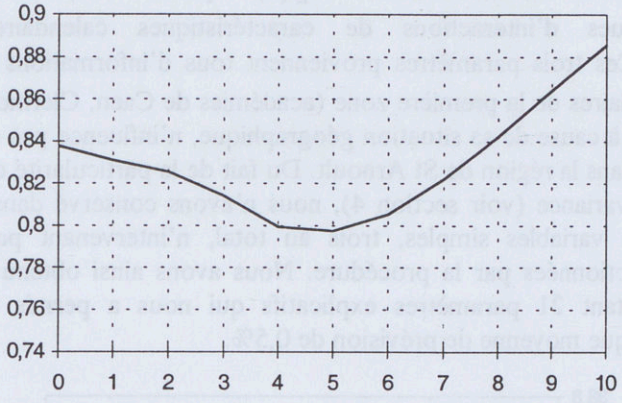
La procédure utilisant le critère $\hat{R}_n(k)$ a permis d'éliminer 3 variables explicatives issues d'interactions de caractéristiques calendaires simples ($r_n(\hat{\kappa}) = 495$). Ces trois paramètres proviennent tous d'informations concernant les vacances scolaires de la première zone (académies de Caen, Clermont-Ferrand, Grenoble...)qui, à cause de sa situation géographique, n'influence pas directement le trafic routier dans la région de St Arnoult. Du fait de la particularité des modèles d'analyse de la variance (voir section 4), nous n'avons conservé dans le modèle définitif que les variables simples, trois au total, n'intervenant pas dans les interactions sélectionnées par la procédure. Nous avons ainsi obtenu un modèle optimal comportant 21 paramètres explicatifs qui nous a permis d'améliorer l'erreur quadratique moyenne de prévision de 0,5%.



Evolution de $\hat{R}_n(k)$ *en fonction du nombre de variables éliminées*

Quant à la procédure mise en place pour la prévision, elle a supprimé plus de paramètres explicatifs : ceux éliminés par la procédure précédente l'ont été également, avec en plus deux autres variables, dont une provenant d'une information calendaire simple ($r_n(\hat{\kappa}^*) = 483$). En fait, toutes les informations relatives aux vacances de la première zone ont disparu du modèle, à l'exception de celle caractérisant la présence ou non de vacances pour la zone. Ceci a permis d'obtenir une amélioration de l'erreur quadratique moyenne de prévision comparable à celle obtenue dans le cas précédent mais avec une stabilité des résultats plus importante à cause d'un nombre de variables explicatives retenues plus restreint. De plus, la méthode en prévision a eu l'avantage de donner un modèle optimal plus rapidement en raison de l'utilisation de calculs numériques plus simples car effectués sur des matrices de taille réduite. Le logiciel MATLAB a ainsi permis d'obtenir un modèle optimal trois fois plus rapidement qu'avec le

critère traditionnel. Ceci n'est donc pas négligeable lorsque ce calcul peut prendre plusieurs heures : avec nos moyens informatiques, il nous en a fallu une trentaine pour trouver le modèle optimal pour le nouveau critère.



Evolution de $\hat{R}_n^*(k)$ *en fonction du nombre de variables éliminées*

6. Conclusions

Nous avons proposé une méthode de sélection de variables basée sur l'étude d'un critère empirique dérivé de ceux de Shibata et de Mallows. Celui-ci est basé sur l'étude de la qualité des prévisions souhaitées. Il a donc l'avantage d'être beaucoup plus rapide d'utilisation car il se calcule à l'aide de matrices et de vecteurs d'une taille réduite par rapport à ceux utilisés par les critères traditionnels qui dépendent de l'estimation des données de tout l'historique. Cela rend ainsi notre critère particulièrement intéressant pour des modèles comportant un grand nombre de paramètres explicatifs, comme dans le cadre de l'exemple présenté. Si la diminution de l'erreur quadratique moyenne de prévision s'est avérée comparable à celle obtenue par le modèle retenu avec le critère de Mallows, la stabilité des prévisions a été améliorée car moins de paramètres explicatifs ont été retenus. De plus, les variables éliminées se sont avérées avoir une interprétation cohérente avec la réalité.

Reste le caractère approximatif dans le choix de ce critère empirique, mais une étude de la convergence de celui-ci vers son pendant théorique, le risque quadratique moyen, est en cours, ce qui permettrait de l'ajuster et d'améliorer les premiers résultats, déjà encourageants.

Bibliographie

Akaike, H. (1970). Statistical predictor identification. *Ann. Inst. Statist. Math.* **22** :203-17.

Akaike, H. (1973). Information theory and an extention of the maximum likelihood principle. In *2nd International Symposium on Information Theory*, Eds B. N. Petrov and F. Csaki, pp. 267-81. Budapest : Akadémia Kiado.

Barron, A., Birgé, L. & Massart, P. (1999). Risk bounds for model selection via penalization. *Probab. Theory Relat. Fields* **113** : 301-413.

Foster, D. P. & George, E. I. (1994). The risk inflation criterion for multiple regression. *Ann. Statist.* **22** : 1947-75.

Hannan, E. J. & Quinn, B. G. (1979). The determination of the order of an autoregression. *J . R. Statist. Soc. B.* **41** : 190-5.

Mallows, C. L. (1973). Some comments on C_p *Technometrics* **15** : 661-75.

Rao, C. R. & Wu, Y. (1989). A strongly consistent procedure for model selection in a regression problem. *Biometrika* **76** : 2, 369-74.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** : 461-4.

Shibata, R. (1981). An optimal selection of regression variables. *Biometrika* **68** : 45-54, Correction **69** : 492.

Shibata, R. (1984). Approximate efficiency of a selection procedure for the number of regression variables. *Biometrika* **71** : 43-49.

Shibata, R. (1986). Selection of the number of regression variables ; a minimax choice of generalized FPE. *Ann. Inst. Statist. Math.* **38** : 459-74.

Zheng, X. & Loh, W. Y. (1995). Consistent variable selection in linear models. *Journal of the American Statistical Association* **90** : 151-156.

Ziani, A. & Danech-Pajouh, M. (1998). Prévision du trafic journalier. Modèle Linéaire généralisé. *Rapport de Convention DSCR / INRETS.*