UNIVERSITE PAUL SABATIER TOULOUSE III

PUBLICATIONS DU LABORATOIRE
DE
STATISTIQUE ET PROBABILITÉS

<u>Cours de DEA :</u>

Sélection de modèle en régression linéaire

Jean-Marc Bardet

Laboratoire de Statistique et Probabilités — UMR CNRS C5583
Université Paul Sabatier — 31062 – Toulouse cedex 4.

Table des matières

Introduction

Dans de nombreux problèmes empiriques se pose la question du choix du modèle. En effet, au cours d'une expérience quelconque, on peut disposer de données relatives à des variables (dites explicatives) dont on ne sait pas a priori si elles ont une influence réelle sur la ou les variables d'intérêt. On peut rajouter ainsi toutes les variables intervenant plus ou moins lors de l'expérience et bricoler un modèle de plus en plus précis pour s'adapter aux données obtenues. Cependant, que ce soit pour comprendre, expliquer, ou prédire, il est absolument nécessaire de sélectionner les variables intervenant vraiment
Prenons pour exemple la clarté d'un béton, donnée par un niveau de gris allant de 0 à 100. De nombreux facteurs peuvent potentiellement avoir un rôle sur cette quantité : le ciment utilisé et sa quantité, les adjuvants, ajouts et leurs quantités, le rapport eau sur liant, la température et le temps du coffrage,... Notons que dans un tel problème se mêlent des variables quantitatives et qualitatives et qu'une analyse de la covariance serait une méthode de choix possible parmi toutes ses variables potentiellement explicatives.
Si aucune information de nature causale, par exemple une loi de la physique, n'est connue quant au problème considéré, il n'y a aucun a priori à avoir vis-à-vis de toutes les variables intervenant dans l'expérience (la seule limite étant qu'il ne faut pas plus de variables explicatives que de données à étudier). Ainsi, une infinité de modèles possibles pourraient servir à représenter la variable d'intérêt en fonction des variables explicatives : des modèles linéaires, non-linéaires, à cartes, des réseaux de neurones, des modèles additifs, des séries chronologiques,... Dans chacune de ces configurations, une fois le type de modèle choisi, il faut encore le décrire en estimant les variables explicatives intervenant "vraiment" (c'est-à-dire avec un certain niveau de confiance) et les paramètres du modèle.

Nous avons choisi d'étudier le modèle linéaire, en raison de sa relative généralité, des nombreuses études existant sur ce sujet, et du fait que de nombreux problèmes concrets, notamment issus de lois physico-chimiques, se ramènent au cas linéaire. Nous avons de plus choisi de ne traiter que le cas de variables quantitatives ; indiquons que l'analyse de la covariance sera un moyen de sélectionner le modèle pour des variables pouvant être quantitatives et qualitatives. En particulier, nous commencerons par étudier 2 critères particuliers, le Cp de Mallows et le critère AIC d'Akaiké. Nous généraliserons alors aux critères avec pénalisation, et aborderons le cas de le cas de la régression fonctionnelle. L'essentiel des résultats et des techniques utilisés sont circonscrites au cas gaussien. Cependant, il pourra exister des généralisations à d'autres distributions de probabilité...

1 Sélection de modèle en régression linéaire paramétrique gaussienne

1.1 Présentation et définition

Hypothèses et notations : On suppose que l'on dispose des n résultats d'une expérience concernant une variable d'intérêt quantitative Y ayant pris n valeurs, $(y_1, y_2, \cdots, y_n)$, et que p variables quantitatives X_i, $i = 1, \cdots, p$, prenant les valeurs $x_{1,i}, \cdots, x_{n,i}$, peuvent être utilisée pour expliquer Y. Par soucis de concision dans les notations, on supposera que $X^1 = (1, \cdots, 1)'$, qui correspond à la partie constante de la régression. Enfin, on définit $\mathcal{M}$ un ensemble de modèles, qui est une famille de sous-ensemble de $\{1, \cdots, p\}$. Deux exemples sont à retenir :

- $\mathcal{M} = \mathcal{P}(\{1, \cdots, p\})$: famille exhaustive de modèles ;
- $\mathcal{M} = \{1, \cdots, j\}_{1 \leq j \leq p}$: famille hiérarchique de modèles.

Par la suite, pour $m \in \mathcal{M}$, on notera $|m| = \mathrm{Card}(m)$ et $X^{(m)}$ la matrice d'ordre $(n, |m|)$ telle que $[X_{i_1}, X_{i_2}, \cdots, X_{i_{|m|}}]$, avec $m = \{i_1, \cdots, i_{|m|}\}$.

Hypothèses sur le vrai modèle : On suppose qu'il existe $m^* \in \mathcal{M}$ tel que le vrai modèle s'écrive :

$$Y = \mu^* + \varepsilon^* = X^{(m^*)}.\beta_{(m^*)} + \varepsilon^*, \quad \text{où} \quad \varepsilon^* \sim \mathcal{N}(0, \sigma_*^2 I_n), \tag{1}$$

et avec m^*, $\mu^* \in \mathbb{R}^n$, $\beta_{(m^*)} \in (\mathbb{R} \setminus \{0\})^{|m^*|}$, ε^* et σ_* inconnus.

Exemple : Le vrai modèle peut par exemple être :

$$Y = \beta_1 + \beta_3 X_3 + \beta_7 X_7 + \varepsilon^*,$$

ce qui signifie que $m^* = \{1, 3, 7\}$.

Remarque : Observons que :

i. Le bruit est un bruit blanc gaussien (les ε_i^* sont indépendants) ;

ii. On suppose que $\beta_{(m^*)} \in (\mathbb{R} \setminus \{0\})^{|m^*|}$; ceci permet l'écriture unique du modèle. En effet, si $m \in \mathcal{M}$ contient strictement $m*$, alors le vrai modèle pourra également s'écrire $Y = X^{(m)}.\beta_{(m)} + \varepsilon^*$, mais au moins une des coordonnées de $\beta_{(m)}$ sera nulle.

Modèles utilisés : Pour modéliser l'expérience, on utilise la famille de modèles suivante, qui est également en correspondance avec $\mathcal{M}$, soit

$$Y = \mu + \varepsilon = X^{(m)}.\beta_{(m)} + \varepsilon, \quad \text{où} \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I_n), \tag{2}$$

avec $m \in \mathcal{M}$, $\mu \in \mathbb{R}^n$, $\beta_{(m)} \in (\mathbb{R} \setminus \{0\})^{|m|}$.

Pour préciser la modélisation, nous utiliserons les définitions suivantes :

Définition : On suppose que m_1, m_2 et m_3 sont trois ensembles non vides et disjoints de $\{1, \cdots, p\}$ et que $m^* = m_1 \cup m_2$. Alors :

- si le modèle utilisé est m_1, on dit que le modèle est sous-ajusté ;
- si le modèle utilisé est $m_1 \cup m_2 \cup m_3$, on dit que le modèle est sur-ajusté.

Exemple : En reprenant l'exemple précédent, le modèle $Y = \beta_1 + \beta_7 X_7 + \varepsilon$ est sous-ajusté, quand le modèle $Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_6 X_6 + \beta_7 X_7 + \varepsilon$ est lui sur-ajusté.

Rappels : On suppose que l'on choisit le modèle $m \in \mathcal{M}$. Alors :

i. L'estimateur de $\beta_{(m)}$ par moindres carrés ordinaires est :

$$\widehat{\beta}_{(m)} = ((X^{(m)})'.X^{(m)})^{-1}(X^{(m)})'.Y \quad \text{d'où} \quad \widehat{Y}_{(m)} = X^{(m)}.\widehat{\beta}_{(m)}.$$

Dans le cadre gaussien précédent, cet estimateur est également l'estimateur par maximum de vraisemblance.

ii. L'estimateur par maximum de vraisemblance de σ^2 est

$$\widehat{\sigma}^2 = \frac{1}{n}\,||Y - \widehat{Y}_{(m)}||^2 = \frac{1}{n}(Y - \widehat{Y}_{(m)})'.(Y - \widehat{Y}_{(m)}).$$

C'est un estimateur biaisé, mais asymptotiquement sans biais. On utilise également l'estimateur non biaisé :

$$\widehat{s}^2 = \frac{1}{(n - |m|)}||Y - \widehat{Y}_{(m)}||^2 = \frac{1}{(n - |m|)}(Y - \widehat{Y}_{(m)})'.(Y - \widehat{Y}_{(m)}).$$

1.2 Distance entre deux modèles

Pour mesurer l'écart entre le vrai modèle (avec m^*) et un modèle candidat (soit $m \in \mathcal{M}$), nous utiliserons les deux distances suivantes.

Distance quadratique

Définition : La distance quadratique entre les modèles M et M^* est :

$$L_2(M, M^*) = \frac{1}{n}\,||X^{(m^*)}.\beta_{(m^*)} - X^{(m)}.\beta_{(m)}||^2 = \frac{1}{n}\,(X^{(m^*)}.\beta_{(m^*)} - X^{(m)}.\beta_{(m)})'.(X^{(m^*)}.\beta_{(m^*)} - X^{(m)}.\beta_{(m)}).$$

Remarquons que le vecteur de paramètres $\beta_{(m)}$ n'est a priori pas plus connu que le vecteur de paramètres $\beta_{(m^*)}$ du vrai modèle. Aussi va-t-on plutôt considérer $L_2(\widehat{M}, M^*)$, où $\widehat{M}$ est obtenu par moindres carrés. Dans le cas où $m = m*$, on a plus précisément

$$\mathbb{E}L_2(\widehat{M}, M^*) = \mathbb{E}\left(\frac{1}{n}\,||X^{(m^*)}.\beta_{(m^*)} - X^{(m)}.\widehat{\beta}_{(m)}||^2\right) = \frac{1}{n}\mathrm{Var}\left(\widehat{Y}_{(m)}\right),$$

ce qui, a une constante près, correspond au risque quadratique portant de l'estimateur $\widehat{Y}_{(m)}$ de μ^*.

Dissemblance de Kullback

Définition : Soit P et P^* deux mesures de probabilité dominées par une même mesure ν. La dissemblance de Kullback entre P et P^* est :

$$K(P, P^*) = \mathbb{E}_{P^*}\left(\log \frac{dP^*}{dP}\right).$$

En particulier, si $p = \dfrac{dP}{d\nu}$ et $p^* = \dfrac{dP^*}{d\nu}$, alors, $K(P, P^*) = \begin{cases} \int p^* \log \dfrac{p^*}{p} d\nu & \text{si } P^* << P; \\ 0 & \text{sinon.} \end{cases}$

On montre par des arguments de convexité que $K(P, P^*) \geq 0$ pour toutes mesures P et P^*, et que $K(P, P^*) = 0$ si et seulement si $P = P^*$ P presque sûrement.

Remarque : Pour des raisons de simplicité d'écriture, nous utiliserons plutôt la dissemblance entre les modèles M et M^* définie par $K(M, M^*) = \dfrac{2}{n}K(P, P^*)$, où P et P^* sont les mesures de probabilité respectives de Y dans le cadre des modèles M et M^*.

Dans le cadre gaussien initialement proposé, on a considère des densités par rapport à la mesure de Lebesgue et :

$$\begin{aligned} f_Y^*(y) &= \frac{1}{(2\pi\sigma_*^2)^{n/2}} \exp\left(\frac{1}{2\sigma_*^2}(y-\mu^*)'.(y-\mu^*)\right) \\ f_Y(y) &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(\frac{1}{2\sigma^2}(y-\mu)'.(y-\mu)\right) \end{aligned}$$

Ainsi, en considérant que Y suit la densité f^*, on a :

$$K(M, M^*) = \log\left(\frac{\sigma_*^2)}{\sigma^2}\right) - \frac{1}{n}\mathbb{E}\left[\frac{1}{\sigma_*^2}||Y-\mu^*||^2 - \frac{1}{\sigma^2}||Y-\mu||^2\right]$$

Mais comme $\mathbb{E}(Y) = \mu^*$, alors $\mathbb{E}\left[(Y-\mu^*)'.(Y-\mu^*) - (Y-\mu)'\right] = \mathbb{E}(||\varepsilon^*||^2) = n\sigma_*^2$.

De plus, $||Y-\mu||^2 = ||Y-\mu^*||^2 + 2(Y-\mu^*)'.(\mu^*-\mu) + ||\mu^*-\mu||^2$, donc avec ce qui précède

$$K(M, M^*) = \log\left(\frac{\sigma^2}{\sigma_*^2}\right) + \frac{\sigma_*^2}{\sigma^2} - 1 + \frac{1}{\sigma^2}L_2(M, M^*).$$

Par la suite, et contrairement à ce que nous avons pu faire avec la distance quadratique, nous utiliserons la dissemblance de Kullback pour $M = \widehat{M}$. Pour ce faire, on remplacera μ et σ^2 par leurs estimations respectives, $\widehat{\mu}$ et $\widehat{\sigma}^2$ qui dépendent alors de Y et en prenant l'espérance (sous la loi du vrai modèle) de cette formule.

1.3 Deux critères pour sélectionner un modèle

Rappels : 1/ **Loi de Fisher :** Soit U_1 et U_2 deux variables indépendantes suivant respectivement des lois $\chi^2(n_1)$ et $\chi^2(n_2)$ avec $(n_1, n_2)(\mathbb{N}^*)^2$. Alors

$$F = \frac{U_1/n_1}{U_2/n_2} \quad \text{suit une loi de Fisher} \quad F(n_1, n_2),$$

telle que $\mathbb{E}F = \dfrac{n_2}{n_2 - 2}$ si $n_2 > 2$ et $\mathrm{Var}F = \dfrac{2n_2^2(n_1 + n_2 - 2)}{n_1(n_2 - 4)(n_2 - 2)^2}$ si $n_2 > 4$.

2/ **Théorème de Cochran :** Soit $X = (X_1, \cdots, X_n)$ un n-échantillon de loi $\mathcal{N}(0, \sigma^2)$ et soit $||.||$ la norme euclidienne classique. Alors :

i. Les composantes de X dans toute base orthonormée de $\mathbb{R}^n$ forment un n-échantillon de loi $\mathcal{N}(0, \sigma^2)$.

ii. Si $\mathbb{R}^n = E_1 + \cdots + E_p$ où les E_i sont p sous-espaces vectoriels deux à deux orthogonaux et de dimensions respectives $d_1, \cdots, d_p$, alors si on note $P_i(X)$ le projeté orthogonal de X sur E_i pour $i = 1, \cdots, p$, les $P_i(X)$ sont indépendants, ainsi que les $||P_i(X)||^2$, dont chacun a pour loi $\sigma^2.\chi^2(d_i)$.

3/ **Définition :** Soit $X = (X_1, \cdots, X_n)$ un n-échantillon de loi P_θ où $\theta \in \mathbb{R}^p$ et soit $\widehat{\theta}_n$ un estimateur de θ. Alors :

i. Le biais de $\widehat{\theta}_n$ est $\mathbb{E}(\widehat{\theta}_n - \theta)$.

ii. Le risque quadratique de $\widehat{\theta}_n$ est $\mathbb{E}\,||\widehat{\theta}_n - \theta||^2$.

Minimisation de la distance quadratique

Une idée naturelle pour sélectionner le "meilleur modèle" est de minimiser le risque quadratique. Cela revient aussi à minimiser l'espérance de la distance L_2 définie un peu plus haut. Dans le cadre dans lequel nous nous sommes placés, pour un modèle $m \in \mathcal{M}$, le risque quadratique s'écrit :

$$\begin{aligned} \mathbb{E}\,||\widehat{Y}_{(m)} - \mu^*||^2 &= n\mathbb{E}\,L_2(\widehat{M}, M^*) = \mathbb{E}\,||X^{(m)}.\widehat{\beta}_{(m)} - \mu^*||^2 \\ &= \mathbb{E}\,||X^{(m)}.\widehat{\beta}_{(m)} - \mu^*_{(m)}||^2 + \mathbb{E}\,||\mu^* - \mu^*_{(m)}||^2, \end{aligned}$$

en appelant $\mu^*_{(m)}$ le projeté orthogonal de μ^* sur le sous-espace vectoriel engendré par les X^i où $i \in m$ (sous-espace que nous noterons $< X^{(m)} >$ par la suite), la dernière égalité découlant du Théorème de Pythagore. Par linéarité de la projection et de l'espérance, il est clair que $\mathbb{E}\,\widehat{Y}_{(m)} = \mu^*_{(m)}$. En conséquence, comme $\widehat{Y}_{(m)}$ est la projection orthogonale de Y sur $< X^{(m)} >$, alors :

$$\begin{aligned} \mathbb{E}\,||\widehat{Y}_{(m)} - \mu^*||^2 &= \mathbb{E}\,||\varepsilon^*_{(m)}||^2 + \mathbb{E}\,||\mu^* - \mu^*_{(m)}||^2 \\ &= |m|\sigma_*^2 + ||\mu^* - \mu^*_{(m)}||^2, \end{aligned}$$

le membre de droite de cette égalité ne nécessitant pas d'espérance puisqu'il est déterministe. Notons que $k\sigma_*^2$ aurait aussi été obtenu dans un cadre non gaussien (ici le Théorème de Cochran peut être utilisé).

Remarque : On obtient ainsi l'expression générale du risque quadratique pour un modèle quelconque. Remarquons qu'il est composé de deux parties comportant des paramètres inconnus. Il n'est donc pas possible de minimiser directement ce risque. On observe également que la première partie est un terme de variance qui augmente avec la dimension du modèle choisi, quant la deuxième partie est un terme de biais, qui diminue quand augmente la dimension du modèle, jusqu'à s'annuler quand celle-ci est supérieure ou égale à la vraie dimension du modèle (soit $|m^*|$). On a donc affaire à un compromis biais-variance, qui heuristiquement s'explique par le fait que plus la dimension du modèle augmente, meilleur est l'ajustement du modèle aux données, mais plus importante est la variabilité du modèle (donc la possible erreur lors d'une prédiction ou d'un lissage).

Pour minimiser la distance quadratique L_2 et ainsi sélectionner un "meilleur modèle" suivant ce critère, on est amené à utiliser des estimations des parties biais et variance. En effet, on a :

$$\begin{aligned}
\mathbb{E}\,||Y-\widehat{Y}_{(m)}||^2 &= \mathbb{E}\,||Y-\mu^*_{(m)}||^2 - \mathbb{E}\,||\widehat{Y}_{(m)}-\mu^*_{(m)}||^2 \quad \text{(Pythagore)}\\
&= \mathbb{E}\,||Y-\mu^*+\mu^*-\mu^*_{(m)}||^2 - \mathbb{E}\,||\widehat{Y}_{(m)}-\mu^*_{(m)}||^2\\
&= \mathbb{E}\,||Y-\mu^*||^2 + ||\mu^*-\mu^*_{(m)}||^2 - |m|\sigma_*^2\\
&= (n-|m|)\sigma_*^2 + ||\mu^*-\mu^*_{(m)}||^2
\end{aligned}$$

(ici encore cette expression est vraie dans un cadre non gaussien). Par suite, la partie biais $||\mu^*-\mu^*_{(m)}||^2$ du risque quadratique est estimée puisque par substitution :

$$\begin{aligned}
n\mathbb{E}\,L_2(\widehat{M},M^*) &= |m|\sigma_*^2 + \mathbb{E}\,||Y-\widehat{Y}_{(m)}||^2 - (n-|m|)\sigma_*^2\\
\text{soit } \frac{\mathbb{E}\,L_2(\widehat{M},M^*)}{\sigma_*^2} &= 2\,\frac{|m|}{n} - 1 + \frac{\widehat{\sigma}^2_{(m)}}{\sigma_*^2}
\end{aligned}$$

en estimant $\mathbb{E}\,||Y-\widehat{Y}_{(m)}||^2$ par $||Y-\widehat{Y}_{(m)}||^2$, ce qui n'est pas illégitime car, par exemple, lorsque $m* \in m$, alors $\widehat{\sigma}^2_{(m)}$ est un estimateur convergent de $\dfrac{\mathbb{E}\,||Y-\widehat{Y}_{(m)}||^2}{n}$. On conserve cependant un terme inconnu dans cette estimation de la distance quadratique, à savoir σ_*, issu de la partie variance du risque quadratique. Or on sait que $m^* \in \mathcal{M}$, donc en considérant $m_p = \{1,\cdots,p\}$, $\widehat{\sigma}_{(m_p)}$ est un estimateur convergent de σ_* tant que p ne varie pas avec n, ce que nous avons supposé. Ainsi nous utiliserons pour minimiser notre distance quadratique une minimisation du critère appelé Cp de Mallows et valant pour $m \in \mathcal{M}$, :

$$Cp(m) = \frac{\widehat{\sigma}^2_{(m)}}{\widehat{\sigma}^2_{(m_p)}} + 2\,\frac{|m|}{n}.$$

En selectionnera donc le modèle $\widehat{m}$ tel que :

$$\widehat{m} = \min_{m\in\mathcal{M}}\{Cp(m)\}.$$

Ce critère a été introduit par Mallows (voir Mallows 1967 et 1995). Nous étudierons ses propriétés un peu après.

Minimisation de la dissemblance de Kullback

Comme précédemment, la sélection du modèle peut se faire en minimisant une distance, ici la dissemblance de Kullback, entre un modèle candidat et le vrai modèle. Mais comme précédemment, cela ne pourra être fait qu'après certaines approximations. Soit $m \in \mathcal{M}$, la dissemblance de Kullback entre la loi de l'estimation issue de ce modèle et le vrai modèle :

$$K(\widehat{M}, M^*) = \mathbb{E}_{M^*}\left[\log\left(\frac{\widehat{\sigma}^2_{(m)}}{\sigma^2_*}\right) + \frac{\sigma^2_*}{\widehat{\sigma}^2_{(m)}} - 1 + \frac{1}{\widehat{\sigma}^2_{(m)}} L_2(\widehat{M}, M^*)\right].$$

Le calcul de cette espérance étant dans le cas général non obtensible, on va faire l'hypothèse que $m^* \in m$. Alors, sous l'hypothèse gaussienne :

$$\sigma^2_{(m)} = \frac{1}{n}||Y - \widehat{Y}_{(m)}||^2 = \frac{1}{n}||P_{(m)^\perp}\varepsilon^*||^2 \sim \frac{\sigma^2_*}{n}\chi^2(n - |m|)$$

d'après le Théorème de Cochran et en appellant $P_{(m)^\perp}$ l'opérateur de projection orthogonale sur $< X^{(m)} >^\perp$. De même,

$$L_2(\widehat{M}, M^*) = \frac{1}{n}||\widehat{Y}_{(m)} - \mu^*||^2 = \frac{1}{n}||P_{(m)}\varepsilon^*||^2 \sim \frac{\sigma^2_*}{n}\chi^2(|m|)$$

et d'après ce même Théorème de Cochran,

$$L_2(\widehat{M}, M^*) \ \text{ est indépendant de } \ \sigma^2_{(m)}.$$

Par suite,

$$K(\widehat{M}, M^*) = \mathbb{E}_{M^*}\left[\log\left(\widehat{\sigma}^2_{(m)}\right)\right] - \log\left(\sigma^2_*\right) - 1 + \frac{n}{n - |m| - 2} + \frac{|m|}{n - |m| - 2},$$

d'après les expressions de l'espérance d'une variable suivant une loi de Fisher. Il nous reste à déterminer l'expression de $\mathbb{E}_{M^*}\left[\log\left(\widehat{\sigma}^2_{(m)}\right)\right]$. Comme on a supposé que $m^* \in m$, cela est possible, mais de ne sera pas forcément probant car dépendant de σ^2_*. On va donc préférer utiliser un estimateur de cette espérance, à savoir $\log\left(\widehat{\sigma}^2_{(m)}\right)$, qui converge. En enlevant les parties constantes, on voit que minimiser $K(\widehat{M}, M^*)$ revient approximativement à minimiser le critère suivant, noté AIC_c (pour AIC corrigé), établi en 1989, et qui s'écrit pour $m \in \mathcal{M}$,

$$AIC_c(m) = \log\left(\widehat{\sigma}^2_{(m)}\right) + \frac{n + |m|}{n - |m| - 2}.$$

Le modèle sélectionné sera donc le modèle $\widehat{m}$ tel que :

$$\widehat{m} = \min_{m \in \mathcal{M}} \{AIC_c(m)\}.$$

Lorsque $n \to \infty$, on peut trouver un équivalent au critère AIC_c, puisque :

$$\frac{n + |m|}{n - |m| - 2} \sim 1 + 2\frac{|m| + 1}{n}.$$

Il conviendra donc asymptotiquement et approximativement à minimiser le critère noté AIC, tel que :

$$AIC(m) = \log\left(\widehat{\sigma}^2_{(m)}\right) + 2\frac{|m| + 1}{n}.$$

Ce critère, à la constante n près, à été introduit par Akaiké en 1973 (AIC pour Akaiké Information Criterium) qui l'a généralisé sous la forme $AIC(m) = -2 \times \log(\text{Vraisemblance}) + 2 \times$ Nombre de paramètres, ce qui ne le limite plus au cas gaussien. Cependant, ce critère donne de très mauvais résultats pour n petit, et on préférera donc utiliser AIC_c.

Plus généralement, la sélection de modèle se fait en utilisant un critère de type vraisemblance pénalisée, c'est-à-dire que pour un modèle $m \in \mathcal{M}$, il s'écrira sous la forme :

$$\text{Crit}(m) = \text{Vraisemblance} + \text{Pen}(|m|),$$

où la vraisemblance dépendra en général de $\widehat{\sigma}^2_{(m)}$ et décroîtra lorsque $|m|$ croîtra, et la pénalisation Pen est une fonction croissante de $|m|$. Le critère BIC par exemple, s'écrira sur le même modèle, puisque

$$BIC(m) = \log\left(\widehat{\sigma}^2_{(m)}\right) + \frac{\log n}{n}|m|.$$

1.4 Probabilités de sur-ajustement par un critère

Sur-ajuster, on l'a déjà évoqué, c'est choisir un modèle “trop gros”, ou plus précisément et en notant Crit le critère utilisé, c'est choisir par ce critère le modèle $m^* \cup m$, où $m^* \cup m \subset \mathcal{M}$, $m \neq \emptyset$ et $m \cap m^* = \emptyset$. La probabilité de sur-ajustement par le modèle $m^* \cup m$ sera donc :

$$P\left[\text{Crit}(m^* \cup m) \leq \text{Crit}(m^*)\right].$$

Nous allons calculer cette probabilité pour deux des critères définis précédemment.

Cp de Mallows

$$\begin{aligned} P\left[\text{Cp}(m^* \cup m) < \text{Cp}(m^*)\right] &= P\left[\frac{\widehat{\sigma}^2_{(m^*\cup m')}}{\widehat{\sigma}^2_{(m_p)}} + 2\,\frac{|m^*| + |m'|}{n} < \frac{\widehat{\sigma}^2_{(m^*)}}{\widehat{\sigma}^2_{(m_p)}} + 2\,\frac{|m^*|}{n}\right] \\ &= P\left[\frac{\widehat{\sigma}^2_{(m^*)} - \widehat{\sigma}^2_{(m^*\cup m)}}{\widehat{\sigma}^2_{(m_p)}} > 2\,\frac{|m|}{n}\right] \end{aligned}$$

Or $\dfrac{\widehat{\sigma}^2_{(m^*)} - \widehat{\sigma}^2_{(m^*\cup m)}}{\widehat{\sigma}^2_{(m_p)}} = \dfrac{||Y - \widehat{Y}_{(m^*)}||^2 - ||Y - \widehat{Y}_{(m^*\cup m)}||^2}{||Y - \widehat{Y}_{(m_p)}||^2}$. D'après le Théorème de Pythagore, comme $Y - \widehat{Y}_{(m^*)} = P_{(m^*)^\perp}(\varepsilon^*)$ et $Y - \widehat{Y}_{(m^*\cup m)} = P_{(m^*\cup m)^\perp}(\varepsilon^*)$, et $< X_{(m^*\cup m)} >^\perp \subset < X_{(m^*)} >^\perp$,

$$\begin{aligned} P_{(m^*)^\perp}(\varepsilon^*) &= P_{(m^*\cup m)^\perp}(\varepsilon^*) + P_{(m^*)^\perp\cap(m)}(\varepsilon^*) \\ \Longrightarrow \quad ||P_{(m^*)^\perp}(\varepsilon^*)||^2 &= ||P_{(m^*\cup m)^\perp}(\varepsilon^*)||^2 + ||P_{(m)}(\varepsilon^*)||^2 \end{aligned}$$

car $(m^*)^\perp \cap (m) = (m)$. De plus, comme on a également $\widehat{\sigma}^2_{(m_p)} = \dfrac{1}{n} P_{(m_p)^\perp}(\varepsilon^*)$ et comme $m \subset m_p$ alors m est orthogonal à $(m_p)^\perp$ et donc d'après la définition de la loi de Fisher

$$F = \left(\frac{n - |m_p|}{|m|}\right) \frac{\widehat{\sigma}^2_{(m^*)} - \widehat{\sigma}^2_{(m^*\cup m)}}{\widehat{\sigma}^2_{(m_p)}} \sim F(|m|, n - |m_p|)$$

donc la probabilité de sur-ajustement par le modèle $m^* \cup m$ est :

$$P\left[\text{Cp}(m^* \cup m) < \text{Cp}(m^*)\right] = P\left[F > 2\,\frac{n - |m_p|}{n}\right].$$

Notons que le résultat eu encore été plus simple si on avait considéré l'estimateur $s^2_{(m_p)}$ sans biais de σ^2_* au lieu de $\widehat{\sigma}^2_{(m_p)}$. Par suite, lorsque $n \to \infty$ et avec m et m^* fixés, on obtient que :

$$P\left[\mathrm{Cp}(m^* \cup m) < \mathrm{Cp}(m^*)\right] \quad \underset{n\to+\infty}{\longrightarrow} \quad P\left[C_m > 2|m|\right] \quad \text{où} \quad C_m \sim \chi^2(|m|)$$

d'après la Loi des Grands Nombres.
La probabilité générale de sur-ajustement qui est supérieure à celle par un sur-modèle quelconque est donc strictement positive. Il est donc possible que le modèle sélectionné soit plus "gros" que le vrai modèle.

AIC et AIC corrigé

En utilisant les mêmes hypothèses et arguments que pour le Cp de Mallows, on a :

$$\begin{aligned}
P\left[\mathrm{AIC}(m^* \cup m) < \mathrm{AIC}(m^*)\right] &= P\left[\log\left(\frac{\widehat{\sigma}^2_{(m^*\cup m')}}{\widehat{\sigma}^2_{(m^*)}}\right) > 2\,\frac{|m|}{n}\right] \\
&= P\left[\frac{\widehat{\sigma}^2_{(m^*\cup m')} - \widehat{\sigma}^2_{(m^*)}}{\widehat{\sigma}^2_{(m^*)}} > \exp\left(2\,\frac{|m|}{n}\right) - 1\right] \\
&= P\left[F > \frac{n-|m^*|-|m|}{|m|}\exp\left(2\,\frac{|m|}{n}\right) - 1\right] \\
& \quad \text{où} \ \ F \sim F(|m|, n-|m^*|-|m|).
\end{aligned}$$

Les calculs pour le critère AIC corrigé sont exactement du même ordre. Ceci nous permet d'évaluer la probabilité asymptotique de sur-ajustement par le modèle $m \cup m$, puisque comme précédemment, lorsque $n \to \infty$ et avec m et m^* fixés,

$$P\left[\mathrm{AIC}(m^* \cup m) < \mathrm{AIC}(m^*)\right] \quad \underset{n\to+\infty}{\longrightarrow} \quad P\left[C_m > 2|m|\right] \quad \text{où} \quad C_m \sim \chi^2(|m|)$$

On obtient ainsi le même type de comportement asymptotique de la probabilité de sur-ajustement par un modèle donné, et également de la probabilité générale de sur-ajustement que pour le Cp de Mallows.

Sans aucune condition sur les variables explicatives X^j, on vient de montrer assez simplement qu'asymptotiquement le modèle sélectionné a une probabilité positive d'être trop "grand". Les probabilités de sous-ajustement sont elles beaucoup plus délicates à obtenir, puisque $||\mu^* - \mu^*_{(m)}||$ va intervenir dans les calculs. Voyons maintenant plus en détail ce qu'il en est.

1.5 Convergence asymptotique du modèle sélectionné

Nous allons étudier si le modèle choisi par un des critères précédemment définis converge bien vers le "vrai" modèle. Le premier à avoir montrer un tel résultat est Nishii (1984) dont nous allons reprendre l'essentiel des hypothèses et preuves.

Pour commencer, on généralise les définitions des critères en considérant $a > 0$ et pour $m \in \mathcal{M}$,

$$\begin{aligned}
Cp^{(a)}(m) &= n.\frac{\widehat{\sigma}^2_{(m)}}{\widehat{\sigma}^2_{(m_p)}} + a\,|m| \\
AIC^{(a)}(m) &= n.\log\left(\widehat{\sigma}^2_{(m)}\right) + a|m|
\end{aligned}$$

(dans cette partie, pour des raisons de simplicité d'écriture, on considérera AIC et non AIC corrigé, et pour permettre l'établissement de lois limites, on a multiplié Cp et AIC par n).

On définit $\widehat{m}$ le modèle minimisant un critère et les deux ensembles :

$$\begin{array}{rcl} \mathcal{M}_1 & = & \{m \in \mathcal{M} \setminus \{m^*\} \ \mid \ m^* \not\subseteq m\} \\ \text{et } \ \mathcal{M}_2 & = & \{m \in \mathcal{M} \setminus \{m^*\} \ \mid \ m^* \subset m\} \end{array}$$

On a le premier résultat suivant :

Lemme 0.1 *Pour $m \in \mathcal{M}_2$, et avec $\xi^{(a)}(k)$ la variable aléatoire telle que $\xi^{(a)}(k) + ak \sim \chi^2(k)$,*

i. $Cp^{(a)}(m^*) - Cp^{(a)}(m) \underset{n\to+\infty}{\xrightarrow{\mathcal{L}}} \xi^{(a)}(|m| - |m^*|)$;

ii. $AIC^{(a)}(m^*) - AIC^{(a)}(m) \underset{n\to+\infty}{\xrightarrow{\mathcal{L}}} \xi^{(a)}(|m| - |m^*|)$.

<u>*Démonstration*</u> : 1/ On a pour $m \in \mathcal{M}_2$:

$$\begin{array}{rcl} Cp^{(a)}(m^*) - Cp^{(a)}(m) & = & n \, \dfrac{\widehat{\sigma}^2_{(m^*)} - \widehat{\sigma}^2_{(m)}}{\widehat{\sigma}^2_{(m_p)}} + a(|m^*| - |m|) \\ & \overset{\mathcal{L}}{\sim} & n \, \dfrac{\chi^2(|m| - |m^*|)}{\chi^2(n - |m_p|)} + a(|m^*| - |m|) \quad \text{(les deux } \chi^2 \text{ étant indépendants),} \\ & \underset{n\to+\infty}{\overset{\mathcal{L}}{\sim}} & \chi^2(|m| - |m^*|) + a(|m^*| - |m|) \end{array}$$

d'après les résultats déjà vus précédemment.

2/ De même, pour $m \in \mathcal{M}_2$:

$$\begin{array}{rcl} AIC^{(a)}(m^*) - AIC^{(a)}(m) & = & n \, \log\left(\dfrac{\widehat{\sigma}^2_{(m^*)}}{\widehat{\sigma}^2_{(m)}}\right) + a(|m^*| - |m|) \\ & = & n \, \log\left(1 + \dfrac{\widehat{\sigma}^2_{(m^*)} - \widehat{\sigma}^2_{(m)}}{\widehat{\sigma}^2_{(m)}}\right) + a(|m^*| - |m|) \\ & \overset{\mathcal{L}}{\sim} & n \, \log\left(1 + \dfrac{\chi^2(|m| - |m^*|)}{\chi^2(n - |m|)}\right) + a(|m^*| - |m|) \quad (\chi^2 \text{ indépendants}), \\ & \underset{n\to+\infty}{\overset{\mathcal{L}}{\sim}} & \chi^2(|m| - |m^*|) + a(|m^*| - |m|). \qquad \blacksquare \end{array}$$

Pour continuer, nous avons besoin d'hypothèses sur les variables explicatives.

Hypothèse 1 : *On suppose qu'il existe une matrice M définie positive telle que :*

$$\lim_{n\to\infty} \frac{1}{n} (X^{(m_p)})'.(X^{(m_p)}) = M.$$

En pratique, cette condition est assez souvent vérifiée. Ainsi, lorsque chaque X^j est constitué de réalisations indépendantes d'une variable aléatoire X_j de carrés integrable, et les differentes X^j étant indépendantes entre elles, alors la Loi des Grands Nombres montre que $\dfrac{1}{n}(X^{(m_p)})'.(X^{(m_p)}) \underset{n\to+\infty}{\longrightarrow} D$, où D est une matrice diagonale dont les termes diagonaux sont les $\mathrm{E}((X_j)^2)$.

Considérons maintenant la probabilité p_n qu'un ensemble soit sélectionné par un critère, soit pour $m \in \mathcal{M}$,

$$p_n(m) = P(\widehat{m} = m).$$

Enfin, on peut définir le risque quadratique que le modèle $m \in \mathcal{M}$ soit le modèle sélectionné, soit

$$R_n(m) = \mathbb{E}_{m^*}\left(||\mu^* - \widehat{Y}_{(m)}||^2\, \mathbb{I}_{\{\widehat{m}=m\}}\right),$$

et le risque quadratique du modèle sélectionné, soit :

$$R_n = \mathbb{E}_{m^*}\left(||\mu^* - \widehat{Y}_{(\widehat{m})}||^2\right).$$

Il est clair que :

$$R_n = \sum_{m\in\mathcal{M}} R_n(m). \tag{3}$$

On peut alors montrer la double propriété suivante :

Propriété 0.1 *i. Pour $m \in \mathcal{M}_1$, si $\lim\limits_{n\to\infty} n\, p_n(m) = 0$ alors $\lim\limits_{n\to\infty} R_n(m) = 0$.*

ii. Pour $m \in \mathcal{M}_2$, si $\lim\limits_{n\to\infty} p_n(m) = 0$ alors $\lim\limits_{n\to\infty} R_n(m) = 0$.

<u>*Démonstration*</u> : On peut encore écrire que $R_n = \mathbb{E}_{m^*}\left(||\mu^* - \widehat{Y}_{(m)}||^2\, \mathbb{I}_{\{\widehat{m}=m\}}\right)$, soit :

$$\begin{aligned} R_n(m) &= \mathbb{E}_{m^*}\left(||\mu^* - \mu^*_{(m)}||^2\, \mathbb{I}_{\{\widehat{m}=m\}} + ||\mu^*_{(m)} - \widehat{Y}_{(m)}||^2\, \mathbb{I}_{\{\widehat{m}=m\}}\right) \\ &= ||\mu^* - \mu^*_{(m)}||^2\, p_n(m) + \mathbb{E}_{m^*}\left(||\mu^*_{(m)} - \widehat{Y}_{(m)}||^2\, \mathbb{I}_{\{\widehat{m}=m\}}\right) \\ &= I_1 + I_2. \end{aligned}$$

i/ En utilisant l'Hypothèse 1 sur les variables explicatives, on peut écrire que :

$$\begin{aligned} ||\mu^* - \mu^*_{(m)}||^2\, p_n(m) &= \frac{1}{n}\,(\mu^* - \mu^*_{(m)})'.(\mu^* - \mu^*_{(m)})'\, n\, p_n(m) \\ &= \frac{1}{n}\,\beta'_{m_p}.(X^{(m_p)})'.(I_n - Q_{(m)}).X^{(m_p)}.\beta_{m_p} n\, p_n(m), \end{aligned}$$

avec I_n la matrice identité d'ordre n et $Q_{(m)}$ la matrice du projecteur orthogonal sur (m). En fait,

$$Q_{(m)} = X^{(m)}.((X^{(m)})'.X^{(m)})^{-1}(X^{(m)})' = X^{(m_p)}D_{(m)}.(D'_{(m)}(X^{(m_p)})'.X^{(m_p)}D_{(m)})^{-1}.D'_{(m)}(X^{(m_p)})',$$

où $D_{(m)}$ est la matrice d'ordre $(p, |m|)$ composée de 0 et de 1 permettant de sélectionner les vecteurs X^j pour $j \in m$ parmi $X^{(m_p)}$. Par exemple, si $p = 5$ et $m = \{1, 3, 4\}$, alors :

$$D_{(m)} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}.$$

Par suite, en utilisant l'Hypothèse 1, on montre que :

$$\begin{aligned}\lim_{n\to\infty}\frac{1}{n}\,\beta'_{m_p}.(X^{(m_p)})'.X^{(m_p)}.\beta_{m_p} &= \beta'_{m_p}.M.\beta_{m_p}\\ \lim_{n\to\infty}\frac{1}{n}\,\beta'_{m_p}.(X^{(m_p)})'.Q_{(m)}.X^{(m_p)}.\beta_{m_p} &= \beta'_{m_p}.M.D_{(m)}.(D'_{(m)}.M.D_{(m)})^{-1}D'_{(m)}.M.\beta_{m_p},\end{aligned}$$

les deux limites étant des réels positifs ne dépendant plus de n. En conséquence,

$$I_1 \underset{n\to+\infty}{\longrightarrow} 0 \quad\text{car}\quad \lim_{n\to\infty} n.p_n(m)=0.$$

Par ailleurs, on peut facilement majorer I_2 en utilisant l'inégalité de Cauchy-Schwartz, puisque :

$$I_2 \leq \left[\mathbb{E}_{m^*}\left((||\mu^*_{(m)}-\widehat{Y}_{(m)}||^2)^2\right)\right]^{1/2}\ [p_n(m)]^{1/2},$$

d'après la définition de $p_n(m)$. Or, on a vu que $||\mu^*_{(m)}-\widehat{Y}_{(m)}||^2 \overset{\mathcal{L}}{\sim} \sigma^2_*\chi^2(|m|)$, donc en utilisant la décomposition d'un χ^2 en somme de carrés de gaussiennes, $\mathbb{E}_{m^*}\left((||\mu^*_{(m)}-\widehat{Y}_{(m)}||^4\right)=(|m|^2+2|m|)\sigma^2_*$. on en déduit donc que comme $p_n(m)\to 0$ quand $n\to\infty$, alors $I_2\to 0$ quand $n\to\infty$.

ii/ En utilisant ce qui précède, dans le cas où $m\in\mathcal{M}_2$, alors on a directement $I_1=0$ car alors $\mu^*_{(m)}=\mu^*$. En utilisant le même raisonnement pour I_2, on voit bien qu'il suffit d'avoir la condition $p_n(m)\to 0$ pour obtenir que $R_n(m)\to 0$. ■

Nous allons montrer maintenant le résultat le plus important de ce chapitre, à savoir le fait qu'asymptotiquement on n'a pas tendance à sous-ajuster, c'est-à-dire que le modèle sélectionné pourra être plus "gros" (au sens de l'inclusion) que le vrai modèle, mais pas plus petit.

Théorème 0.1 *i. Pour tout modèle $m\in\mathcal{M}_1$, alors $\forall h>0$,*

$$\lim_{n\to\infty} n^h.p_n(m)=0.$$

ii. Pour tout modèle $m\in\mathcal{M}_2$, alors

$$\lim_{n\to\infty} p_n(m)=p(m)=P\left(\xi^{(a)}(|m^*|-|m|)\geq \xi^{(a)}(|m^*|-|l|)\ \forall l\in\mathcal{M}_2\cup\{m^*\}\right).$$

<u>*Démonstration*</u> : Nous allons choisir de travailler avec le critère AIC mais comme son comportement asymptotique est le même que celui du Cp de Mallows, on en déduira les résultats pour ce dernier. Comme de nombreuses étapes ont déjà été effectuées au cours des preuves précédentes, nous ne donnons que les lignes principales de la démonstration.

i. Soit $m\in\mathcal{M}_1$. Alors :

$$\begin{aligned}p_n(m) &\leq P\left(AIC(m)\leq AIC(m^*)\right)\\ &\leq P\left(n(\widehat{\sigma}^2_{(m)}-\widehat{\sigma}^2_{(m^*)})\leq n\left[\exp\left(\frac{a(|m^*|-|m|)}{n}\right)-1\right]\widehat{\sigma}^2_{(m^*)}\right)\\ &\leq P\Big(2\left((Y-\widehat{Y}_{(m)})-(\mu^*-\mu^*_{(m)}))'.(\mu^*-\mu^*_{(m)})\right)+\left(||\widehat{Y}_{(m^*)}-\mu^*||^2-||\widehat{Y}_{(m)}-\mu^*_{(m)}||^2\right)\\ &\qquad\qquad +||\mu^*-\mu^*_{(m)}||^2\leq b_n\,\widehat{\sigma}^2_{(m^*)}\Big)\\ &\leq P\left(X_n+Y_n+c_n\leq b_n\,\widehat{\sigma}^2_{(m^*)}\right),\end{aligned}$$

où

- le réel b_n est tel que $b_n = n\left[\exp\left(\frac{a(|m^*|-|m|)}{n}\right)-1\right] \underset{n\to+\infty}{\longrightarrow} b = a(|m^*|-|m|)$;
- la variable aléatoire $X_n = 2\left((Y-\widehat{Y}_{(m)})-(\mu^*-\mu^*_{(m)}))'.(\mu^*-\mu^*_{(m)})\right)$;
- la variable aléatoire $Y_n = \left(||\widehat{Y}_{(m^*)}-\mu^*||^2 - ||\widehat{Y}_{(m)}-\mu^*_{(m)}||^2\right)$;
- et le réel $c_n = ||\mu^*-\mu^*_{(m)}||^2$.

La variable aléatoire X_n est clairement une variable gaussienne car, avec les notations précédentes, $X_n = 2.\left(Q_{(m^\perp)}.\varepsilon^*\right)'.Q_{(m^\perp)}.\mu^* = 2.\left(Q_{(m^\perp)}.\mu^*\right)'.Q_{(m^\perp)}.\varepsilon^* \overset{\mathcal{L}}{\sim} \mathcal{N}(0,\gamma_n^2)$. On peut déterminer le comportement asymptotique de $\gamma_n^2 = 4\sigma_*^2.(\mu^*)'.Q_{(m^\perp)}.\mu^*$, puisque l'on a vu qu'avec l'Hypothèse 1,

$$\frac{1}{n}\gamma_n^2 \underset{n\to+\infty}{\longrightarrow} \gamma^2, \quad \text{où} \quad \gamma > 0.$$

Maintenant, on a bien-sûr $\frac{X_n}{\gamma_n} \overset{\mathcal{L}}{\sim} \mathcal{N}(0,1)$ et on peut écrire que

$$\begin{array}{rcl} p_n & \leq & P\left(\frac{X_n}{\gamma_n} \leq -\frac{c_n}{\gamma_n} - \frac{Y_n}{\gamma_n} + \frac{b_n}{\gamma_n}\widehat{\sigma}^2_{(m^*)}\right) \\ \text{ou encore } p_n & \leq & P\left(\frac{X_n}{\gamma_n} \leq \left[2n^{1/4}-\frac{c_n}{\gamma_n}\right] + \left[-n^{1/4}-\frac{Y_n}{\gamma_n}\right] + \left[\frac{b_n}{\gamma_n}\widehat{\sigma}^2_{(m^*)} - n^{1/4}\right]\right). \end{array}$$

Or, pour des événements quelconques F, G et H, on a

$$P(F) \leq P(F\cap G\cap H) + P(G^c) + P(H^c)$$

puisque $P(F) = P(F\cap G\cap H)+P(F\cap G\cap H^c)+P(F\cap G^c)$. Donc en considérant les événements :

- $F =$ " $\frac{X_n}{\gamma_n} \leq \left[2n^{1/4}-\frac{c_n}{\gamma_n}\right] + \left[-n^{1/4}-\frac{Y_n}{\gamma_n}\right] + \left[\frac{b_n}{\gamma_n}\widehat{\sigma}^2_{(m^*)} - n^{1/4}\right]$ " ;
- $G =$ " $\left[-n^{1/4}-\frac{Y_n}{\gamma_n} \leq 0\right]$ " ;
- $H =$ " $\left[\frac{b_n}{\gamma_n}\widehat{\sigma}^2_{(m^*)} - n^{1/4} \leq 0\right]$ " ;

on obtient que

$$p_n \leq P\left(\frac{X_n}{\gamma_n} \leq 2n^{1/4}-\frac{c_n}{\gamma_n}\right) + P\left(-n^{1/4}-\frac{Y_n}{\gamma_n} \geq 0\right) + P\left(\frac{b_n}{\gamma_n}\widehat{\sigma}^2_{(m^*)} - n^{1/4} \geq 0\right).$$

Il ne reste plus maintenant qu'à étudier le comportement asymptotique de chacune de ces trois probabilités. En premier lieu, et toujours d'après l'Hypothèse 1, $\frac{1}{n}||\mu^*-\mu^*_{(m)}||^2 \underset{n\to+\infty}{\longrightarrow} c > 0$, donc $\frac{c_n}{\gamma_n} \underset{n\to+\infty}{\sim} \frac{c}{\gamma}n^{1/2}$ et ainsi $2n^{1/4}-\frac{c_n}{\gamma_n} \underset{n\to+\infty}{\sim} -\frac{c}{\gamma}n^{1/2}$. Comme $\frac{X_n}{\gamma_n}$ suit une loi gaussienne centrée réduite, en appliquant l'inégalité de Bienaymé-Tchébitchev,

$$\begin{array}{rcl} P\left(\frac{X_n}{\gamma_n} \leq 2n^{1/4}-\frac{c_n}{\gamma_n}\right) & = & P\left(\exp\left(\frac{X_n}{\gamma_n}\right) \leq \exp\left(2n^{1/4}-\frac{c_n}{\gamma_n}\right)\right) \\ & = & o\left(e^{-n^{1/4}}\right), \end{array}$$

l'espérance de l'exponentielle d'une loi gaussienne centrée réduite étant une constante.
Quand à la seconde probabilité, de la même manière, on a :

$$\begin{aligned} P\left(\frac{Y_n}{\gamma_n} \leq -n^{1/4}\right) &= P\left(\exp\left(-\frac{Y_n}{\gamma_n}\right) \geq \exp\left(n^{1/4}\right)\right) \\ &\leq \mathbb{E}\left[\exp\left(-\frac{Y_n}{\gamma_n}\right)\right] e^{-n^{1/4}}. \end{aligned}$$

Mais $Y_n = \left(Q_{(m^\perp)}.\varepsilon^*\right)'.Q_{(m^\perp)}.\varepsilon^*$ donc

$$\begin{aligned} \mathbb{E}\left[\exp\left(-\frac{Y_n}{\gamma_n}\right)\right] &= \frac{1}{(2\pi.\sigma_*^2)^{n/2}} \int_{\mathbb{R}^n} \exp\left(-\frac{1}{2}\frac{x'.x}{\sigma_*^2} - \gamma_n^{-1}.\left(Q_{(m^\perp)}.x\right)'.Q_{(m^\perp)}.x\right) dx \\ &= \frac{1}{(2\pi)^{n/2}} \int_{\mathbb{R}^n} \exp\left(-\frac{1}{2} x'.\left(I_n + 2\,\sigma_*^2.\gamma_n^{-1}.Q_{(m^\perp)}\right).x\right) dx \\ &= \left|I_n + 2\,\sigma_*^2.\gamma_n^{-1}.Q_{(m^\perp)}\right|^{-1/2} = \left|I_n + 2\,\sigma_*^2.\gamma_n^{-1}.J_{(m^\perp)}\right|^{-1/2}, \end{aligned}$$

où $J_{(m^\perp)}$ est la matrice identité à laquelle on a enlevé $|m|$ unités sur la diagonale (c'est la matrice $Q_{(m^\perp)}$ diagonalisée). Comme $\gamma_n \underset{n\to+\infty}{\longrightarrow} 0$, on en déduit que $\mathbb{E}\left[\exp\left(-\frac{Y_n}{\gamma_n}\right)\right] \underset{n\to+\infty}{\longrightarrow} 1$ et donc

$$P\left(\frac{Y_n}{\gamma_n} \leq -n^{1/4}\right) = \mathcal{O}(e^{-n^{1/4}}).$$

Enfin, $\widehat{\sigma}^2_{(m^*)} \underset{n\to+\infty}{\overset{\mathcal{L}}{\longrightarrow}} 1$ donc toujours en utilisant l'inégalité de Bienaymé-Tchébitchev,

$$P\left(\frac{b_n}{\gamma_n}\,\widehat{\sigma}^2_{(m^*)} - n^{1/4} \geq 0\right) \leq \mathbb{E}\left[e^{\widehat{\sigma}^2_{(m^*)}}\right].e^{-b_n.\gamma_n.n^{1/4}} = o(e^{-n^{1/4}}).$$

On vient bien de montrer que $p_n(m)$ décroˆplus vite que toute fonction puissance quand $n \to \infty$ et quand $m \in \mathcal{M}_1$.

ii. Soit $m \in \mathcal{M}_2 \cup \{m^*\}$. Alors :

$$\begin{aligned} p_n(m) &= P(\widehat{m} = m) \\ &\simeq P\left(AIC(m) \leq AIC(m') \text{ pour tout } m' \in \mathcal{M}_2 \cup \{m^*\}\right) \\ &\simeq P\left(AIC(m^*) - AIC(m) \geq AIC(m^*) - AIC(m') \text{ pour tout } m' \in \mathcal{M}_2 \cup \{m^*\}\right) \\ &\underset{n\to+\infty}{\longrightarrow} P\left(\xi^{(a)}(|m| - |m^*|) \geq \xi^{(a)}(|m'| - |m^*|) \text{ pour tout } m' \in \mathcal{M}_2 \cup \{m^*\}\right) \end{aligned}$$

ce qui achève la preuve. ■

Conséquence : Sous l'Hypothèse 1, le critère de sélection de modèle (AIC ou Cp) ne sous-ajuste pas asymptotiquement mais sur-ajuste. De plus, le risque quadratique obtenu à partir du modèle sélectionné est tel que

$$R_n \underset{n\to+\infty}{\longrightarrow} \sum_{m\in\mathcal{M}_2\cup\{m^*\}} R_n(m) \geq \mathbb{E}\left(||\mu^* - \widehat{Y}_{m^*}||^2\right);$$

on n'a pas pu obtenir asymptotiquement qu'un risque quadratique supérieur à celui que l'on aurait obtenu en ayant la connaissance a priori du vrai modèle.

2 Sélection de modèle en régression linéaire fonctionnelle gaussienne

2.1 Présentation

On suppose maintenant que le vrai modèle est toujours

$$Y = \mu^* + \varepsilon^*, \quad \text{où} \quad \varepsilon^* \sim \mathcal{N}(0, \sigma_*^2 I_n), \tag{4}$$

mais cette fois-ci

$$\mu_i^* = s^*(x_i) \quad \text{où } s^* \text{ est une fonction de } I \text{ dans } \mathbb{R},$$

avec $(x_1, \cdots, x_n) \in I^n$ connus. En général, on supposera que I est un intervalle de $\mathbb{R}$, et en particulier $I = [0, 1]$. On suppose également que $s^* \in \mathcal{H}$, où $\mathcal{H}$ est un espace fonctionnel connu, possédant une base orthonormée dénombrable $(\phi_j)_{j\in\mathbb{N}^*}$ et ainsi il existe une famille unique de réels $\beta = (\beta_j^*)_{j\in\mathbb{N}^*}$ telle que :

$$s^* = \sum_{j=1}^{\infty} \beta_j^*.\phi_j.$$

Remarque : Une grande partie du travail suivant dépendra du choix de l'espace fonctionnel $\mathcal{H}$. Un exemple souvent rencontré est celui d'un espace de Hilbert. En revanche, si $\mathcal{H}$ est de dimension finie, on est ramené au cas paramétrique, puisqu'alors

$$s^*(x_i) = \sum_{j=1}^{p} \beta_j^*.\phi_j(x_i) = \sum_{j=1}^{p} \beta_j^*.x_{ij}.$$

et on retrouve la matrice $X^{(m_p)}$ précédente.

On supposera donc que la famille de modèles utilisée est

$$Y = \mu + \varepsilon, \quad \text{où} \quad \mu = (s(x_i))_{1\le i\le n} \quad \text{et} \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I_n), \tag{5}$$

où $s \in \mathcal{H}$ est une fonction inconnue que l'on cherche à estimer. Pour ce faire, on utilisera sa décomposition dans la base $(\phi_j)_{j\in\mathbb{N}^*}$, ce qui revient à estimer la famille $\beta^* = (\beta_j^*)_{j\in\mathbb{N}^*}$. Cependant, comme l'on ne dispose que d'un nombre n d'observations, on ne peut au mieux estimer qu'un nombre inférieur à n de paramètres. On considérera donc à nouveau une famille finie de sous-ensembles de $\mathbb{N}^*$ dont le cardinal de chacun est inférieur à $\mathrm{rg}(X)$ avec $X = (\phi_j(x_i))_{1\le i\le n, j\in\mathbb{N}^*}$, soit

$$\mathcal{M}_n \subset \{m \subset \mathbb{N}^*, \ |m| \le \mathrm{rg}(X)\} \quad \text{et} \quad |\mathcal{M}_n| < \infty.$$

On estimera donc β par $\widehat{\beta}(m)$ avec $m \in \mathcal{M}_n$, où comme dans le cas paramétrique

$$\widehat{\beta}(m) = ((X^{(m)})'.X^{(m)})^{-1}(X^{(m)})'.Y$$

La différence avec le cas paramétrique est qu'en général on ne peut estimer le bon ensemble m, puisque celui-ci en général ne sera pas de cardinal fini. On ne pourra donc au mieux qu'estimer une fonction approchant la fonction s^*, et qui sera :

$$\widehat{s} = \sum_{j=1}^{|m|} (\widehat{\beta}(m))_j.\phi_{m_j} \quad \text{où} \quad m = \{m_j, \ j = 1, \cdots, |m|\}.$$

2.2 Quelques résultats

Cas des décompositions ℓ^2

On s'intéresse ici aux résultats proposés par Shibata (1981). Par ailleurs, on passera sans repréciser de l'écriture de β sous la forme $(\beta(m))_{1\leq j\leq |m|}$ à celle sous la forme $(\beta_j)_{j\in\mathbb{N}^*}$ suivant le contexte. On commence par supposer que :

Hypothèse ℓ^2 : Les suites $(x_{ij})_{j\in\mathbb{N}^*} = (\phi_j(x_i))_{j\in\mathbb{N}^*}$ appartiennent à ℓ^2 pour tout $i = 1,\cdots,n$, soit $\sum_{j\in\mathbb{N}^*} \phi_j(x_i)^2 < \infty$.

On définit alors la norme $||.||_n$ telle que pour toute suite $\beta = (\beta)_{n\in\mathbb{N}^*} \in \ell^2$:

$$||\beta||_n = \sum_{i=1}^{n} \left(\sum_{j=1}^{\infty} x_{ij}.\beta_j \right)^2 .$$

Cette norme généralise en quelque sorte la norme matricielle précédemment définie. On définit alors le risque quadratique pour $m \in \mathcal{M}_n$:

$$R_n(m) = \mathbb{E}\left(||\widehat{\beta}(m) - \beta^*||_n^2 \right) = \mathbb{E}\left(||\beta^* - \beta_n^{(m)}||_n^2 \right) + |m|.\sigma^2,$$

avec $\beta_n^{(m)} = \text{Argmin}_{\beta\in<X^{(m)}>} ||\beta^* - \beta||_n^2$. On définit alors le "meilleur modèle possible" pour un n fixé par la formule :

$$m_n^* = \text{Argmin}_{m\in\mathcal{M}_n}\ R_n(m).$$

Ce modèle est là-encore inconnaissable, mais on essaye de l'estimer par un critère de type Cp de Mallows, soit pour $m \in \mathcal{M}_n$:

$$S_n(m) = \frac{1}{n}\, ||Y - \widehat{\beta}(m)||^2\, (n + 2|m|)$$

et ainsi on estime le meilleur modèle par :

$$\widehat{m}_n = \text{Argmin}_{m\in\mathcal{M}_n}\ S_n(m).$$

On en arrive alors au théorème suivant :

Théorème : Sous les notations et hypothèses précédentes, et si :

i. $\text{rg}\left((X^{(m)})'.(X^{(m)})\right) = |m|$ pour tout $m \in \mathcal{M}_n$;

ii. $\max_{m\in\mathcal{M}_n} |m| = o(n)$;

iii. $\forall\delta \in [0,1[$, alors $\sum_{m\in\mathcal{M}_n} \delta^{R_n(m)} \underset{n\to+\infty}{\longrightarrow} 0$;

$$\text{alors} \quad \lim_{n\to\infty} \frac{||\widehat{\beta}(\widehat{m}_n) - \beta^*||_n^2}{R_n(m_n^*)} = \lim_{n\to\infty} \frac{||\widehat{\beta}(\widehat{m}_n) - \beta^*||_n^2}{\inf_{m\in\mathcal{M}_n} ||\widehat{\beta}(m) - \beta^*||_n^2} = 1.$$

Ce théorème nous indique donc qu'asymptotiquement on sélectionne un modèle minimisant le risque quadratique pour la famille de modèles choisie. La condition iii. du théorème reste cependant un peu obscure ; sur les exemples suivants, on voit qu'elle peut être facilement vérifiée :

- Si $I = [0,1[$, $x_i = \dfrac{i-1}{n}$ pour $i = 1, \cdots, n$ et $f(x) = \sum_{k=1}^{\infty} \beta_k.x^k$, avec la famille $\mathcal{M}_n = \{\{1, \cdots, k\}\}_{1 \le k \le k_n}$ où $k_n = o(n)$ alors le Théorème est vérifié dès que β^* admet une infinité de coordonnées non nulles. Par exemple, pour $f(x) = \log(1+x)$, pour $f(x) = \exp x$,..., mais pas pour f polynôme !
- Si $I =]-1,1[$, $x_i = \dfrac{i}{n}$ pour $i = -(n-1), \cdots, (n-1)$ et avec $\mathcal{H}$ est l'ensemble des fonctions de I dans $\mathbb{R}$ qui sont $\mathcal{C}^2$. On utilise la base suivante de $\mathcal{H}$:

$$\phi_{2k}(x) = \frac{\sin(\pi.k.x)}{k+1}, \quad \phi_{2k+1}(x) = \frac{\cos(\pi.k.x)}{k+1}$$

  (la partie fractionnelle découlant de l'intégration d'une série de Fourier pour une fonction $\mathcal{C}^1$). et le Théorème est vérifié dès que β^* admet une infinité de coordonnées non nulles.

Li (1987) et Kneib (1994) ont améliorer ces résultats en passant au cas non gaussien.

Cas du développement dans $\mathbb{L}^2([0,1)])$

On reprend ici les résultats proposés par Polyak et Tsybakov (1990). On se place dans le même cadre que précédemment, mais en fixant $x_i = \dfrac{i}{n}$ pour $i = 1, \cdots, n$ et la base $(\phi_j)_j$ que l'on pose telle que :

$$\phi_{2k}(x) = \sqrt{2}.\sin(2\pi.k.x), \quad \phi_{2k+1}(x) = \sqrt{2}.\cos(2\pi.k.x), \text{ et } \quad \phi_1(x) = 1.$$

Hypothèse : On suppose que $s^*(x) = \sum_{j=1}^{\infty} c_j.\phi_j(x)$ avec $(c_j)_j$ une suite telle que pour tout $j \in \mathbb{N}^*$, $|c_j| \le d_j$ avec $(j.d_j)_j$ une suite décroissante et $\sum_{j=1}^{\infty} d_j < \infty$.

On définit $f_N^n = \sum_{j=1}^{N} c_j^n.\phi_j(x)$ avec $N \le n$, c'est-à-dire que l'on considère la famille de modèles $\mathcal{M}_n = \{\{1, \cdots, k\}\}_{1 \le k \le n}$. On va estimer cette fonction par

$$\widehat{f}_N^n(x) = \sum_{j=1}^{N} \widehat{c}_j^n.\phi_j(x) \quad \text{où} \quad \widehat{c}_j^n = \frac{1}{n}\sum_{i=1}^{n} y_i.\phi_j(x_i).$$

Enfin, on définit le modèle à partir d'un critère de type Cp de Mallows :

$$\widehat{N}(n) = \text{Argmin}_{1 \le N \le n} \left(\frac{1}{n} \sum_{i=1}^{n} (y_i - \widehat{f}_N^n(x_i))^2 + 2\frac{\sigma_n^2.N}{n} \right),$$

où σ_n^2 est un estimateur de σ^2. Si on définit le risque quadratique par $R_n(N) = \dfrac{1}{n} \sum_{i=1}^{n} (s(x_i) - \widehat{f}_N^n(x_i))^2$, on a le théorème suivant :

Théorème : Si s admet un développement avec une infinité de coefficients non nuls, alors :

$$\frac{R_n(\widehat{N}(n))}{\min_{1 \le N \le n} R_n(N)} \xrightarrow[n \to +\infty]{\mathcal{P}} 1.$$

2.3 Estimation adaptative

Le problème des estimations précédentes tient dans le fait que l'on pose une base a priori pour des fonctions appartenant à un espace fonctionnel fixé ; ces estimations sont donc totalement liée au choix de cet espace. De plus le cadre paramétrique et non-paramétrique sont distincts et les résultats donnés ne sont qu'asymptotiques. Nous allons présenter une méthodologie permettant de dépasser ses limitations et d'intégrer l'ensemble de ces problèmes dans un cadre unique.
Cependant, avant d'évoquer en détail ces méthodes, nous allons voir quelques rappels et définitions qui nous seront utiles.

Rappels

Produit scalaire : Soit X un espace vectoriel réel ou complexe. Cet espace X est dit pré-hilbertien s'il existe une forme bilinéaire $(.,.)$ appelée produit scalaire et vérifiant les prorpiétés suivantes :

i. $\forall(x,y) \in X^2$, $\forall\alpha \in \mathbb{C}$, alors $(\alpha.x, y) = \alpha(x,y)$;
ii. $\forall(x,y) \in X^2$, $(x,y) = (y,\overline{x})$;
iii. $\forall(x,y,z) \in X^3$, $(x+y,z) = (x,z)+(y,z)$;
iv. et $\forall x \in X$ alors $(x,x) \geq 0$; $(x,x) = 0 \Longleftrightarrow x = 0$.

Espace de Hilbert : Un espace préhilbertien complet (c'est-à-dire tel que toute suite de Cauchy converge) est appelé un espace de Hilbert.

Exemples : • l'espace $\mathbb{L}^2(S, \mathcal{B}(S), m)$ où S est un espace topologique, avec le produit scalaire

$$(f,g) = \int_S f.\overline{g}.dm.$$

• l'espace $\ell^2(\mathbb{R})$ pour les suites de nombres réels muni du produit scalaire

$$((x_n)_n, (y_n)_n) = \sum_{n=1}^{\infty} x_n.y_n.$$

Définition : Soit $(\Omega, \mathcal{A}, P)$ un espace de probabilité, H_1 un espace de Hilbert et $Y : \Omega \to H_1$ une fonction aléatoire telle que $Y = Y_s$ où $s \in \mathcal{H}$ un espace de Hilbert, de norme $||.||$. Alors, pour $\mathcal{H}_\alpha$ un sous-espace connu de $\mathcal{H}$, le risque quadratique minimax d'estimation de s sur $\mathcal{H}_\alpha$ est

$$R(\mathcal{H}_\alpha) = \inf_{\widehat{s}=f(Y_s)} \sup_{s\in\mathcal{H}_\alpha} \mathbb{E}\left(||\widehat{s}-s||^2\right), \quad \text{où la fonction } f \text{ est mesurable.}$$

Remarque : On peut écrire que si $\widehat{s} \in \mathcal{H}_\alpha$, $\mathbb{E}\left(||\widehat{s}-s||^2\right) = ||s_{\mathcal{H}_\alpha} - s||^2 + \mathbb{E}\left(||\widehat{s}-s_{\mathcal{H}_\alpha}||^2\right)$, d'où

$$R(\mathcal{H}_\alpha) \leq \sup_{s\in\mathcal{H}_\alpha} \left(\mathbb{E}\left(||\widehat{s}-s||^2\right) + ||s_{\mathcal{H}_\alpha} - s||^2\right).$$

Le risque minimax n'est pas exprimable en général, mais il peut être majoré ou minoré.

Exemple : On se place dans le cadre du problème de régression fonctionnelle,

$$y_i = s(x_i) + \varepsilon_i \quad \text{pour } i = 1, \cdots, n, \quad \text{avec } x_i = \frac{i}{n},$$

et avec les ε_i qui sont des variables gaussiennes centrées indépendantes de même variance σ^2. Pour une fonction s, on note $||.||_n$ la norme telle que :

$$||s||_n = \frac{1}{n}\sum_{i=1}^{n} s^2(x_i).$$

Enfin, on définit l'espace des fonctions holderiennes

$$\mathcal{H}_\alpha(L) = \left\{s : [0,1] \leftarrow \mathbb{R} \mid |s(x) - s(y)| \leq L.|x-y|^\alpha,\ \forall (x,y) \in [0,1]^2\right\},$$

avec $\alpha > 0$ et $L > 0$. On peut, après un travail certain..., montrer que le risque quadratique minimax pour cet espace fonctionnel s'écrit :

$$R_n(\mathcal{H}_\alpha(L)) = C.\left(\left(\frac{\sigma^2}{n}\right)^\alpha .L\right)^{\frac{2}{1+\alpha}} \quad \text{où} \quad C > 0.$$

Deux attitudes sont possibles pour estimer la fonction s^*.

i. On peut directement faire des estimations de s^* à partir du sous-espace $\mathcal{H}_\alpha(L)$. Ainsi, on note :

$$\tilde{s}_n = \text{Argmin}_{t\in\mathcal{H}_\alpha(L)} \sum_{i=1}^{n}(y_i - t(x_i))^2,$$

et ainsi $\tilde{s}_n$ dépend de α et de L. On montre alors (Van de Geer, 1990) que :

$$\mathbb{E}(||s^* - \tilde{s}_n||_n^2) \sim C'.R_n(\mathcal{H}_\alpha(L)) \quad \text{où} \quad C' > 0.$$

ii. On peut également estimer s sur un espace plus grand que $\mathcal{H}_\alpha(L)$. Ainsi, en travaillant sur $\mathbb{L}^2([0,1],\mu_n)$ avec $\mu_n = \frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}$, on peut estimer s sur S_m sous-espace de $\mathbb{L}^2([0,1],\mu_n)$ défini pour $m \in \mathbb{N}^*$ par $S_m = \left\langle \left\{\mathbb{I}_{\left[\frac{j-1}{m},\frac{j}{m}\right]}\right\}_{1\leq j\leq m}\right\rangle$. Ainsi, pour $m \in \mathbb{N}^*$, soit :

$$\widehat{s}_m = \text{Argmin}_{t\in S_m} \sum_{i=1}^{n}(y_i - t(x_i))^2 .$$

$$\begin{aligned} \text{Alors}\ \ \mathbb{E}(||s^* - \widehat{s}_m||_n^2) &= ||s^* - s_m^*||_n^2 + \mathbb{E}(||\widehat{s}_m - s_m^*||_n^2) \\ &= ||s^* - s_m^*||_n^2 + \frac{m}{n}\sigma^2. \end{aligned}$$

Ainsi, lorsque s^* appartient à $\mathcal{H}_\alpha(L)$, alors $||s^* - s_m^*||_n^2 \leq \frac{L^2}{m^{2\alpha}}$. Donc, si on minimalise le risque quadratique précédent, soit maintenant $\frac{m}{n}\sigma^2 + \frac{L^2}{m^{2\alpha}}$, on obtient l'estimation suivante de m :

$$\widehat{m}_{\alpha,L} = \left(\left(\frac{L^2}{\sigma^2}n\right)^{\frac{1}{1+2\alpha}}\right).$$

En utilisant cette estimation, on obtient encore :

$$\mathbb{E}(||s^* - \widehat{s}_{\widehat{m}_{\alpha,L}}||_n^2) \sim C''.R_n(\mathcal{H}_\alpha(L)) \quad \text{où} \quad C'' > 0.$$

Définition : On dira ainsi qu'un estimateur $\widehat{s}$ est adaptatif dans le sens du risque quadratique minimax si $\widehat{s}$ ne dépend pas de θ, paramètre d'un espace fonctionnel $\mathcal{H}_\theta$, et si
$\mathbb{E}(||s^* - \widehat{s}||^2) \leq C.R(\mathcal{H}_\alpha)$ où $C > 0$.

Bibliographie

Ne sont présentées ici que les références supplémentaires par rapport à celles spécifiées à la fin du polycopié de Ph. Besse. Les références avec des $*$ sont des articles, leur nombre correspondant à la difficulté (estimée...) de l'article. Tout est normalement présent à la bibliothèque de maths.

- [1*] Akaike, H. *Information theory and an extension of the maximum likelihood principle.* Second International Symposium on Information Theory (Tsahkadsor, 1971), pp. 267–281. Akadémiai Kiadó, Budapest, 1973.

- [2***] Baraud, Yannick. *Model selection for regression on a fixed design.* Probab. Theory Related Fields 117 (2000), no. 4, 467–493.

- [3***] Barron, Andrew ; Birgé, Lucien ; Massart, Pascal. *Risk bounds for model selection via penalization.* Probab. Theory Related Fields 113 (1999), no. 3, 301–413.

- [4***] Birgé, Lucien ; Massart, Pascal. *Gaussian model selection.* J. Eur. Math. Soc. (JEMS) 3 (2001), no. 3, 203–268.

- [5] McQuarrie, Allan D. R. ; Tsai, Chih-Ling. Regression and time series model selection. World Scientific Publishing Co., Inc., River Edge, NJ, 1998

- [6*] Mallows, Colin L. *More comments on C_p.* Technometrics 37 (1995), no. 4, 362–372.

- [7] Montgomery, Douglas C. ; Peck, Elizabeth A. Introduction to linear regression analysis. Second edition. Wiley Series in Probability and Mathematical Statistics : Applied Probability and Statistics. A Wiley-Interscience Publication. John Wiley & Sons, Inc., New York, 1992.

- [8**] Nishii, Ryuei. *Asymptotic properties of criteria for selection of variables in multiple regression.* Ann. Statist. 12 (1984), no. 2, 758–765.

- [9] Rencher, Alvin C. Linear models in statistics. Wiley Series in Probability and Statistics. A Wiley-Interscience Publication. John Wiley & Sons, Inc., New York, 2000.

- [10**] Shibata, Ritei. *An optimal selection of regression variables.* Biometrika 68 (1981), no. 1, 45–54.

- [11] Wetherill, G. Barrie ; Duncombe, P. ; Kenward, M. ; Köllerström, J. ; Paul, S. R. ; Vowden, B. J. Regression analysis with applications. Monographs on Statistics and Applied Probability. Chapman & Hall, London, 1986.