Université de Montréal

# Méthodes de prévision en régression linéaire multivariée

par

Ana GUEORGUIEVA

Département de mathématiques et de statistique

Faculté des arts et des sciences

Mémoire présenté à la Faculté des études supérieures

en vue de l'obtention du grade de

Maître ès sciences (M.Sc.)
en Statistique

avril 2006

# Université de Montréal

Faculté des études supérieures

Ce mémoire intitulé

## Méthodes de prévision en régression linéaire multivariée

présenté par

## Ana GUEORGUIEVA

a été évalué par un jury composé des personnes suivantes :

*Christian LÉGER*
(président-rapporteur)

*Martin BILODEAU*
(directeur de recherche)

*Yves LEPAGE*
(membre du jury)

Mémoire accepté le:
05/04/06

# SOMMAIRE

Nous étudions le problème de prévoir plusieurs variables dépendantes à partir d'un même ensemble des variables indépendantes dans le modèle de la régression linéaire multivariée. Nous considérons une famille d'estimateurs, appelés estimateurs à rétrécissement, qui nous permettent de prévoir les variables dépendantes par des combinaisons linéaires des prévisions obtenues par la méthode des moindres carrés.

Notre étude est motivée par les travaux suivants :

[3] où Breiman et Friedman proposent deux estimateurs du type "Curds and Whey" et les comparent aux autres estimateurs existants ;

[1] où Bilodeau propose des modifications aux estimateurs considérés dans [3] ;

[10] où Srivastava et Solanky proposent deux estimateurs minimax KS et SS et les comparent aux autres estimateurs existants.

Nous considérons aussi l'estimateur "full cross validation" introduit par Breiman et Friedman [3]. Nous le calculons de façon différente en suivant la méthode proposée par P. D. Sasieni dans les discussions sur le même article.

Tous ces estimateurs sont comparés dans deux simulations (avec des données générées de la loi normale et de la loi $t$ de Student) à l'aide des deux critères définis dans [3] et [10].

**MOTS–CLÉS** : régression linéaire multivariée, prévisions, estimateur à rétrécissement

# SUMMARY

We study the problem of predicting several response variables from the same set of explanatory variables in the model of multivariate linear regression. We consider a group of estimators, known as shrinkage estimators, where the dependent variables are predicted as linear combinations of the ordinary least squares predictions.

This study is motivated by the following articles :

[3] where Breiman and Friedman propose two estimators of type "Curds and Whey" and compare them to the existing estimators;

[1] where Bilodeau proposes modifications of the estimators in [3];

[10] where Srivastava and Solanky propose two minimax estimators and compare them to the existing estimators.

We also consider the "full cross validation" estimator, proposed by Breiman and Friedman [3] and compute it by using a method suggested by P.D. Sasieni in the discussion section of the same paper.

All estimators are compared in two simulations (data generated from normal and Student's $t$ distributions) using the two criteria defined in [3] and [10].

**KEY WORDS :** multivariate linear regression, predictions, shrinkage estimators

# TABLE DES MATIÈRES

# LISTE DES FIGURES

# LISTE DES TABLEAUX

# LES LOGICIELS

Nous avons utilisé les logiciels Splus (version 6.2), SPSS (version 11.5) et Fortran Power Station (version 4.0), dans l'environnement Windows XP Pro. Le Fortran Power Station (FPS 4.0) a été utilisé pour la programmation en Fortran 90.

Dans le chapitre 4, nous faisons deux simulations avec 168 000 répétitions en tout. Les nombreuses répétitions et la lourdeur des calculs (7 estimateurs et 12 critères pour chaque répétition) imposent l'utilisation d'un logiciel permettant une plus grande vitesse que celle des logiciels statistiques. Pour cette raison nous avons choisi Fortran 90 qui nous offre également plusieurs sous-routines de minimisation quadratique nécessaire pour calculer l'estimateur FCV (présenté au chapitre 3).

À toutes les étapes de la programmation des simulations, nous avons fait les essais nécessaires pour s'assurer que les résultats obtenus avec Fortran 90 sont identiques aux résultats obtenus par Splus (utilisé comme base de comparaison). Par exemple, nous avons remarqué que dans Fortran 90, la sous-routine "CANCR" qui calcule la matrice de coefficients canoniques $C$ (présentée dans la section 1.3) nous donne des résultats différents que la fonction respective "cancor" de Splus. Il s'agit d'une différence de l'ordre $10^{-2}$. Nous avons alors choisi de calculer $C$ à l'aide de la méthode matricielle présentée dans la section 1.3, donnant une différence de l'ordre $10^{-10}$ entre les résultats des deux logiciels.

Les simulations ont été faites comme suit :

- en FPS 4.0 nous avons écrit et compilé des programmes qui génèrent les données (par les méthodes présentées dans les sections 4.1 et 4.2) et qui calculent les estimateurs $(K)$ (présentés dans les chapitres 2 et 3) ainsi que les critères $A_{(K)}$ et $W_{(K)}$ (présentés dans la section 4.3) ;

- l'exécution des programmes a été faite sur deux ordinateurs de configuration :

  (1) INTEL(R),

  Pentium(R) 4 CPU, 1.6 GHz ,

  256 MB RAM,

  20 GB IDE Harddisk

  pour les petites valeurs des paramètres ($p \leq 30$ ou $N \leq 300$) et

  (2) INTEL IA32 Architecture,

  Dual P4 XEON CPU, 3.2 GHz , 2MB L2 Cache,

  1 GB RAM,

  120 GB SATA / IDE Harddisk

  pour les grandes valeurs des paramètres ($p = 50$ ou $N \geq 400$) ;

- les sorties (les valeurs des critères) ont été traitées avec SPSS 11.5 pour obtenir des fichiers txt compatibles avec les logiciels statistiques et pour créer les boxplots dans l'annexe 1 ;
- nous avons utilisé Splus pour calculer les médianes et pour désigner les figures dans la section 4.4.

Les deux exemples dans le chapitre 5 ont été élaborés avec Splus. Dans Splus, il n'y a pas de fonction de minimisation quadratique nécessaire pour calculer l'estimateur FCV (voir le chapitre 3) mais la compatibilité entre Fortran 90 et Splus nous a permit de créer les fonctions nécessaires. En FPS 4.0 nous avons fait une programmathèque de type .dll (dynamic link library) contenant la sous-routine de minimisation. Ensuite, en ajoutant cette programmathèque comme un chapitre de Splus, nous avons créé la fonction nécessaire pour la minimisation quadratique.

# REMERCIEMENTS

Avant tout, je tiens à remercier mon directeur de recherche Monsieur Martin Bilodeau pour m'avoir proposé le sujet de ce mémoire, pour sa direction et pour le temps qu'il m'a consacré.

Merci à mon père et ma sœur qui m'ont aidé à programmer et réaliser les nombreuses simulations.

Merci à mon mari pour sa patience.

# Chapitre 1

# LA RÉGRESSION LINÉAIRE MULTIVARIÉE

## 1.1. Le modèle

Considérons le modèle de régression linéaire multivariée avec $p$ variables explicatives $\mathbf{x} = (x_1, \ldots, x_p)^\top$ et $q$ variables dépendantes $\mathbf{y} = (y_1, \ldots, y_q)^\top$, $q \leq p$. Les matrices des observations prennent la forme suivante :

$$X = \begin{pmatrix} \mathbf{x}_1^\top \\ \vdots \\ \mathbf{x}_N^\top \end{pmatrix} = (x_{ij}), \quad Y = \begin{pmatrix} \mathbf{y}_1^\top \\ \vdots \\ \mathbf{y}_N^\top \end{pmatrix} = (y_{ij}),$$

où $X$ (respectivement $Y$) est une matrice $N \times p$ (respectivement $N \times q$) dont la $n$-ième ligne correspond à la $n$-ième observation indépendante.

Les vecteurs $\mathbf{y}_n$, $n = 1, \ldots, N$, sont indépendants de moyenne conditionnelle $E(\mathbf{y}_n|\mathbf{x}_n) = A\mathbf{x}_n$ (où $A = (\mathbf{a}_1, \ldots, \mathbf{a}_q)^\top$ est la matrice $q \times p$ des coefficients inconnus) et de matrice de covariance conditionnelle $\Sigma$, supposée symétrique et définie positive. La forme matricielle du modèle s'écrit alors

$$Y = XA^\top + E,$$

ou de manière équivalente

$$\mathbf{y}_n = A\mathbf{x}_n + \boldsymbol{\varepsilon}_n, \quad n = 1, \ldots, N.$$

Ici, la matrice d'erreur $E = (\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_N)^\top$ est $N \times q$ et les termes d'erreur $\boldsymbol{\varepsilon}_n$ sont supposés indépendants de loi $N_q(0, \Sigma)$, $\Sigma$ est inconnue.

## 1.2. Les estimateurs

Soit $M = X(X^\top X)^{-1}X^\top$ la matrice de projection (symétrique et idempotente, $MM = M$) et $S = Y^\top(I_q - M)Y$, une matrice symétrique définie positive. L'estimateur sans biais de $\Sigma$ est donné par

$$\widehat{\Sigma} := \frac{1}{N-p}S.$$

Considérons la méthode des moindres carrés (notée par l'acronyme MC). L'estimateur MC des coefficients de la régression $A$ est défini par

$$\widehat{A} := Y^\top X(X^\top X)^{-1}.$$

Désignons par

$$\widehat{Y} := X\widehat{A}^\top = MY$$

les valeurs prévues par l'estimateur MC de $Y$. La prévision pour la $n$-ième observation est donc $\hat{\mathbf{y}}_n = \widehat{A}\mathbf{x}_n$. Posons enfin

$$\widehat{E} := Y - \widehat{Y},$$

où $\widehat{E}$ désigne la matrice des résidus.

En termes de la matrice $M$ nous avons les identités suivantes :

$$\begin{aligned} \widehat{Y}^\top\widehat{Y} &= (MY)^\top(MY) = Y^\top MMY = Y^\top MY, \\ \widehat{E} &= Y - \widehat{Y} = Y - X\widehat{A}^\top = Y - MY = (I_q - M)Y. \end{aligned} \tag{1}$$

Il en résulte que la matrice $S$ peut être réécrite comme

$$S = Y^\top(I_q - M)Y = \Big((I_q - M)Y\Big)^\top(I_q - M)Y = \widehat{E}^\top\widehat{E}.$$

## 1.3. Les matrices $C$ et $F$ et les corrélations canoniques

Nous allons trouver une matrice $q \times q$ non-singulière $C$ qui satisfait aux deux conditions suivantes :

$$C^\top SC = I_q \quad \text{et} \quad C^\top(Y^\top MY)C = F, \tag{2}$$

où $F = \operatorname{diag}(f_1, \ldots, f_q)$ est une matrice diagonale et $f_1 \geq f_2 \geq \cdots \geq f_q$ sont ordonnés.

Notons que les conditions données en (2) s'expriment de manière équivalente comme suit :

$$\begin{aligned} S &= C^{-\top}C^{-1}, \\ Y^\top MY &= C^{-\top}FC^{-1}, \end{aligned} \tag{3}$$

où la notation $C^{-\top}$ signifie $(C^{-1})^\top = (C^\top)^{-1}$.

Soit $S = H_S\Lambda_S H_S^\top$ la décomposition spectrale de $S$, où $H_S$ est une matrice orthogonale (i.e. $H_S^{-1} = H_S^\top$) et $\Lambda_S$ est une matrice diagonale dont les éléments sont dans un ordre décroissant. Posons

$$\begin{aligned} S^{-\frac{1}{2}} &= H_S\Lambda_S^{-\frac{1}{2}}H_S^\top, \\ S^{\frac{1}{2}} &= H_S\Lambda_S^{\frac{1}{2}}H_S^\top. \end{aligned}$$

Notons que les matrices $S^{-\frac{1}{2}}$ et $S^{\frac{1}{2}}$ sont symétriques, définies positives et satisfont aux identités $\left(S^{\frac{1}{2}}\right)^{-1} = S^{-\frac{1}{2}}$, $\left(S^{-\frac{1}{2}}\right)^{-1} = S^{\frac{1}{2}}$ et $S^{-\frac{1}{2}}SS^{-\frac{1}{2}} = I_q$.

Introduisons la matrice symétrique $G$ définie par

$$G := S^{-\frac{1}{2}}(Y^\top MY)S^{-\frac{1}{2}},$$

ou de manière équivalente

$$Y^\top MY = S^{\frac{1}{2}}GS^{\frac{1}{2}}. \tag{4}$$

Soit

$$G = H_G\Lambda_G H_G^\top$$

la décomposition spectrale de $G$, où $H_G$ est orthogonale et $\Lambda_G$ est diagonale dont les éléments sont ordonnés dans un ordre décroissant.

Nous allons vérifier que la matrice $C$ définie par

$$C := S^{-\frac{1}{2}}H_G \tag{5}$$

satisfait aux conditions (2). Ceci repose sur les identités suivantes (obtenues à l'aide de (4) et (5)) :

$$\begin{aligned} C^\top SC &= H_G^\top(S^{-\frac{1}{2}}SS^{-\frac{1}{2}})H_G = H_G^\top I_q H_G = I_q, \\ C^\top(Y^\top MY)C &= (H_G^\top S^{-\frac{1}{2}})(S^{\frac{1}{2}}GS^{\frac{1}{2}})(S^{-\frac{1}{2}}H_G) = \Lambda_G = F. \end{aligned}$$

Les corrélations canoniques échantillonnales s'obtiennent à l'aide de la diagonalisation de la matrice

$$\widehat{Q} = (Y^\top Y)^{-1} Y^\top X (X^\top X)^{-1} X^\top Y,$$

où $X$ et $Y$ sont centrées, i.e. chaque colonne est centrée à sa moyenne. En termes des matrices $C$ et $F$, cette matrice s'écrit

$$\begin{aligned}
\widehat{Q} &= (Y^\top Y)^{-1}(Y^\top M Y) = \left[(Y^\top M Y)^{-1}(S + Y^\top M Y)\right]^{-1} \\
&= \left[(Y^\top M Y)^{-1} S + I_q\right]^{-1} = \left[(C^{-\top} F C^{-1})^{-1}(C^{-\top} C^{-1}) + I_q\right]^{-1} \\
&= \left[(C F^{-1} C^\top)(C^{-\top} C^{-1}) + I_q\right]^{-1} = \left[C(F^{-1} + I_q)C^{-1}\right]^{-1}.
\end{aligned}$$

On obtient finalement

$$\widehat{Q} = C(F^{-1} + I_q)^{-1} C^{-1} = C C^2_{corr} C^{-1}. \tag{6}$$

Alors $C$ est en effet la matrice des coefficients canoniques et la relation entre $F$ et les corrélations canoniques est donnée par

$$C^2_{corr} = \operatorname{diag}(c_1^2, c_2^2, \ldots, c_q^2) = (F^{-1} + I_q)^{-1} = \operatorname{diag}\Big(\frac{f_1}{f_1 + 1}, \ldots, \frac{f_q}{f_q + 1}\Big).$$

Comme les éléments $f_i$ sont ordonnés de manière que $f_1 \geq f_2 \geq \cdots \geq f_q$, nous avons aussi $c_1^2 \geq c_2^2 \geq \cdots \geq c_q^2$.

# Chapitre 2

## LES ESTIMATEURS À RÉTRÉCISSEMENT

Rappelons que les vecteurs-lignes $\widehat{\mathbf{a}}_j$ de l'estimateur des moindres carrés $\widehat{A} = Y^\top X(X^\top X)^{-1}$ sont donnés par la formule $\widehat{\mathbf{a}}_j = (X^\top X)^{-1}X^\top \omega_j$, où le vecteur $\omega_j$ désigne la $j$-ième colonne de $Y$. La matrice $\widehat{A}$ est obtenue en faisant la régression MC pour chacune des variables $\omega_j$ séparément, et donc sans tenir compte des corrélations. Notons que les prévisions MC, $\widehat{Y} = X\widehat{A}^\top$, ne dépendent pas de $\Sigma$.

Breiman et Friedman [**3**] proposent de faire les prévisions en utilisant une combinaison linéaire de toutes les prévisions MC des $q$ variables comme suit :

$$\widetilde{\mathbf{y}}_n = B\widehat{\mathbf{y}}_n = B\widehat{A}\mathbf{x}_n,$$
$$\widetilde{Y} = X(B\widehat{A})^\top = X\widetilde{A}^\top,$$

où la matrice $B$, $q \times q$ est appelée *matrice de rétrécissement* et $\widetilde{A} = B\widehat{A}$, une matrice $q \times p$ est l'estimateur à rétrécissement des coefficients $A$. En général, $B$ peut être choisie de façon différente. Nous allons considérer ici le cas

$$B = B_K = \left(I_q - CH^{(K)}C^{-1}\right)^\top = \left[C\left(I_q - H^{(K)}\right)C^{-1}\right]^\top, \tag{7}$$

où $C$ est la matrice des coefficients canoniques (introduite à la section 1.3) et

$$H^{(K)} = \operatorname{diag}\left(h_1^{(K)}, \ldots, h_q^{(K)}\right)$$

est une matrice diagonale dont les éléments sont des fonctions des éléments $f_1, \ldots, f_q$ définis ci-haut. Comme on le verra dans les sections suivantes, plusieurs choix des éléments diagonaux $h_i^{(K)}$ ont été proposés dans la littérature. L'indice $(K)$ indiquera le choix effectué.

## 2.1. L'ESTIMATEUR "CURDS AND WHEY" (C&W)

Notons $(\mathbf{y}_0, \mathbf{x}_0)$ une observation future dont la prévision MC est $\widehat{\mathbf{y}}_0 = \widehat{A}\mathbf{x}_0$. Nous supposons bien sûr que cette observation future est indépendante des observations $(Y, X)$. Dans le cas idéal, la distribution conjointe de $(\mathbf{y}, \mathbf{x})$ est supposée

$$\begin{pmatrix} \mathbf{y} \\ \mathbf{x} \end{pmatrix} \sim N_{p+q}\left( \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \Sigma + AVA^\top & AV \\ VA^\top & V \end{pmatrix} \right).$$

La matrice de rétrécissement $B$ est alors choisie pour minimiser le risque de prévision

$$\arg\min_B E[(\mathbf{y}_0 - B\widehat{A}\mathbf{x}_0)^\top \Omega^{-1}(\mathbf{y}_0 - B\widehat{A}\mathbf{x}_0)],$$

où $\Omega$ et une matrice symétrique définie positive arbitraire. La valeur optimale de $B$ est donc

$$B = E[\mathbf{y}_0\widehat{\mathbf{y}}_0^\top]E[\widehat{\mathbf{y}}_0\widehat{\mathbf{y}}_0^\top]^{-1} \tag{8}$$

et elle ne dépend pas de $\Omega$.

Breiman et Friedman **[3]** utilisent les expressions suivantes pour les espérances dans (8) :

$$E[\mathbf{y}_0\widehat{\mathbf{y}}_0^\top] = E[(A\mathbf{x}_0 + \boldsymbol{\varepsilon}_0)(\mathbf{x}_0^\top \widehat{A}^\top)] = AE(\mathbf{x}_0\mathbf{x}_0^\top)A^\top = AVA^\top,$$
$$E[\widehat{\mathbf{y}}_0\widehat{\mathbf{y}}_0^\top] = AVA^\top + \frac{p}{N}\Sigma.$$

Nous verrons à la section 2.3 que la valeur $\frac{p}{N}$ ci-dessus est erronée même pour des distributions multinormales. Ils obtiennent donc que

$$B = \left(I_q + r\Sigma G^{-1}\right)^{-1},$$

où $G = AVA^\top$ et $r = p/N$. Comme les paramètres $\Sigma$ et $G$ sont inconnus, on prend les estimateurs échantillonaux $S$ et $Y^\top MY$ et l'estimateur de $B$ devient

$$\widehat{B}_{OPT} = [I_q + rS(Y^\top MY)^{-1}]^{-1}.$$

En utilisant (3), l'estimateur s'écrit en termes de $C$ et $F$ comme suit :

$$\widehat{B}_{OPT} = \left[I_q + rC^{-\top}C^{-1}(CF^{-1}C^\top)\right]^{-1} = \left[I_q + rC^{-\top}F^{-1}C^\top\right]^{-1}$$
$$= \left[C^{-\top}\left(I_q + rF^{-1}\right)C^\top\right]^{-1} = C^{-\top}\left(I_q + rF^{-1}\right)^{-1}C^\top.$$

Il résulte de cette identité que la matrice $(I_q+rF^{-1})^{-1}$ est diagonale avec éléments diagonaux

$$\Big(1+\frac{r}{f_i}\Big)^{-1}=\frac{f_i}{r+f_i}=1-\frac{r}{r+f_i},$$

ce qui nous permet d'écrire l'estimateur sous la forme

$$\begin{aligned}\widehat{B}_{OPT}&=C^{-\top}\big(I_q-H^{(OPT)}\big)C^{\top}=\big(I_q-CH^{(OPT)}C^{-1}\big)^{\top},\\ H^{(OPT)}&=\operatorname{diag}\Big(\frac{r}{r+f_1},\ldots,\frac{r}{r+f_q}\Big).\end{aligned}\tag{9}$$

## 2.2. L'ESTIMATEUR C&W–GCV

Notons que la matrice $B$ dans (8) est obtenue à partir d'une observation future $\mathbf{y}_0$ qui ne fait pas partie de l'échantillon. Cette procédure peut être approchée par la méthode de la validation croisée, notée par l'acronyme CV. Pour ce faire, on enlève chaque observation et on la traite comme une observation future. L'analogue CV de (8) devient donc

$$B=Y^{\top}\widehat{Y}_{\backslash n}\big(\widehat{Y}_{\backslash n}^{\top}\widehat{Y}_{\backslash n}\big)^{-1},\tag{10}$$

où $\widehat{Y}_{\backslash n}=\begin{pmatrix}\widehat{\mathbf{y}}_{\backslash 1}^{\top}\\ \vdots\\ \widehat{\mathbf{y}}_{\backslash N}^{\top}\end{pmatrix}$, $\widehat{\mathbf{y}}_{\backslash k}=\widehat{A}_{\backslash k}\mathbf{x}_k$ est l'estimation MC avec la $k$-ième observation enlevée. Breiman et Friedman [3] proposent une simplification de plus en calculant les $\widehat{\mathbf{y}}_{\backslash k}$ à l'aide de la formule :

$$\widehat{\mathbf{y}}_{\backslash k}=\frac{1}{1-m_{kk}}(\widehat{\mathbf{y}}_k-m_{kk}\mathbf{y}_k)\approx\frac{1}{1-r}(\widehat{\mathbf{y}}_k-r\mathbf{y}_k),$$

où les éléments diagonaux $m_{kk}$ de la matrice $M$ sont remplacés par la moyenne :

$$m_{kk}\approx\frac{1}{N}\operatorname{trace}(M)=r.\tag{11}$$

Cette approximation donne lieu à la validation croisée généralisée ou GCV. Nous avons donc

$$\widehat{Y}_{\backslash n}\approx\frac{1}{1-r}\big(\widehat{Y}-rY\big)$$

que nous substituons dans (10) pour recalculer $B$ (à l'aide de (1)) :

$$\begin{aligned} B &= \frac{1}{1-r}(Y^\top \widehat{Y} - rY^\top Y)\Big[\frac{1}{(1-r)^2}\Big(\widehat{Y}^\top\widehat{Y} - r(Y^\top\widehat{Y} + \widehat{Y}^\top Y) + r^2 Y^\top Y\Big)\Big]^{-1} \\ &= (1-r)\Big(Y^\top MY - rY^\top Y\Big)\Big(Y^\top MY - 2rY^\top MY + r^2 Y^\top Y\Big)^{-1} \\ &= (1-r)\Big(Y^\top MY - rY^\top Y\Big)\left[(1-2r)Y^\top MY + r^2 Y^\top Y\right]^{-1}. \end{aligned}$$

Puisque les matrices $Y^\top Y$ et $Y^\top MY$ sont symétriques et $\widehat{Q}^\top = (Y^\top MY)(Y^\top Y)^{-1}$, on a

$$\begin{aligned} B &= (1-r)\Big(\widehat{Q}^\top - rI_q\Big)\Big(Y^\top Y\Big)\left\{\left[(1-2r)\widehat{Q}^\top + r^2 I_q\right]\Big(Y^\top Y\Big)\right\}^{-1} \\ &= (1-r)\Big(\widehat{Q}^\top - rI_q\Big)\left[(1-2r)\widehat{Q}^\top + r^2 I_q\right]^{-1}. \end{aligned}$$

Pour exprimer $B$ en termes des matrices $C$ et $F$, nous utilisons l'identité (6) et le fait que $I_q = C^{-\top}C^\top$ :

$$\begin{aligned} B &= (1-r)C^{-\top}\left[(I_q + F^{-1})^{-1} - rI_q\right]C^\top C^{-\top}\left[(1-2r)(I_q + F^{-1})^{-1} + r^2 I_q\right]^{-1}C^\top \\ &= (1-r)C^{-\top}\left[(I_q + F^{-1})^{-1} - rI_q\right]\left[(1-2r)(I_q + F^{-1})^{-1} + r^2 I_q\right]^{-1}C^\top \\ &= C^{-\top}DC^\top, \end{aligned}$$

où nous avons posé

$$D = (1-r)\left[(I_q + F^{-1})^{-1} - rI_q\right]\left[(1-2r)(I_q + F^{-1})^{-1} + r^2 I_q\right]^{-1}.$$

Puisque les matrices $I_q$ et $F$ sont diagonales, $D = \mathrm{diag}(d_1, \ldots, d_q)$ l'est aussi et ses éléments s'expriment :

$$\begin{aligned} d_i &= (1-r)\left[\Big(1 + \frac{1}{f_i}\Big)^{-1} - r\right]\left[(1-2r)\Big(1 + \frac{1}{f_i}\Big)^{-1} + r^2\right]^{-1} \\ &= (1-r)\frac{(1-r)f_i - r}{(1-r)^2 f_i + r^2}. \end{aligned}$$

Ceci entraîne la forme cherchée de $B$ :

$$\begin{aligned} B &= C^{-\top}\left[I_q - (I_q - D)\right]C^\top = C^{-\top}\big(I_q - H^{(CWG)}\big)C^\top \\ &= \left[C\big(I_q - H^{(CWG)}\big)C^{-1}\right]^\top, \end{aligned}$$

où $H^{(CWG)} = \text{diag}\Big(h_1^{(CWG)}, \ldots, h_q^{(CWG)}\Big)$, $h_i^{(CWG)} = 1 - d_i = \frac{r}{(1-r)^2 f_i + r^2}$. Pour l'estimateur C&W généralisé on obtient donc

$$\begin{aligned} B_{CWG} &= \Big(I_q - CH^{(CWG)}C^{-1}\Big)^\top, \\ H^{(CWG)} &= \text{diag}\Big(h_1^{(CWG)}, \ldots, h_q^{(CWG)}\Big), \\ h_i^{(CWG)} &= \frac{r}{(1-r)^2 f_i + r^2}. \end{aligned} \tag{12}$$

## 2.3. L'ESTIMATEUR C&W–GCV OPTIMALE

Calculons la matrice $B$ de (8) en corrigeant l'erreur introduite par Breiman et Friedman **[3]**. La deuxième espérance est une matrice

$$E[\widehat{\mathbf{y}}_0\widehat{\mathbf{y}}_0^\top] = E[\widehat{A}\mathbf{x}_0\mathbf{x}_0^\top\widehat{A}^\top] = E[\widehat{A}V\widehat{A}^\top] = \Big(E[\widehat{\mathbf{a}}_i^\top V\widehat{\mathbf{a}}_j]\Big)_{i,j=1,\ldots,q}$$

dont l'élément $(i,j)$ est

$$\begin{aligned} E[\widehat{\mathbf{a}}_i^\top V\widehat{\mathbf{a}}_j] &= tr[VE(\widehat{\mathbf{a}}_i\widehat{\mathbf{a}}_j^\top)] = tr\{VE[(X^\top X)^{-1}X^\top\mathbf{y}_i\mathbf{y}_j^\top X(X^\top X)^{-1}]\} \\ &= tr\{VE[(X^\top X)^{-1}X^\top(\sigma_{ij}I_p + X\mathbf{a}_i\mathbf{a}_j^\top X^\top)X(X^\top X)^{-1}]\} \\ &= \sigma_{ij}\ tr\{VE[(X^\top X)^{-1}]\} + \mathbf{a}_i^\top V\mathbf{a}_j, \end{aligned}$$

où $\sigma_{ij}$ sont les éléments de $\Sigma$. Puisque la matrice $X^\top X \sim W_p(N,V)$ suit la distribution de Wishart, $E[(X^\top X)^{-1}] = \frac{1}{(N-p-1)}V^{-1}$ et l'élément $(i,j)$ s'exprime comme

$$E[\widehat{\mathbf{a}}_i^\top V\widehat{\mathbf{a}}_j] = \frac{1}{(N-p-1)}\sigma_{ij} + \mathbf{a}_i^\top V\mathbf{a}_j.$$

Il en résulte que la deuxième espérance dans (8) est

$$E[\widehat{\mathbf{y}}_0\widehat{\mathbf{y}}_0^\top] = AVA^\top + \frac{p}{(N-p-1)}\Sigma.$$

On notera que cette espérance repose sur la normalité des $p$ variables explicatives. Posons $r_1 = \frac{p}{N-p-1}$. La matrice optimale $B$ prend en effet la forme

$$B = \big(I_q + r_1\Sigma G^{-1}\big)^{-1}.$$

Dans [1], Bilodeau a montré que l'estimateur CWG (12) n'est pas minimax et il propose d'utiliser dans (11) l'approximation $m_{kk} \approx r_1$, ce qui donne l'estimateur

$$\begin{aligned} B_{CWGopt} &= \left(I_q - CH^{(CWGopt)}C^{-1}\right)^\top, \\ H^{(CWGopt)} &= \operatorname{diag}\left(h_1^{(CWGopt)}, \ldots, h_q^{(CWGopt)}\right), \\ h_i^{(CWGopt)} &= \frac{r_1}{(1-r_1)^2 f_i + r_1^2}. \end{aligned} \tag{13}$$

## 2.4. Les estimateurs minimax

Notons par

$$\rho_2^{(K)} = E\left[\operatorname{tr}\Sigma^{-1}(A - \widetilde{A}_K)X^\top X(A - \widetilde{A}_K)^\top\right]$$

la fonction de risque pour l'estimation de $A$, où $\widetilde{A}_K$ est un estimateur des coefficients $A$. Pour l'estimateur MC on a $\rho_2^{(\widehat{A})} = pq$ et les estimateurs minimax satisfont donc $\Delta = \rho_2^{(K)} - pq \leq 0$.

Srivastava et Solanky [10] proposent un estimateur sans biais de $\Delta$ et, en considérant les estimateurs de la forme

$$\begin{aligned} \widetilde{A}_K &= \left[I_q - CH^{(K)}C^{-1}\right]^\top \widehat{A}, \\ H^{(K)} &= \operatorname{diag}\left(h_1^{(K)}, \ldots, h_q^{(K)}\right), \end{aligned}$$

ils cherchent à déterminer les $\{h_i^{(K)}\}$ (considérés comme des fonctions de $\{f_1, \ldots, f_q\}$) pour que l'estimateur $\widetilde{A}_K$ soit minimax. Ils obtiennent ainsi deux estimateurs que nous allons introduire maintenant.

Le premier estimateur, noté $\widetilde{A}_{KS}$, est déterminé en cherchant $\{h_i^{(K)}\}$ de la forme $h_i = d_i/f_i$, où $d_i$ sont des constantes positives et ordonnées :

$$\begin{aligned} B_{KS} &= \left(I_q - CH^{(KS)}C^{-1}\right)^t, \\ H^{(KS)} &= \operatorname{diag}\left(h_1^{(KS)}, \ldots, h_q^{(KS)}\right), \\ h_i^{(KS)} &= \frac{p+q-2i-1}{(N-p-q+2i+1)} \cdot \frac{1}{f_i}. \end{aligned} \tag{14}$$

Cet estimateur est minimax si $p \geq q+1$ et $N \geq p-q-1$

Si on cherche les $\{h_i^{(K)}\}$ de la forme $h_i = d/f_i$, où $d$ est une constante, ils obtiennent l'autre estimateur $\widetilde{A}_{SS}$ comme

$$\begin{aligned} B_{SS} &= \left(I_q - CH^{(SS)}C^{-1}\right)^\top, \\ H^{(SS)} &= \operatorname{diag}\left(h_1^{(SS)}, \ldots, h_q^{(SS)}\right), \\ h_i^{(SS)} &= \frac{N(p+q-1)}{(N-p)(N-p+q+1)}\left(\frac{1}{f_i}\right). \end{aligned} \tag{15}$$

Ce dernier estimateur est minimax si $N \geq \frac{2p(p-q-1)}{p-3q-1}$ et $p \geq 3q+1$.

## 2.5. Les parties positives des estimateurs

Considérons les estimateurs du type (7). Leur facteur de rétrécissement est la matrice diagonale $(I_q - H^{(K)})$ dont les éléments peuvent prendre des valeurs négatives. La pratique nous montre que ces estimateurs sont dominés par les estimateurs où les matrices $H^{(K)}$ sont remplacées par la partie positive i.e.

$$\begin{aligned} H_+^{(K)} &= \operatorname{diag}\left(h_{1+}^{(K)}, \ldots, h_{q+}^{(K)}\right), \\ h_{i+}^{(K)} &= \min\{1, h_i^{(K)}\} \end{aligned}$$

(voir les articles de Srivastava et Solanky [**10**] et Breiman et Friedman [**3**]).

Il y a plusieurs autres estimateurs à rétrécissement. Dans [**11**], van der Merwe et Zidek introduisent l'estimateur minimax FICYREG (noté MZ). Bilodeau et Kariya [**2**] proposent un estimateur du type Efron-Morris (noté EM), qui domine uniformément le MZ (en termes du risque $\rho_2$), mais qui n'est pas invariant. Konno [**6**] suggère un estimateur du type Efron-Morris (noté EMI) qui est invariant. Dans [**10**], Srivastava et Solanky prennent en considération les trois estimateurs MZ, EM et EMI, ainsi que l'estimateur KC (introduit dans leur article) et montrent qu'ils sont inférieurs aux estimateurs KS+ et SS+ dans le sens des critères de la section 4.3. Dans [**3**], Breiman et Friedman comparent les estimateurs CWG+ et OPT aux estimateurs MC et MZ (ainsi qu'aux estimateurs obtenus par les méthodes *Separate Ridge Regression*(Hoerl et Kennard, [**4**]), *Reduced Rank Regression* (Izenman, [**5**]), *PLS Regression* (Wold, [**12**])) et les trouvent supérieurs toujours au sens des critères de la section 4.3.

Pour étudier le comportement de l'estimateur FCV (introduit dans la section suivante), par des simulations et des exemples, nous allons donc utiliser l'estimateur OPT (défini dans (9)) et les trois estimateurs CWG+, KS+ et SS+ (les parties positives des estimateurs définis dans (12), (14) et (15)). Nous allons aussi considérer la modification CWGopt+ proposée par Bilodeau [1] (les parties positives de l'estimateur défini dans (13)). Finalement, nous allons inclure l'estimateur MC comme référence de base. On remarquera que les éléments $h_i^{(OPT)}$ de l'estimateur OPT sont toujours inférieurs à 1 ; il est donc inutile de prendre la partie positive.

# Chapitre 3

## L'ESTIMATEUR FCV ("FULL CROSS VALIDATION")

Dans leur article [3], Breiman et Friedman proposent l'estimateur FCV, $\widetilde{A}_{FCV} = B\widehat{A}$, tel que $B$ prend la forme

$$B = C^{-\top} D C^{\top} = \left(C D C^{-1}\right)^{\top},$$

où la matrice diagonale $D$ est

$$\begin{aligned} D &= \arg\min_{\Delta=\text{diag}} \sum_{i=1}^{q} \sum_{n=1}^{N} \left[ y_{ni} - \left( C_{\setminus n}^{-\top} \Delta C_{\setminus n}^{\top} \widehat{\mathbf{y}}_{\setminus n} \right)_i \right]^2 \\ &= \arg\min_{\Delta=\text{diag}} \sum_{n=1}^{N} \| \mathbf{y}_n - C_{\setminus n}^{-\top} \Delta C_{\setminus n}^{\top} \widehat{\mathbf{y}}_{\setminus n} \|^2. \end{aligned} \tag{16}$$

L'expression $C_{\setminus n}$ désigne la matrice des coefficients canoniques pour les données avec la $n$-ième observation enlevée et $\widehat{\mathbf{y}}_{\setminus n}$ est l'estimation MC avec la $n$-ième observation enlevée. La minimisation est faite sur toutes les matrices $\Delta = \text{diag}(\delta_1, \ldots, \delta_q)$. Le résultat est donc une matrice diagonale $D = \text{diag}(d_1, \ldots, d_q)$ dont les éléments ne sont pas a priori ordonnés et peuvent prendre des valeurs négatives. Dans [3], Breiman et Friedman proposent de régler ce problème en remplaçant $D$ par la matrice "la plus convenable" dont les éléments sont ordonnés et positifs. Nous considérons ici une méthode alternative, où les contraintes sont imposées lors de la minimisation.

Dans la discussion sur l'article [3], P. D. Sasieni propose une minimisation directe, avec les mêmes contraintes, à savoir on minimise (16) sous les contraintes

linéaires

$$\delta_1 \geq \delta_2 \geq \cdots \geq \delta_q \geq 0.$$

Ce problème de programmation quadratique peut être résolu à l'aide de différents logiciels mathématiques. Par exemple, Fortran 90 offre les sousroutines **QPROG** (dans la librairie IMSL pour Windows) et **E04NFF** ou **nag_qp_sol** (dans la bibliothèque NAG pour UNIX).

Pour calculer $D$, nous allons écrire les contraintes linéaires sous une forme matricielle et nous allons transformer la fonction objective

$$\Phi := \sum_{n=1}^{N} \|\mathbf{y}_n - C_{\backslash n}^{-\top} \Delta C_{\backslash n}^{\top} \widehat{\mathbf{y}}_{\backslash n}\|^2$$

de façon que la minimisation soit faite sur le vecteur $\delta := (\delta_1, \ldots, \delta_q)$ au lieu de la matrice $\Delta$.

Les contraintes sont :

$$\delta_1 \geq \delta_2 \geq \ldots \geq \delta_q \geq 0 \Leftrightarrow$$

$$\delta_1 - \delta_2 \geq 0,\ \delta_2 - \delta_3 \geq 0, \ldots, \delta_{q-1} - \delta_q \geq 0 \text{ et } \delta_q \geq 0 \Leftrightarrow$$

$$\begin{pmatrix} 1 & -1 & 0 & \ldots & 0 \\ 0 & 1 & -1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \delta \geq 0 \ \Leftrightarrow\ H\delta \geq 0.$$

La fonction objective s'écrit :

$$\begin{aligned} \Phi &= \sum_{n=1}^{N} \|\mathbf{y}_n - C_{\backslash n}^{-\top} \Delta C_{\backslash n}^{\top} \widehat{\mathbf{y}}_{\backslash n}\|^2 \\ &= \sum_{n=1}^{N} \left(\mathbf{y}_n - C_{\backslash n}^{-\top} \Delta C_{\backslash n}^{\top} \widehat{\mathbf{y}}_{\backslash n}\right)^{\top} \left(\mathbf{y}_n - C_{\backslash n}^{-\top} \Delta C_{\backslash n}^{\top} \widehat{\mathbf{y}}_{\backslash n}\right) \\ &= \sum_{n=1}^{N} \left(\mathbf{y}_n^{\top} - \widehat{\mathbf{y}}_{\backslash n}^{\top} C_{\backslash n} \Delta C_{\backslash n}^{-1}\right) \left(\mathbf{y}_n - C_{\backslash n}^{-\top} \Delta C_{\backslash n}^{\top} \widehat{\mathbf{y}}_{\backslash n}\right) \end{aligned}$$

$$= \sum_{n=1}^{N} \left( \mathbf{y}_n^\top \mathbf{y}_n - \widehat{\mathbf{y}}_{\setminus n}^\top C_{\setminus n} \Delta C_{\setminus n}^{-1} \mathbf{y}_n - \mathbf{y}_n^\top C_{\setminus n}^{-\top} \Delta C_{\setminus n}^\top \widehat{\mathbf{y}}_{\setminus n} + \widehat{\mathbf{y}}_{\setminus n}^\top C_{\setminus n} \Delta C_{\setminus n}^{-1} C_{\setminus n}^{-\top} \Delta C_{\setminus n}^\top \widehat{\mathbf{y}}_{\setminus n} \right)$$

$$= \sum_{n=1}^{N} \Big( \mathbf{y}_n^\top \mathbf{y}_n - (C_{\setminus n}^\top \widehat{\mathbf{y}}_{\setminus n})^\top \Delta (C_{\setminus n}^{-1} \mathbf{y}_n) - (C_{\setminus n}^{-1} \mathbf{y}_n)^\top \Delta (C_{\setminus n}^\top \widehat{\mathbf{y}}_{\setminus n}) + \\ + \widehat{\mathbf{y}}_{\setminus n}^\top C_{\setminus n} \Delta C_{\setminus n}^{-1} C_{\setminus n}^{-\top} \Delta C_{\setminus n}^\top \widehat{\mathbf{y}}_{\setminus n} \Big).$$

En utilisant $a^\top \Delta b = b^\top \Delta a$ on obtient

$$\Phi = \sum_{n=1}^{N} \left[ \mathbf{y}_n^\top \mathbf{y}_n - 2 \widehat{\mathbf{y}}_{\setminus n}^\top C_{\setminus n} \Delta C_{\setminus n}^{-1} \mathbf{y}_n + \widehat{\mathbf{y}}_{\setminus n}^\top C_{\setminus n} \Delta C_{\setminus n}^{-1} C_{\setminus n}^{-\top} \Delta C_{\setminus n}^\top \widehat{\mathbf{y}}_{\setminus n} \right]$$

$$= tr(YY^\top) - 2 \sum_{n=1}^{N} \left( C_{\setminus n}^\top \widehat{\mathbf{y}}_{\setminus n} \right)^\top \Delta \left( C_{\setminus n}^{-1} \mathbf{y}_n \right) + \\ + \sum_{n=1}^{N} \left[ \left( C_{\setminus n}^\top \widehat{\mathbf{y}}_{\setminus n} \right)^\top \Delta \left( C_{\setminus n}^{-1} C_{\setminus n}^{-\top} \right) \Delta \left( C_{\setminus n}^\top \widehat{\mathbf{y}}_{\setminus n} \right) \right].$$

Pour transformer $\Phi$, nous utiliserons les équations :

$$\mathbf{a}^\top \Delta = (\Delta \mathbf{a})^\top = (\delta_1 a_1, \ldots, \delta_q a_q) = \delta^\top \Delta_a = (\Delta_a \delta)^\top,$$

où $\mathbf{a} = (a_1, \ldots, a_q)^\top$ et $\Delta_a = \mathrm{diag}(\mathbf{a})$. Posons $\Psi_{\setminus n} := \mathrm{diag}(C_{\setminus n}^\top \widehat{\mathbf{y}}_{\setminus n}) = \Psi_{\setminus n}^\top$ une matrice diagonale. Nous avons alors

$$\Phi = tr(YY^\top) - 2 \sum_{n=1}^{N} \delta^\top \Psi_{\setminus n} C_{\setminus n}^{-1} \mathbf{y}_n + \sum_{n=1}^{N} \delta^\top \Psi_{\setminus n} C_{\setminus n}^{-1} C_{\setminus n}^{-\top} \Psi_{\setminus n} \delta$$

$$= tr(YY^\top) - 2 \sum_{n=1}^{N} \left[ \left( C_{\setminus n}^{-\top} \Psi_{\setminus n} \right)^\top \mathbf{y}_n \right]^\top \delta + \sum_{n=1}^{N} \delta^\top \left( C_{\setminus n}^{-\top} \Psi_{\setminus n} \right)^\top \left( C_{\setminus n}^{-\top} \Psi_{\setminus n} \right) \delta$$

$$= tr(YY^\top) - 2 \Big[ \sum_{n=1}^{N} \left( C_{\setminus n}^{-\top} \Psi_{\setminus n} \right)^\top \mathbf{y}_n \Big]^\top \delta + \delta^\top \Big[ \sum_{n=1}^{N} \left( C_{\setminus n}^{-\top} \Psi_{\setminus n} \right)^\top \left( C_{\setminus n}^{-\top} \Psi_{\setminus n} \right) \Big] \delta.$$

Soit

$$\alpha := \sum_{n=1}^{N} \left( C_{\setminus n}^{-\top} \Psi_{\setminus n} \right)^\top \left( C_{\setminus n}^{-\top} \Psi_{\setminus n} \right)$$

une matrice $q \times q$, symétrique définie positive, et posons

$$\beta := \sum_{n=1}^{N} \left( C_{\setminus n}^{-\top} \Psi_{\setminus n} \right)^\top \mathbf{y}_n$$

un vecteur $q \times 1$. Notons aussi $D = \text{diag}(d) = \text{diag}(d_1, \cdots, d_q)$, où $d = (d_1, \cdots, d_q)^\top$ est un vecteur $q \times 1$. L'équation (16) se transforme alors comme suit :

$$\begin{aligned} d &= \arg \min_{\delta:\ H\delta \geq 0} \left[\delta^\top \alpha \delta - 2\beta^\top \delta + \text{tr}(Y^\top Y)\right] \\ &= \arg \min_{\delta:\ H\delta \geq 0} \left(\delta^\top \alpha \delta - 2\beta^\top \delta\right). \end{aligned}$$

L'estimateur FCV prend la forme

$$\widetilde{A}_{FCV} = \widehat{A}B, \quad B = (CDC^{-1})^\top, \quad D = \text{diag}(d),$$

$$d = \arg \min_{\delta:\ H\delta \geq 0} (\delta^\top \alpha \delta - 2\beta^\top \delta).$$

Écrite dans la forme générale, $B$ devient

$$\begin{aligned} B &= \Big(CDC^{-1}\Big)^\top = \Big[C\Big(I_q - I_q + D\Big)C^{-1}\Big]^\top = \Big[I_q - C(I_q - D)C^{-1}\Big]^\top \\ &= \Big[I_q - CH^{(FCV)}C^{-1}\Big]^\top, \\ H^{(FCV)} &= I_q - D = I_q - \text{diag}(d) = \text{diag}(1 - d_1, \ldots, 1 - d_q). \end{aligned}$$

Puisque les $d_i$ sont positifs, notons que nous avons $h_i^{(FCV)} \leq 1$, i.e.

$$H^{(FCV)} = H_+^{(FCV)},$$

où $H_+^{(FCV)}$ est la partie positive de $H^{(FCV)}$ définie comme à la section 2.5.

La force de l'estimateur FCV réside dans le fait qu'on estime les coefficients sans supposer que les données sont normales. Cette méthode ne dépend pas de la distribution de $X$ ni de celle de $Y$.

# Chapitre 4

# LES SIMULATIONS

Nous choisissons d'étudier le comportement de l'estimateur $\widehat{A}_{FCV}$ en le comparant avec les six estimateurs $\widehat{A}$, $\widehat{A}_{OPT}$, $\widehat{A}_{CWGopt+}$, $\widehat{A}_{CWG+}$, $\widehat{A}_{KS+}$ et $\widehat{A}_{SS+}$ (notés par (K)). Le modèle de la simulation est pris de l'article de Srivastava et Solanky [**10**] (voir aussi Breiman et Friedman [**3**]).

## 4.1. Les données de la loi normale

Dans les différentes simulations, le nombre $p$ des variables explicatives est $p = 10, 20, 30$ et 50. Pour chaque $p$, nous choisissons un nombre différent de variables dépendantes. La taille échantillonnale est choisie en considérant des petites et des grandes tailles par rapport au nombre des variables (voir dans le tableau 4.1).

| p | q | N |
|---|---|---|
| 10 | 3,4,5,6,7,8 | 30,50,100,200,300 |
| 20 | 5,10,15,18 | 50,100,200,300,400 |
| 30 | 5,10,15,20,25,28 | 100,200,300,400,500 |
| 50 | 10,20,30,40,45 | 100,200,300,400,500 |

Tab. 4.1. Les situations de la simulation normale

Pour chaque répétition aléatoire de chaque situation, les $p$ variables explicatives sont générées de la loi normale, i.e. chaque observation $\mathbf{x}_n$ est générée

$$\mathbf{x}_n \sim N_p\Big(0, V\Big),\ n = 1, \ldots, N.$$

La matrice de covariance est elle-même aléatoire avec une réalisation différente à chaque répétition :

$$V = (v_{ij}) = \Big(r^{|i-j|}\Big),\ i, j = 1, \ldots, p,$$

où $r$ est généré de la loi $r \sim Unif[-1, 1]$. Le choix aléatoire de $r$ nous garantit un degré différent de collinéarité pour les différentes répétitions. Quand $r \approx 1$ la collinéarité entre les variables explicatives est forte, alors que $r \approx 0$ indique qu'elles sont presque indépendantes. Les réponses sont générées selon le modèle :

$$\mathbf{y}_n = A\mathbf{x}_n + \boldsymbol{\varepsilon}_n,\ n = 1, \ldots, N$$

ou la forme matricielle

$$Y = XA^\top + E. \tag{17}$$

Les termes d'erreurs sont générés de la loi normale

$$\boldsymbol{\varepsilon}_n \sim N_p\Big(0, \Sigma\Big),\ n = 1, \ldots, N$$

avec deux structures de covariance, $\Sigma = \sigma^2 I_q$ (les variances des erreurs associées à chaque réponse sont les mêmes) et $\Sigma = \sigma^2 \text{diag}(q^2, ..., 2^2, 1^2)$ (les variances sont très différentes). Pour chaque répétition, $\sigma^2$ est choisi de façon aléatoire entre les valeurs $\{1, 3, 10\}$.

Pour chaque répétition, les paramètres de la régression $A = (a_{ij})$ sont générés selon le modèle

$$A = CG^\top,$$

où $G$ est une matrice $p \times 10$ dont les éléments sont

$$g(j, k) = h_k\Big((l_k - \mid j - j_k \mid)_+\Big)^2,\ j = 1, \ldots, p,\ k = 1, \ldots, 10.$$

Les valeurs $h_k$ sont ajustées pour que

$$\sum_{j=1}^{p} g(j, k) = 1,\ k = 1, \ldots, 10.$$

Les valeurs $j_k$ et $l_k$ sont des entiers aléatoires des lois $Unif\{1, \ldots, 50\}$ et $Unif\{1, \ldots, 6\}$, respectivement.

La matrice $C$, de dimension $q \times 10$, a des vecteurs-colonnes générés indépendamment selon la loi

$$(c_{1k}, \ldots, c_{qk}) \sim N_q\Big(0, \Gamma\Big), \ k = 1, \ldots, 10$$

avec matrice de covariance

$$\Gamma = \Big(\gamma_{ij}\Big) = \Big(\rho^{|i-j|}\Big), \ i, j = 1, \ldots, q,$$

où $\rho$ est généré de la loi $\rho \sim Unif[-1, 1]$.

Pour chaque situation $(p, q, N, \Sigma)$, nous avons fait 500 répétitions (ce qui donne 105 000 répétitions en tout).

## 4.2. Les données de la loi $t$

Contrairement aux autres estimateurs, le FCV ne suppose pas que les données soient normales. Une deuxième simulation a comme but d'étudier son comportement sur des données provenant de la distribution $t$ de Student. Les matrices $A$ et $X$ sont générées comme avant, la matrice $Y$ se calcule du modèle (17), mais les termes d'erreurs sont générés de la loi $t$ ,

$$\varepsilon_n \sim t_{q,\nu}\big(0, \Sigma\big), \ n = 1, \ldots, N$$

avec le même choix de $\Sigma$. Pour obtenir la matrice $E$, nous avons besoin de générer une matrice $Z$ de dimension $N \times q$, telle que

$$Z = \big(z_{ij}\big), \ z_{ij} \sim N\big(0, 1\big), \ i = 1, \ldots, N, \ j = 1, \ldots, q,$$

et d'un vecteur $\omega$, $N \times 1$, tel que

$$\omega = \big(\omega_i\big), \ \ \omega_i \sim \chi^2_\nu, \ i = 1, \ldots, N.$$

La matrice

$$T := \sqrt{\nu} \begin{pmatrix} z_{11}/\sqrt{\omega_1} & z_{12}/\sqrt{\omega_1} & \cdots & z_{1q}/\sqrt{\omega_1} \\ z_{21}/\sqrt{\omega_2} & z_{22}/\sqrt{\omega_2} & \cdots & z_{2q}/\sqrt{\omega_2} \\ \vdots & \vdots & \ddots & \vdots \\ z_{N1}/\sqrt{\omega_N} & z_{N2}/\sqrt{\omega_N} & \cdots & z_{Nq}/\sqrt{\omega_N} \end{pmatrix}$$

représente un échantillon de $N$ vecteurs (les lignes de $T$) provenant de la loi $t_{q,\nu}(0, I_q)$. Comme la matrice $\Sigma$ est diagonale, la matrice $E$ s'obtient par

$E = \sigma T$, dans le cas où $\Sigma = \sigma^2 I_q$,

$E = \sigma T \mathrm{diag}(q, \ldots, 2, 1)$, dans le cas où $\Sigma = \sigma^2 \mathrm{diag}(q^2, \ldots, 2^2, 1^2)$.

Comme le but de cette simulation additionnelle est de comparer les comportements des estimateurs sur les données non-normales, nous ignorons les grandes tailles $N$. Les situations étudiées sont les suivantes :

| p | q | N |
|---|---|---|
| 10 | 3,4,5,6,7,8 | 30,50,100 |
| 20 | 5,10,15,18 | 50,100,200 |
| 30 | 5,10,15,20,25,28 | 100,200,300 |
| 50 | 10,20,30,40,45 | 100,200,300 |

TAB. 4.2. Les situations de la simulation $t_{p,2}$

Pour chaque situation $(p, q, N, \Sigma)$, nous avons fait 500 répétitions (ce qui donne 63 000 répétitions en tout).

## 4.3. LES CRITÈRES

La performance des estimateurs est évaluée à l'aide des deux critères utilisés dans [**3**] et [**10**].

Considérons l'estimateur (K). Pour chaque réplicat l'erreur quadratique moyenne de la $j$-ième variable est donnée par

$$e_j(K)^2 = \big(\mathbf{a}_j - \widetilde{\mathbf{a}}_j(K)\big)^t V \big(\mathbf{a}_j - \widetilde{\mathbf{a}}_j(K)\big),$$

où la notation $\mathbf{a}_j(K)^\top$ correspond aux vecteur-lignes de la matrice $\widehat{A}_K$ (qui représente l'estimation obtenue par la méthode (K)). Les $\mathbf{a}_j^\top$ sont les vecteur-lignes de $A$, i.e. les vrais coefficients générés dans la simulation. Les deux critères pour chaque méthode $(K)$ s'écrivent comme suit :

$$A_{(K)} = \frac{\sum_{j=1}^q e_j(FCV)^2}{\sum_{j=1}^q e_j(K)^2},$$

$$W_{(K)} = \max_{j \in \{1,\ldots,q\}} \left\{ \frac{e_j(FCV)^2}{e_j(K)^2} \right\},$$

où $(K)$ est $\{MC, CWG+, CWGopt+, OPT, KS+, SS+\}$ (ce qui donne douze comparaisons en tout).

L'inégalité $A_{(K)} < 1$ signifie que la somme des carrés des erreurs obtenues par l'estimateur FCV est plus petite que la somme des carrés des erreurs obtenues par l'estimateur (K). Alors, pour cette répétition, l'estimateur FCV donne une estimation des coefficients $A$ plus exacte que celle donnée par l'estimateur (K). Par contre, l'inégalité $A_{(K)} > 1$ indique que l'estimateur (K) est plus efficace que l'estimateur FCV.

Le critère $W_{(K)}$ calcule le maximum des ratios $e_j(FCV)^2/e_j(K)^2$. C'est une mesure de la pire erreur donnée par FCV par rapport à (K). L'inégalité $W_{(K)} < 1$ signifie que l'estimateur FCV estime tous les coefficients mieux que l'estimateur (K). De façon réciproque, l'inégalité $W_{(K)} > 1$ indique qu'il existe au moins une variable réponse dont l'estimation respective par FCV est moins bonne que celle donnée par (K). Cependant, l'inégalité $W_{(K)} > 1$ ne signifie pas nécessairement que FCV est moins efficace que (K) sur l'ensemble de toutes les variables (par exemple dans le cas où $A_{(K)} < 1$). Notons qu'un grand nombre de variables réponse augmente la possibilité que les deux estimateurs FCV et (K) donnent ses pires erreurs sur des variables distinctes. De plus, pour certaines répétitions les erreurs $e_j^2$ sont très petites (d'ordre $10^{-4}$) et dans cette situation une valeur élevée de $W_{(K)}$ ne correspond pas à une grande différence réelle. Le critère $W_{(K)}$ devrait donc être considéré en combinaison avec $A_{(K)}$.

Comme tous les critères sont faits par rapport à la même référence (celle de FCV), les résultats nous permettent de comparer les comportements de tous les estimateurs (K). Par exemple, pour deux estimateurs (K) et (L) nous avons

$$A_{(K)} < A_{(L)} \Leftrightarrow \frac{\sum_{j=1}^{q} e_j(FCV)^2}{\sum_{j=1}^{q} e_j(K)^2} < \frac{\sum_{j=1}^{q} e_j(FCV)^2}{\sum_{j=1}^{q} e_j(L)^2}$$
$$\Leftrightarrow \sum_{j=1}^{q} e_j(K)^2 > \sum_{j=1}^{q} e_j(L)^2.$$

Ceci montre que l'estimateurs (L) est plus efficace que l'estimateur (K).

## 4.4. Les résultats

Les valeurs des critères $A_{(K)}$ et $W_{(K)}$ sont illustrées à l'aide de boxplots présentés dans l'annexe 1. La partie de gauche représente le critère $A_{(K)}$ tandis que celle de droite $W_{(K)}$, où $(K)$ désigne l'estimateur utilisé. Chaque figure contient les résultats de tous les douze critères, obtenus dans une situation particulière (marquée au-dessus). La variable $Index$ a été créée pour nous indiquer le choix de $\Sigma$ :

$Index = 1$, si $\Sigma = \sigma^2 I_q$,

$Index = 2$, si $\Sigma = \sigma^2 \mathrm{diag}(\mathrm{q}^2, \ldots, 2^2, 1^2)$.

Au début, nous avons fait un boxplot pour chaque situation $(p, q, N, Index)$ et ensuite les résultats semblables ont été réunis et présentés dans une même figure avec sa description au-dessus. La convention utilisée sera d'agréger tous les résultats sur les paramètres absents de la description. Par exemple, le premier boxplot de l'annexe 1 avec $p = 10,\ q = 3$ signifie que les résultats sont agrégés sur toutes les valeurs de $N$ et $Index$ et le dernier boxplot avec $p = 10,\ Index = 2$ signifie que les résultats sont agrégés sur toutes les valeurs de $q$ et $N$. Les boxplots (présentés dans l'annexe 1) sont les suivants :

- les situations $(p, q)$ : toutes les tailles $N$ et choix de $\Sigma$,
- les situations $(p, q, Index = 1)$ et $(p, q, Index = 2)$ : toutes les tailles $N$, car les deux choix de $\Sigma$ nous donnent toujours des résultats différents,
- certaines situations $(p, q, N)$ pour lesquelles les résultats obtenus pour cette taille $N$ sont différents des résultats de la situation respective $(p, q)$
- il y a quelques figures additionnelles qui contiennent uniquement les boxplots des critères $A_{(K)}$, quand les $W_{(K)}$ prennent des valeurs trop grandes.

Les deux simulations sont traitées séparément.

La ligne horizontale tracée à une hauteur de 1 est la ligne de référence qui nous aide à comparer les estimateurs (K) à l'estimateur FCV. Les boîtes au-dessous de la ligne indiquent que les estimateurs respectifs sont moins précis que le FCV. La boîte la plus haute nous montre le meilleur estimateur parmi les estimateurs (K) : MC, CWG+, OPT, CWGopt+, KS+ et SS+.

### 4.4.1. Conclusions de la simulation normale

Les estimateurs OPT et KS+ sont supérieurs à l'estimateur MC, mais ils sont toujours inférieurs aux autres estimateurs. Dépendamment des différentes situations le meilleur estimateur est parmi les FCV, CWG+, CWGopt+ et SS+.

En général, le comportement de l'estimateur CWG+ est inférieur au comportement de FCV, de CWGopt+ et de SS+. Dans certaines situations, CWG+ est équivalent à CWGopt+ et à SS+, mais il n'est jamais supérieur à FCV. Il atteint sa meilleure efficacité pour des petits $q$ : $q = 3, 4$, $p = 10$ (voir les boxplots à la page 38) ou $q = 5$, $p = 20, 30$ (pages 40 et 42), mais la seule situation où il peut être considéré comme un des meilleurs est le cas où $p = 10$, $q = 3$, $N = 30$ (voir le boxplot à la page 39).

Nous avons choisi la médiane comme une mesure robuste qui nous permet de voir facilement les tendances de comportement des estimateurs. Les figures suivantes désignent la médiane des critères $A_{(K)}$ en fonction de $q$. La convention utilisée est la même que pour les boxplots. La médiane est donc calculée sur 5000 observations pour les figures 4.1, 4.2-gauche et 4.3. Dans la figure 4.2-droite, la médiane est calculée sur 1000 observations de la situation $p = 30$, $N = 100$ (voir les boxplots à la page 43).

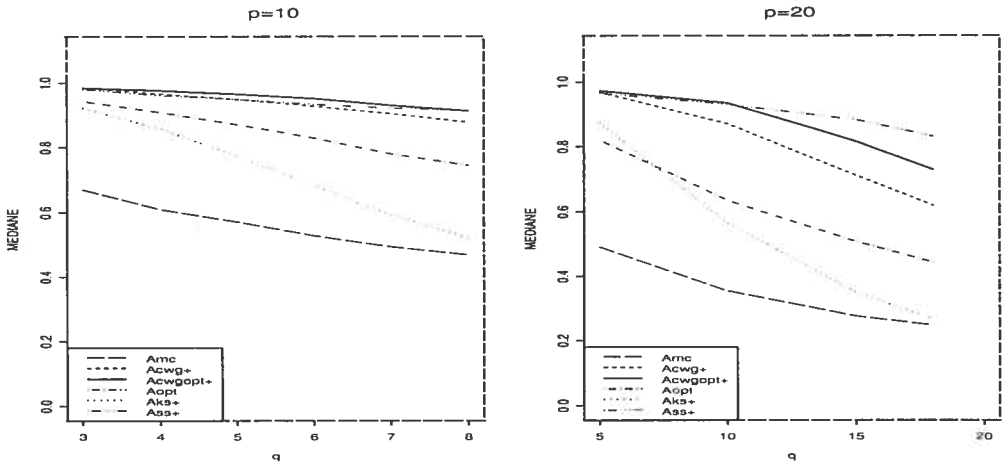



FIG. 4.1. La médiane des $A_{(K)}$ ; simulation normale ; situations $p = 10$ (à gauche) et $p = 20$ (à droite)

L'estimateur SS+ est toujours supérieur aux estimateurs MC, OPT et KS+. Il est fait pour des petites valeurs de $q$ quand il est un des meilleurs : voir par exemple les boxplots pour les situations $p = 10,\ q = 3, 4, 5$ (page 38) ; $p = 20,\ q = 5, 10$ (page 40) et $p = 30,\ q = 5$ (page 42). L'estimateur SS+ reste également un des meilleurs estimateurs pour des plus grandes valeurs de $q$, si la taille échantillonnale est assez grande. Par exemple, en considérant les grandes tailles $N$, dans des situations où $p = 20,\ q = 15, 18,\ N \geq 100$ (voir les boxplots à la page 41) et $p = 50,\ q \leq 30,\ N \geq 200$ (page 47), SS+ est l'un des deux meilleurs estimateurs (avec le CWGopt+) parmi les estimateurs (K), tandis que dans le cas où $p = 30,\ q \geq 10$ (page 42), il est le meilleur estimateur parmi les estimateurs (K). Cependant, pour les mêmes $(p, q)$ et des petites tailles $N$, l'estimateur SS+ devient inférieur aux estimateurs CWGopt+ et FCV : voir les boxplots des situations $p = 30,\ q \geq 15,\ N = 100$ (page 43) ; $p = 20,\ q = 15, 18,\ N = 50$ (page 41) ou $p = 50,\ N = 100$ (page 45 et 46). Les différents comportements de l'estimateur SS+ par rapport à la taille $N$ se voient bien en comparant les graphiques dans la figure 4.2.

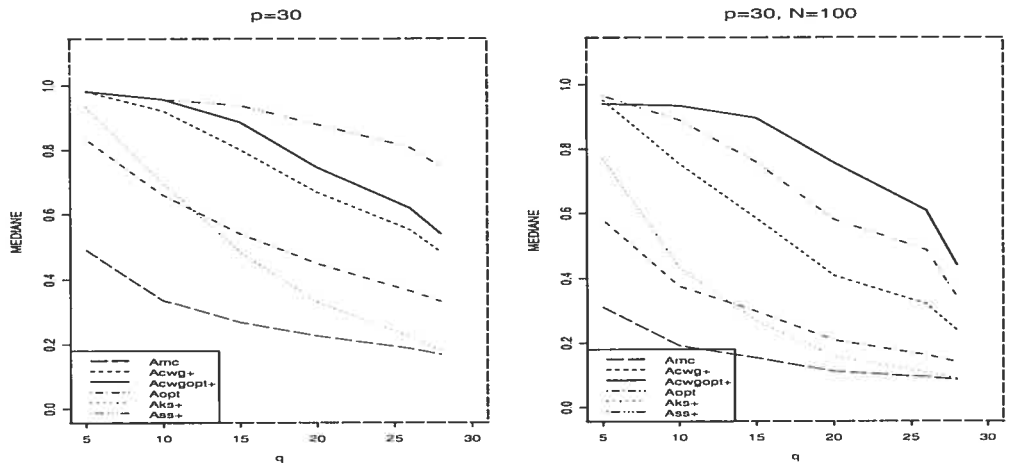



FIG. 4.2. La médiane des $A_{(K)}$ ; simulation normale ; situations $p = 30$ (à gauche) et $p = 30,\ N = 100$ (à droite)

Le SS+ est plus efficace dans les cas où $Index = 1$ (quand il est souvent équivalent à CWGopt+) et il est légèrement inférieur à l'estimateur FCV. En général, l'estimateur SS+ est inférieur à FCV. Le seul cas où SS+ a le même comportement que FCV (selon le critère $A_{(SS+)}$) est quand $p = 10,\ q = 3$ ;

puisque $W_{(SS+)} > 1$, SS+ peut être considéré meilleur que FCV et équivalent aux estimateurs CWG+ et CWGopt+ par des raisonnements similaires (voir les boxplots à la page 38).

L'estimateur CWGopt+ est un des meilleurs. Son comportement est toujours supérieur aux comportements des MC, OPT et KS+. Il est supérieur à CWG+ (sauf dans le cas $p = 10,\ q = 3$ où ils sont équivalents). Bien que dans certaines situations CWGopt+ est inférieur à SS+ (pour des grandes tailles $N$), il devient largement le meilleur parmi les estimateurs (K) pour des petits $N$, où il reste seul en concurrence avec FCV (voir les boxplots pour les situations $p = 10,\ N = 30, 50$ à la page 39 ; $p = 30,\ N = 100$ à la page 43 ou $p = 50,\ q \geq 20,\ N = 100$ à la page 46). Cette dépendance de la taille $N$ se voit bien en comparent les graphiques dans la figure 4.2. Pour $p = 30$ et toutes les tailles $N$, l'estimateur CWGopt+ est inférieur à SS+ et à FCV (4.2-gauche) ; pour $N = 100$ (4.2-droite), il devient le meilleur parmi les estimateurs (K).

L'estimateur CWGopt+ est uniformément un des meilleurs si $Index$ est égal à 1 ou 2. Comme on peut le voir dans les figures, il est un estimateur qui se comporte bien pour toutes les valeurs de $p$ et $q$ et il est très efficace dans presque toutes les situations (bien qu'en général il reste inférieur à l'estimateur FCV).

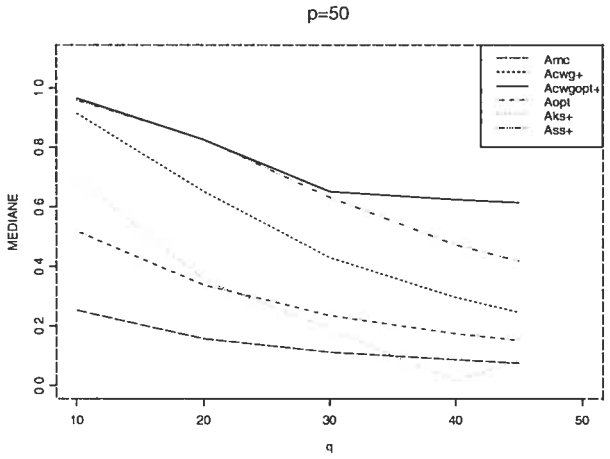



FIG. 4.3. La médiane des $A_{(K)}$ ; simulation normale ; situations $p = 50$

Le FCV est un très bon estimateur qui est presque toujours le meilleur. Son efficacité augmente avec la croissance des nombres des variables $p, q$ ou $p + q$.

Si le nombre des variables explicatives $p$ est fixé, son efficacité augmente avec la croissance du nombre des variables réponse $q$ (cette tendance se voit bien dans les figures). Pour des petits $q$ — par exemple $p = 10$, $q = 3, 4$ (voir les boxplots à la page 38) et $p = 20, 30$, $q = 5$ (pages 40 et 42) — le FCV est équivalent à CWGopt+ et à SS+. Avec la croissance de $q$, le FCV devient le meilleur. Avec des grandes valeurs $p$, il est le meilleur pour tout $q$ (voir les boxplots pour $p = 50$ à la page 44 ou bien la figure 4.3).

La taille d'échantillon $N$ n'influence pas beaucoup son comportement. Il est très efficace pour toutes valeurs de la taille $N$, mais le plus grand écart avec les autres estimateurs s'obtient pour des petits $N$ car FCV n'exige pas la normalité des données. Considérons par exemple le cas où $p = 10$. Pour $q = 3$ et 4, le FCV est équivalent à CWG+, à CWGopt+ et à SS+ (selon les critères $A$), mais en considérant seulement la taille $N = 30$, nous constatons que FCV est légèrement supérieur à ces trois estimateurs si $q = 3$ ; il est clairement supérieur à CWG+ et SS+ dans le cas où $q = 4$ (voir les boxplots à la page 39). Pour $q = 5, 6, 7$ et 8, les différences sont plus grandes et évidentes. Nous avons des résultats semblables pour $p = 20$ et 30. Pour $p = 50$, le meilleur estimateur est FCV (suivi par CWGopt+ pour des petits $N$ et par SS+ pour des grands $N$).

Dans les situations où $Index = 1$, l'estimateur FCV est parmi les meilleurs (il est équivalent à CWGopt+ et à SS+). Il devient encore plus performant quand $Index = 2$ (où il est le meilleur).

### 4.4.2. Conclusions de la simulation $t_{p,2}$

Dans cette simulation les comportements des estimateurs (K) restent semblables aux comportements observés dans la simulation précédente. Les estimateurs OPT et KS+ sont toujours supérieurs à MC et inférieurs aux autres. Ici, l'estimateur CWG+ est moins efficace. Il est supérieur à MC, à OPT et à KS+, mais inférieur à FCV, à CWGopt+ et à SS+. L'estimateur SS+ reste efficace dans les mêmes situations que dans la simulation précédente sauf que maintenant il devient strictement inférieur à FCV. Les estimateurs SS+ et CWGopt+ sont les meilleurs parmi les estimateurs (K) et inférieurs à FCV. Voir la figure suivante qui désigne la médiane des 3000 observations des critères $A_{(K)}$ en fonction de $q$ pour chacune des valeurs de $p$ séparément.

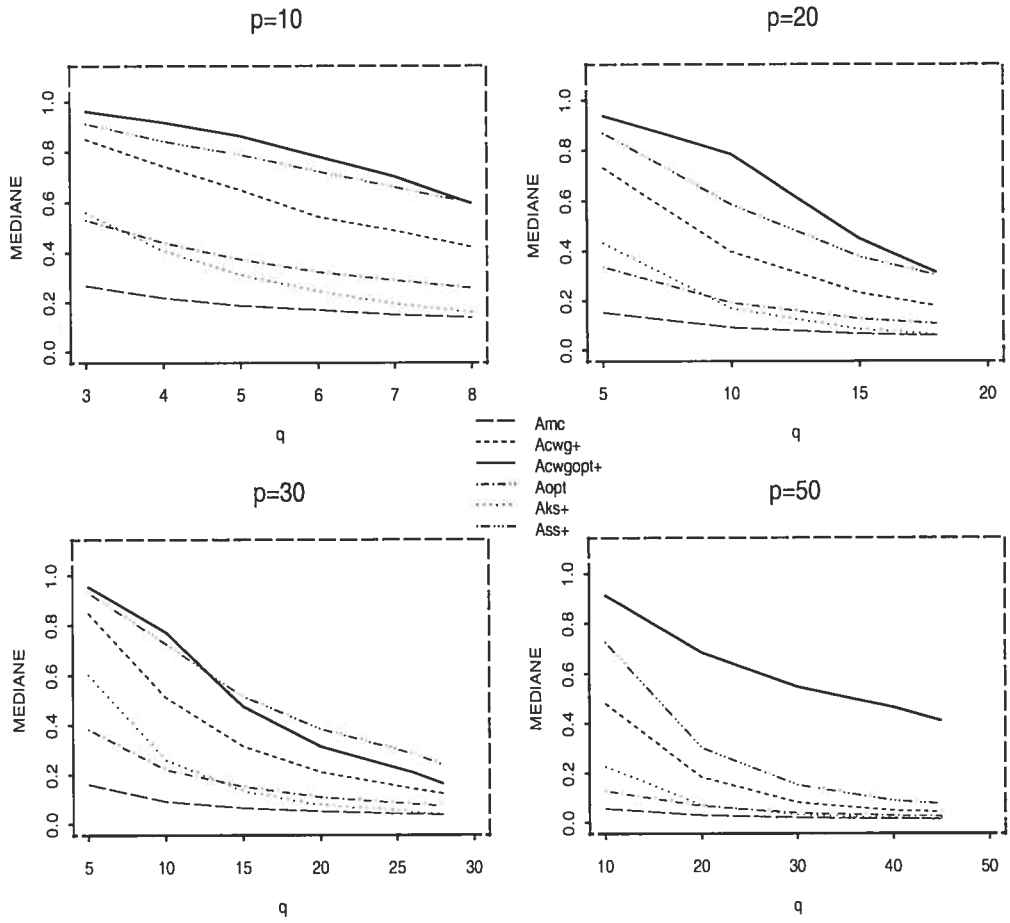



FIG. 4.4. La médiane des $A_{(K)}$ de la simulation t pour des différents $p$

L'estimateur CWGopt+ est le plus efficace pour les petits $q$ et les petites tailles $N$ — $p = 10,\ q = 3,4,\ N = 30$ (voir les boxplots à la page 48) ; $p = 20,\ q = 5,\ N = 50$ (page 50) et $p = 30,\ q = 5,\ N = 100$ (page 51) — où il est équivalent à FCV. Dans les situations $p = 50,\ q = 10, 20, 30,\ N = 100$ (voir les boxplots à la page 54) il est légèrement inférieur à FCV, selon le critère $A_{(CWGopt+)}$, mais avec $W_{(CWGopt+)} \gg 1$. Dans toutes les autres situations il est strictement inférieur à l'estimateur FCV.

L'estimateur FCV est le meilleur. Dans cette simulation il nous montre une meilleure efficacité. Comparons par exemple (dans la figure 4.5) les résultats obtenus des deux simulations dans la situation $p = 10,\ N = 30, 50$ et 100, où on trouve le plus petit écart avec les autres estimateurs. La figure 4.5-gauche (données de la loi normale) et la figure 4.5-droite (données de la loi t) désignent la médiane des 3000 observations des critères $A_{(K)}$ pour cette situation. Il est évident que l'estimateur FCV se comporte mieux pour les données des la loi $t_{10,2}$.

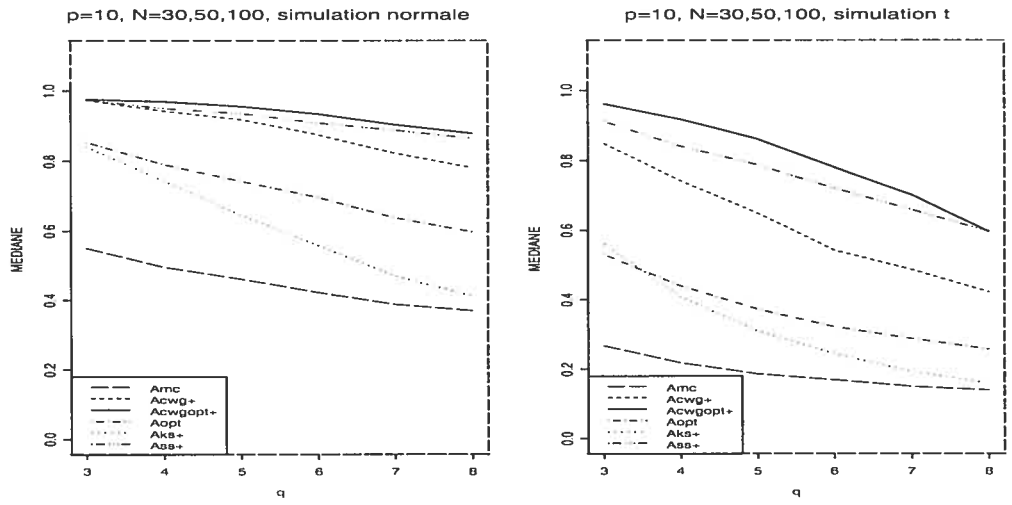



FIG. 4.5. La médiane des $A_{(K)}$ ; situations $p = 10,\ N = 30, 50, 100$ de la simulation normale (à gauche) et de la simulation t (à droite)

Dans la simulation $t_{p,2}$, on trouve les mêmes tendances du comportement de l'estimateur FCV. Son efficacité augmente avec la croissance du nombre des variables réponse $q$ (si $p$ est fixé) ou, en général, avec le nombre des variables $p+q$ (voir la figure 4.4). Ici, il dépend moins de la valeur de $Index$, mais il est plus efficace si $Index = 2$. Il reste sans aucune concurrence dans les cas des grands $q$

et $Index = 2$ (voir les boxplots des situations $p = 10,\ q = 5, 6, 7, 8,\ Index = 2$ à la page 49 ; $p = 20,\ q = 10, 15, 18,\ Index = 2$ à la page 50 et $p = 30,\ q = 15, 20, 25, 28,\ Index = 2$ à la page 52).

Un désavantage de l'estimateur FCV est qu'il n'estime pas les coefficients de façon équilibrée. Dans les deux simulations, les critères $W_{(K)}$ prennent toujours des valeurs plus grandes que 1 même si $A_{(K)} << 1$ et FCV est le meilleur. Dans une telle situation (où $A_{(K)} < 1$ et $W_{(K)} > 1$), il y a (au moins) une variable pour laquelle l'erreur d'estimation $e_j(FCV)^2$ est beaucoup plus grande que les erreurs $e_j(K)^2$ donnés par (K), tandis que les autres variables sont très bien estimées par FCV. Cette tendance existe dans les deux simulations, mais elle est beaucoup plus légère pour les données de la loi $t_{p,2}$ — les valeurs maximales de $W_{(K)}$ sont plus petites.

En conclusion, on peut dire que l'estimateur FCV est le meilleur choix quand nous devons prédire plusieurs variables réponse d'une distribution non-normale.

# Chapitre 5

# LES EXEMPLES

## 5.1. CHIMIOMÉTRIE

Considérons les données de Skagerberg, MacGregor et Kiparissides [**9**]. Cet exemple est étudié également par Breiman et Friedman [**3**] et Srivastava et Solanky [**10**]. Il y a $p = 22$ variables explicatives, $q = 6$ variables dépendantes et $N = 56$ observations. Les variables ne proviennent pas de la distribution normale et nous considérons certaines transformations. Sur les variables dépendantes on applique une transformation logarithmique et ensuite elles sont standardisées. Sur les variables explicatives on considère deux transformations : la transformation logarithmique $\ln(X + 0.03)$ et la transformation $\sqrt{X + 0.03}$ (qui donne les meilleurs résultats). Les résultats complets sont mis dans l'annexe 2. La matrice de corrélation entre les variables dépendantes (voir le tableau 5.1) nous suggère une corrélation forte entre $y_1$ et $y_2$, ainsi que entre $y_4, y_5$ et $y_6$.

| | $y_1$ | $y_2$ | $y_3$ | $y_4$ | $y_5$ | $y_6$ |
|---|---|---|---|---|---|---|
| $y_1$ | 1.000 | 0.957 | 0.065 | 0.254 | 0.255 | 0.259 |
| $y_2$ | 0.957 | 1.000 | -0.128 | 0.282 | 0.266 | 0.276 |
| $y_3$ | 0.065 | -0.128 | 1.000 | -0.500 | -0.484 | -0.479 |
| $y_4$ | 0.254 | 0.282 | -0.500 | 1.000 | 0.974 | 0.978 |
| $y_5$ | 0.255 | 0.266 | -0.484 | 0.974 | 1.000 | 0.976 |
| $y_6$ | 0.259 | 0.276 | -0.479 | 0.978 | 0.976 | 1.000 |

TAB. 5.1. Les corrélations entre les variables réponse de l'exemple Chimiométrie

Comme dans les deux articles, les erreurs de prédiction sont calculées par la méthode CV, à savoir en enlevant successivement chacune des 56 observations et en calculant chaque fois les estimateurs sur les 55 observations restantes. Les erreurs de prédiction sont calculées pour cette observation :

$$\mathbf{e}_{\backslash n}(K) = \mathbf{y}_n - \widetilde{A}_{(K)}\mathbf{x}_n,\ n = 1, \ldots, 56,$$

où $(K)$ est une des méthodes MC, CWG+, CWGopt+, OPT, KS+ et SS+. On prend finalement les moyennes

$$\mathbf{e}(K) = \frac{1}{56}\sum_{n=1}^{56}\mathbf{e}_{\backslash n}(K).$$

Les calculs sont faits pour chaque transformation des données $(Y, X)$ ; $(Y, \ln(X + 0.03))$ et $(Y, \sqrt{X + 0.03})$ et les meilleurs résultats sont obtenus pour $(Y, \sqrt{X + 0.03})$ (voir le tableau 5.2).

| | $y_1$ | $y_2$ | $y_3$ | $y_4$ | $y_5$ | $y_6$ | moyenne |
|---|---|---|---|---|---|---|---|
| MC | 0.0980 | 0.3334 | 0.1974 | 0.0967 | 0.2290 | 0.1714 | 0.1876 |
| CWG+ | 0.1489 | 0.1898 | 0.1923 | 0.0965 | 0.2113 | 0.1575 | 0.1661 |
| CWGopt+ | 0.1524 | 0.1844 | 0.2292 | 0.0985 | 0.2119 | 0.1551 | 0.1719 |
| OPT+ | 0.1208 | 0.2423 | 0.1885 | 0.0947 | 0.2092 | 0.1581 | 0.1689 |
| KS+ | 0.1331 | 0.2202 | 0.1932 | 0.0965 | 0.2093 | 0.1596 | 0.1687 |
| SS+ | 0.1491 | 0.1882 | 0.1924 | 0.0969 | 0.2104 | 0.1562 | 0.1655 |
| FCV | 0.1876 | 0.1661 | 0.1719 | 0.1689 | 0.1687 | 0.1655 | 0.1715 |

TAB. 5.2. Les erreurs de prédiction par CV pour $(Y, \sqrt{X + 0.03})$ de l'exemple Chimiométrie

L'estimateur FCV a la meilleure performance pour les variables $y_2, y_3$ et $y_5$, mais en moyenne le meilleur estimateur est SS+. Ceci est en accord avec les résultats des simulations où $SS+$ est un des meilleurs estimateurs pour des petites valeurs de $q$.

Les résultats de l'annexe 2 révèlent un phénomène important relié à la sélection de modèles. La variable réponse étant la même dans les trois analyses selon le choix des régresseurs $X$, $\ln(X + 0.03)$ et $\sqrt{X + 0.03}$, on observe que la méthode

classique MC par le choix du modèle $\sqrt{X + 0.03}$ procure une réduction d'environ 50% des erreurs de prévision comparativement à toutes les méthodes de prévision pour le choix du modèle original $X$. L'effet du choix de modèle est beaucoup plus probant que le choix de la méthode de prévision.

## 5.2. Pulp-Fiber

Considérons l'exemple étudié par Rousseeuw, van Aelst,van Driessen et Agulló [8] (les données sont prises de Lee [7]). Il y a $p = 4$ variables explicatives, $q = 4$ variables dépendantes et $N = 62$ observations. Dans cet article les auteurs cherchent les valeurs aberrantes, tandis que nous voulons améliorer les estimations faites sur toutes les observations. La matrice de corrélation entre les variables dépendantes (voir le tableau 5.3) nous montre que toutes les variables sont très corrélées.

| | $y_1$ | $y_2$ | $y_3$ | $y_4$ |
|---|---|---|---|---|
| $y_1$ | 1.000 | 0.914 | 0.984 | 0.988 |
| $y_2$ | 0.914 | 1.000 | 0.942 | 0.875 |
| $y_3$ | 0.984 | 0.942 | 1.000 | 0.975 |
| $y_4$ | 0.988 | 0.875 | 0.975 | 1.000 |

Tab. 5.3. Les corrélations entre les variables réponse de l'exemple Pulp-Fiber

Les erreurs de prédiction sont calculées par la méthode CV (comme dans l'exemple précédent). Les fortes corrélations entre les variables réponse, la condition $p = q$ et les données qui ne sont pas normales sont les conditions sous lesquelles l'estimateur FCV se comporte le mieux. Il est le meilleur estimateur en moyenne (voir le tableau 5.4), ainsi que pour toutes variables sauf $y_4$.

| | $y_1$ | $y_2$ | $y_3$ | $y_4$ | moyenne |
|---|---|---|---|---|---|
| MC | 2.5413 | 0.1436 | 0.4844 | 0.1382 | 0.8269 |
| CWG+ | 2.5322 | 0.1420 | 0.4822 | 0.1384 | 0.8237 |
| CWGopt+ | 2.5309 | 0.1418 | 0.4819 | 0.1384 | 0.8233 |
| OPT+ | 2.5285 | 0.1423 | 0.4822 | 0.1380 | 0.8228 |
| KS+ | 2.5314 | 0.1427 | 0.4832 | 0.1377 | 0.8238 |
| SS+ | 2.5305 | 0.1412 | 0.4813 | 0.1385 | 0.8229 |
| FCV | 2.5082 | 0.1406 | 0.4795 | 0.1377 | 0.8165 |

Tab. 5.4. Les erreurs de prédiction par CV de l'exemple Pulp-Fiber

Une fois que les données sont rendues normales, l'estimateur FCV n'est plus le meilleur en moyenne. À titre de comparaison, nous avons donné dans l'annexe 2 les erreurs de prédiction calculées après une transformation Box-Cox.

Ici encore nous pouvons observer qu'une simple transformation de Box-Cox (la variable réponse demeurant toujours la même) permet à la méthode MC de réduire d'environ 10% les erreurs de prévision par rapport à toutes les autres méthodes appliquées sans aucune transformation. La sélection d'un autre modèle permet encore une fois d'obtenir la plus grande amélioration dans les erreurs de prévision.

# BIBLIOGRAPHIE


[1] BILODEAU, M. (2000), *Multivariate flattening for better predictions*, Canad. J. Statist., 28(1), 159-170.

[2] BILODEAU, M., KARIYA,T (1989), *Minimax estimators in the normal MANOVA model*, J. Multivariate Anal., 28(2), 260-270.

[3] BREIMAN, L., FRIEDMAN, J. H. (1997), *Predicting multivariate responses in multiple linear regression*, J. Roy. Statist. Soc., Ser. B, 59(1), 3-54.

[4] HOERL, A. E., KENNARD, R. W. (1970), *Ridge regression : biased estimation for nonorthogonal problems*, Technometrics, 8, 27-51.

[5] IZENMAN, A. J. (1975), *Reduced-rank regression for multivariate linear model*, J. Multivariate Anal., 5, 248-264

[6] KONNO, Y. (1991), *On estimation of a matrix of normal means with unknown covariance matrix*, J. Multivariate Anal., 36(1), 44-55

[7] LEE, J. (1992), *Relationships between properties of pulp-fiber and paper*, Ph. D. thesis, University of Toronto, Faculty of Forestry.

[8] ROUSSEEUW, P. J., VAN DRIESSEN, K., VAN AELST, S., AGULLÓ, J. (2004), *Robust multivariate regression*, Technometrics, 46(3), 293-305.

[9] SKAGERBERG, B., MACGREGOR, J., KIPARISSIDES, C. (1992), *Multivariate data analysis applied to low-density polyethylene reactors*, Chemometr. Intell. Lab. Syst., 14, 341-356.

[10] SRIVASTAVA, M. S., SOLANKY, T. K. S. (2003), *Predicting multivariate response in linear regression model*, Comm. Statist. Simulation Comput., 32(2), 389-409.

[11] VAN DER MERWE, A., ZIDEK, J. V. (1980), *Multivariate regression analysis and canonical variates*, Canad. J. Statist., 8(1), 27-39.

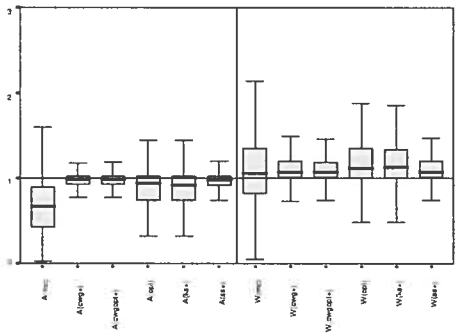
[12] WOLD, H. (1975), *Soft modelling by latent variables ; the non-linear iterative partial least squares (NIPALS) approach*, Dans Perspectives in probability and statistics

(papers in honour of M. S. Bartlett on the occasion of his 65th birthday), édités par J. Gani, Academic Press, New York, 117-142.

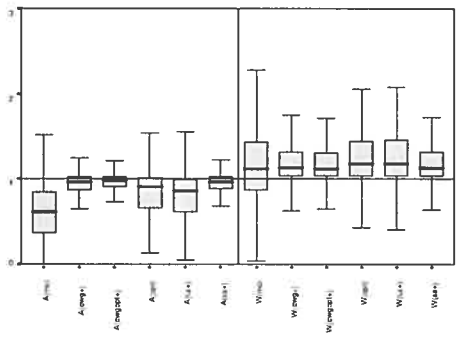
# .1. Annexe 1

## .1.1. Boxplots de la simulation normale

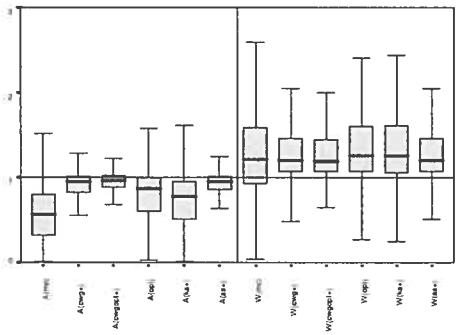
p=10 q=3



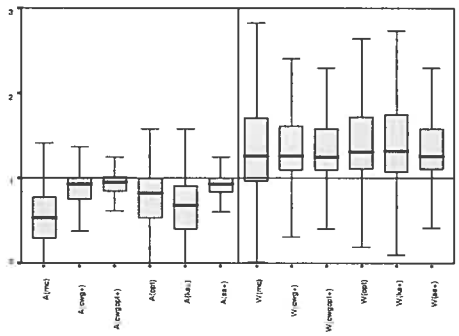
p=10 q=4



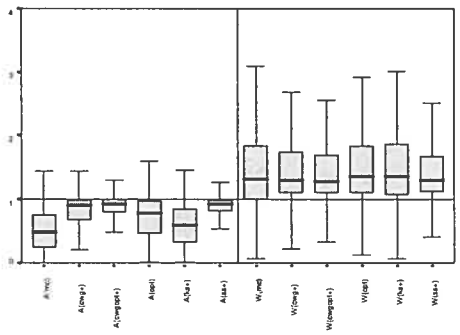
p=10 q=5



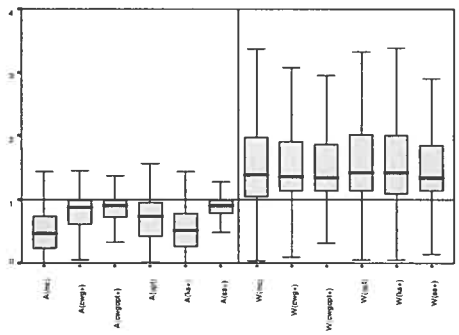
p=10 q=6



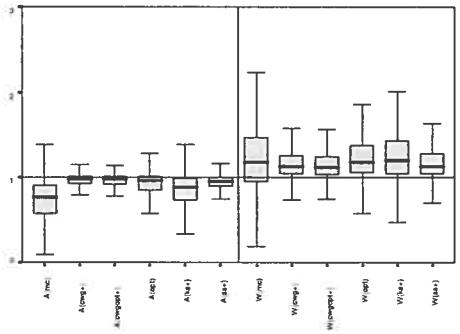
p=10 q=7



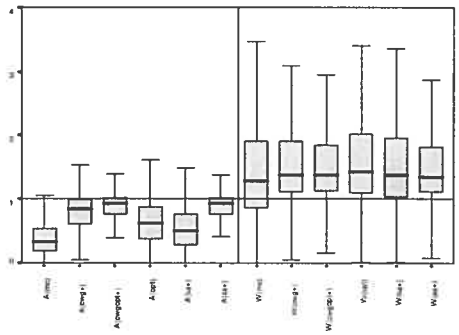
p=10 q=8



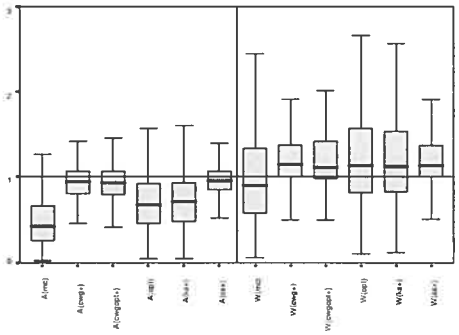
p=10 Index=1



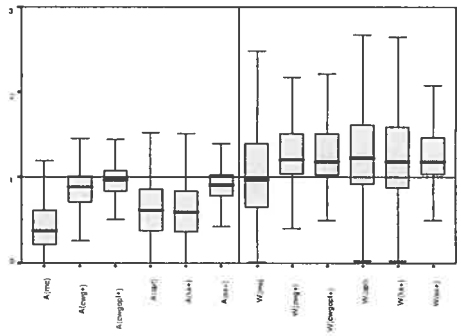
p=10 Index=2



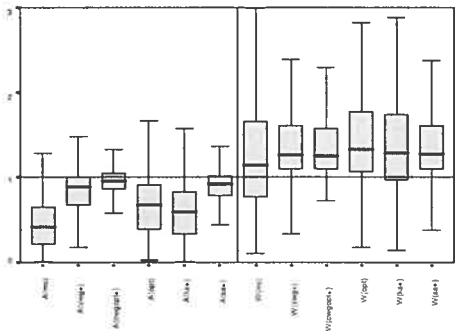
p=10 q=3 N=30



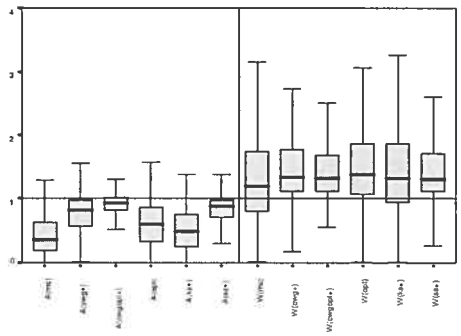
p=10 q=4 N=30



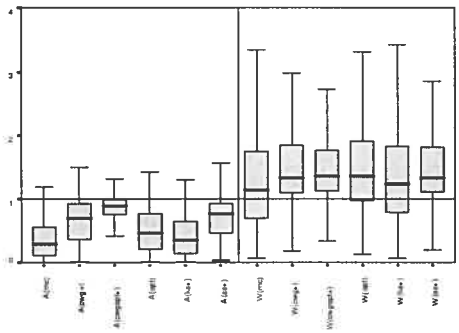
p=10 q=5 N=30,50



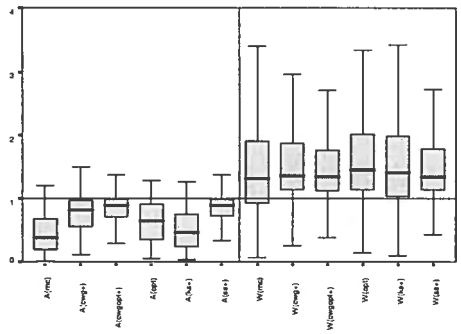
p=10 q=6 N=30,50



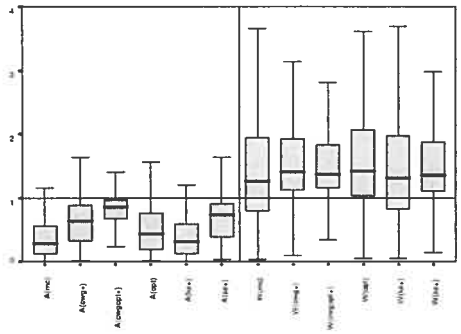
p=10 q=7 N=30



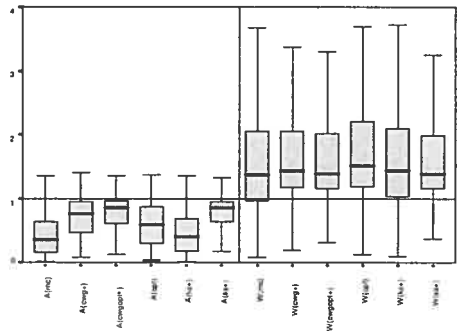
p=10 q=7 N=50



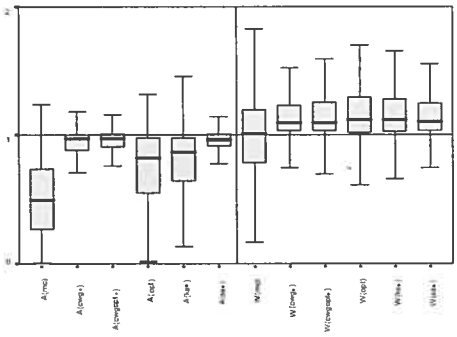
p=10 q=8 N=30



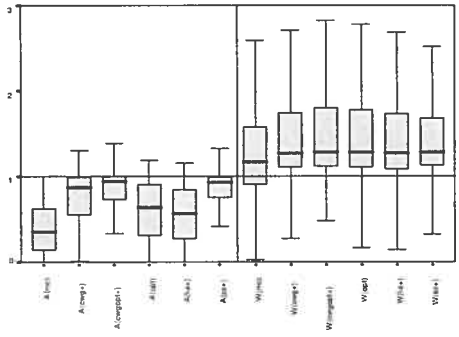
p=10 q=8 N=50



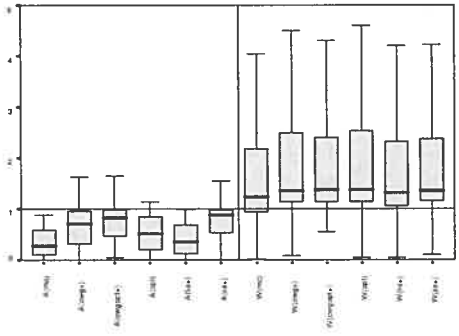
p=20 q=5



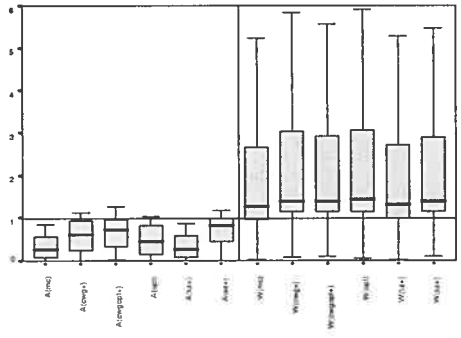
p=20 q=10



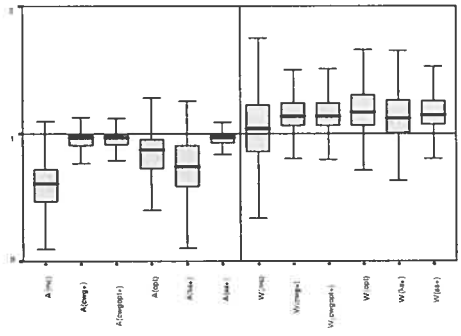
p=20 q=15



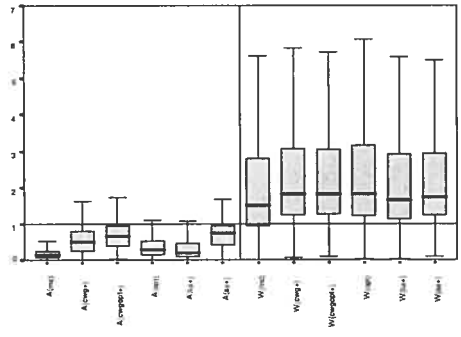
p=20 q=18



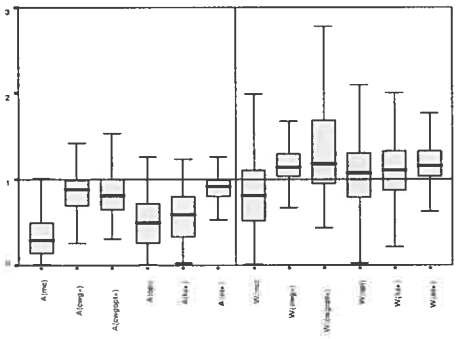
p=20 Index=1



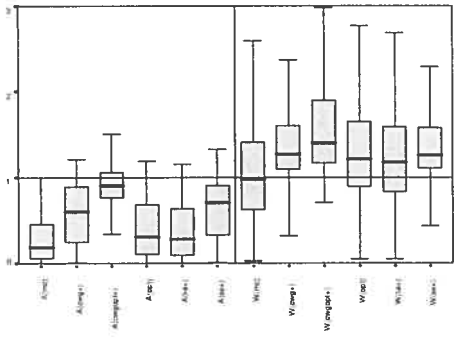
p=20 Index=2



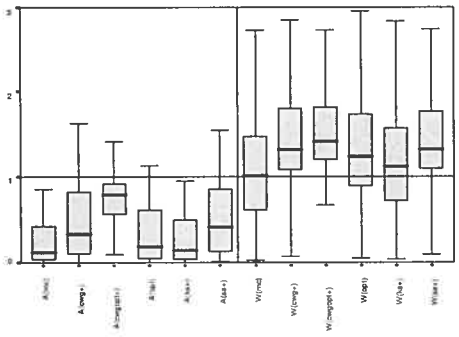
p=20 q=5 N=50



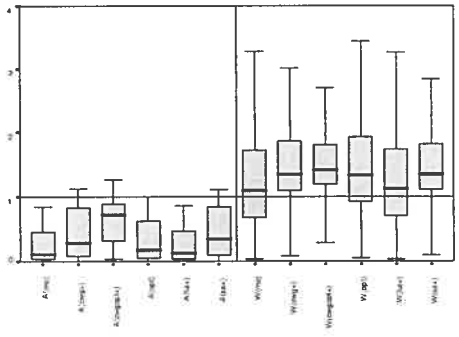
p=20 q=10 N=50



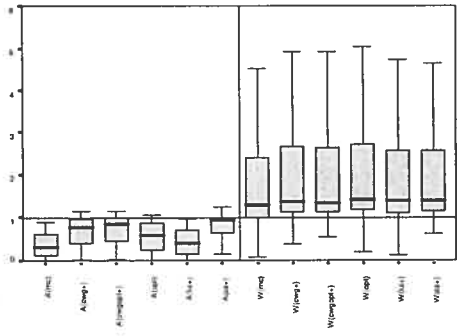
p=20 q=15 N=50



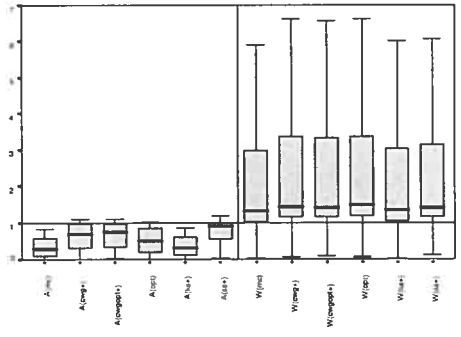
p=20 q=18 N=50



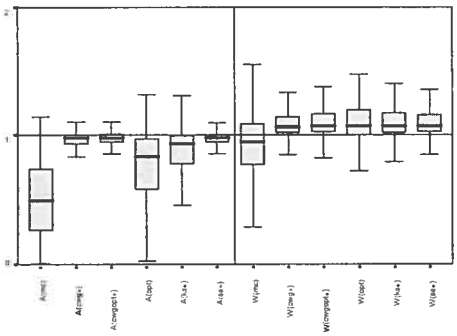
p=20 q=15 N>=100



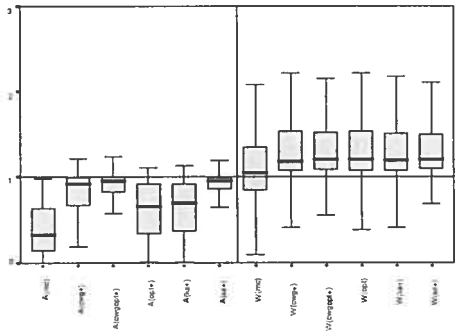
p=20 q=18 N>=100



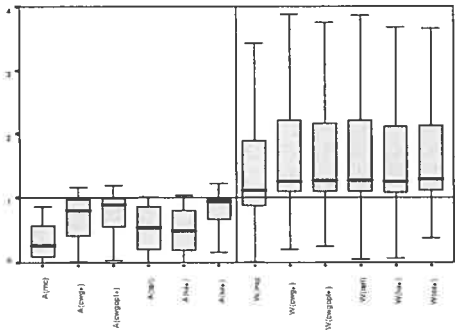
p=30 q=5



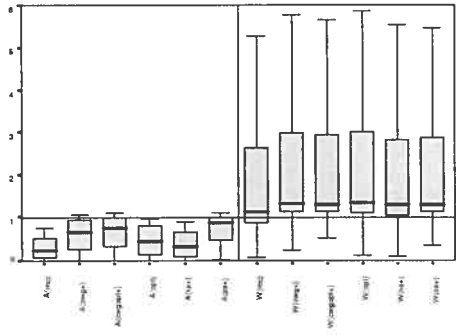
p=30 q=10



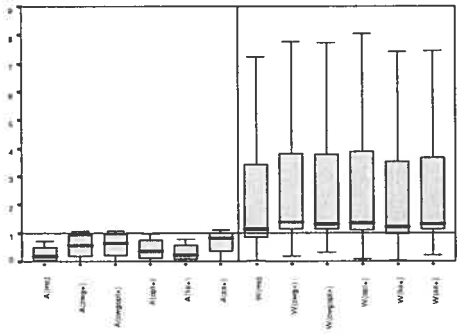
p=30 q=15



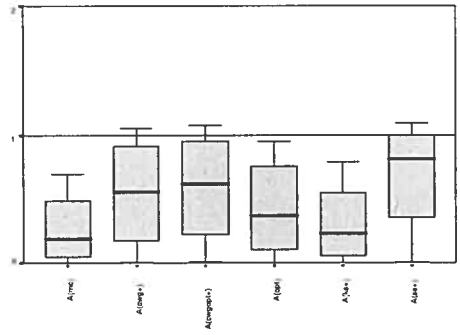
p=30 q=20



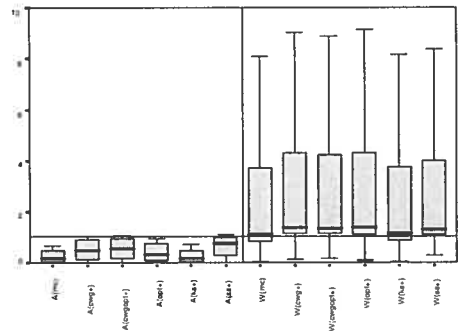
p=30 q=25



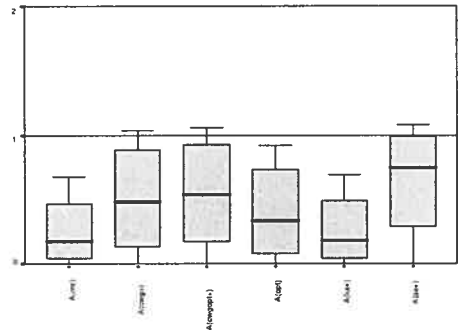
p=30 q=25



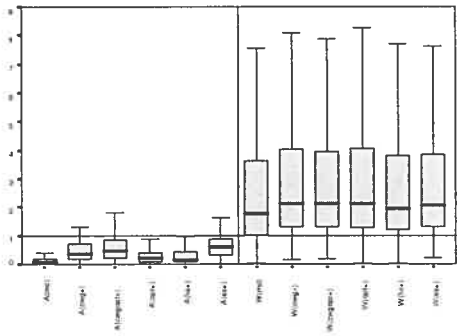
p=30 q=28



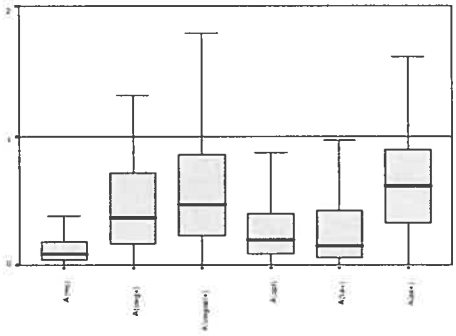
p=30 q=28



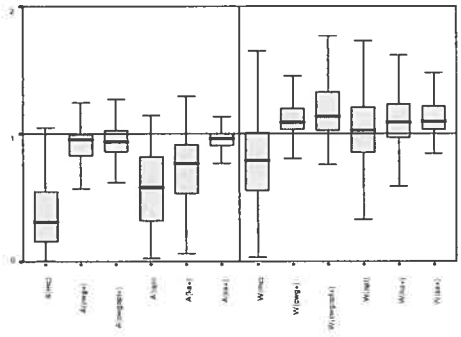
p=30 Index=2



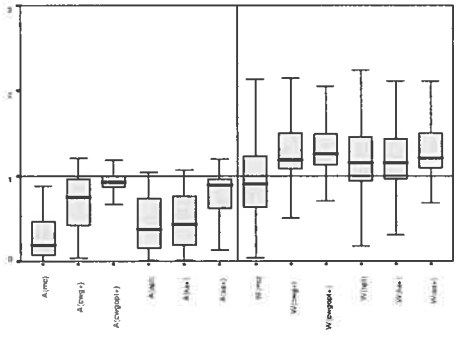
p=30 Index=2



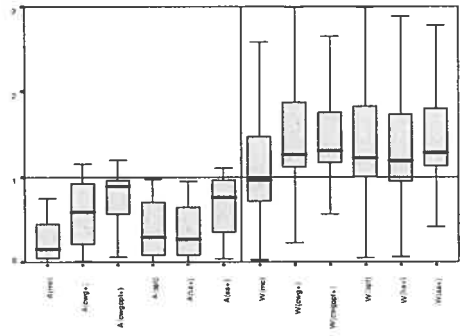
p=30 q=5 N=100



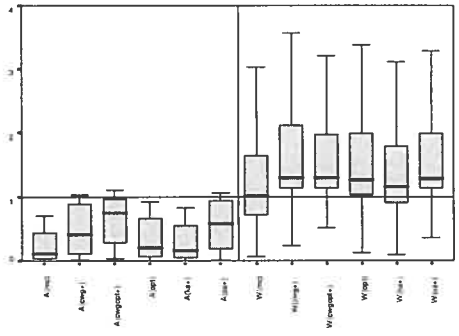
p=30 q=10 N=100



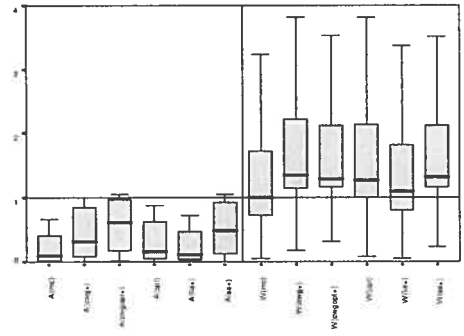
p=30 q=15 N=100



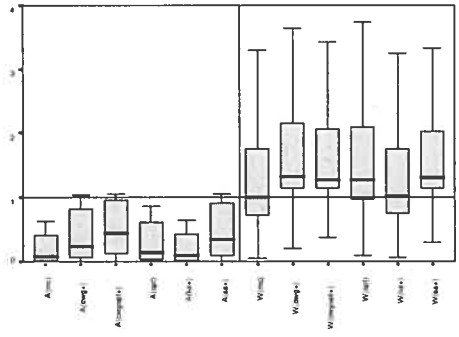
p=30 q=20 N=100



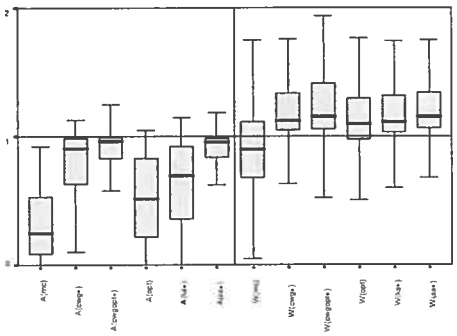
p=30 q=25 N=100



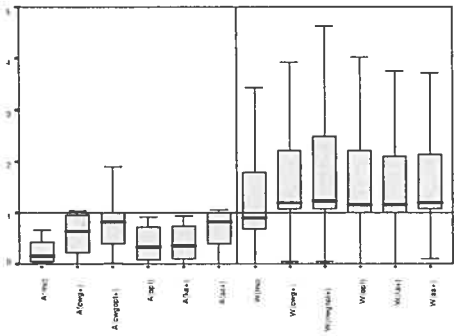
p=30 q=28 N=100



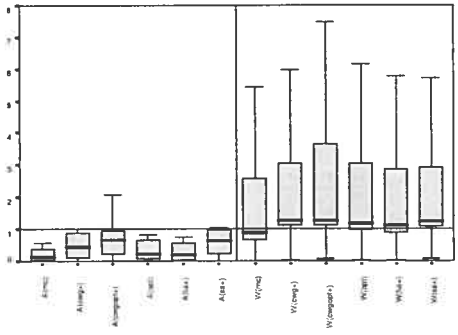
p=50 q=10



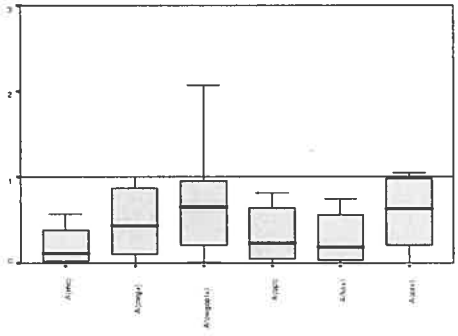
p=50 q=20



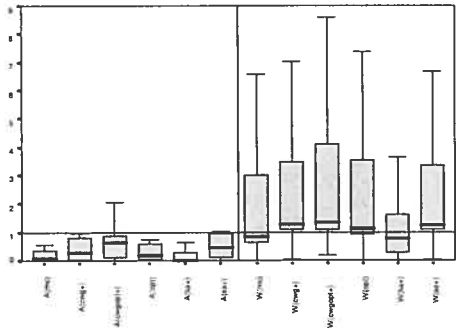
p=50 q=30



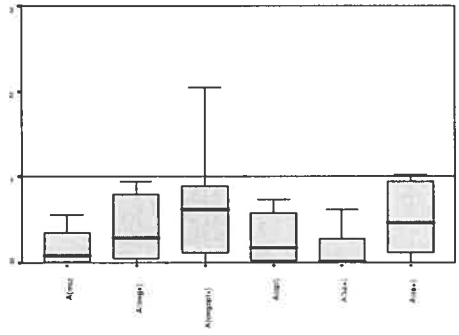
p=50 q=30



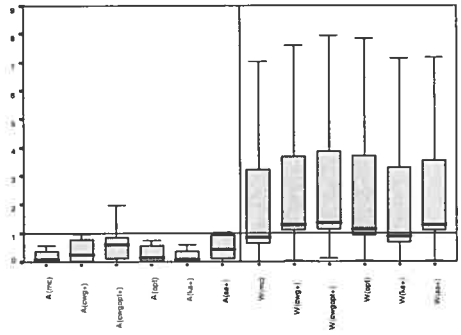
p=50 q=40



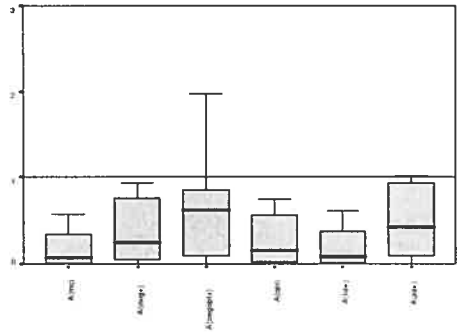
p=50 q=40



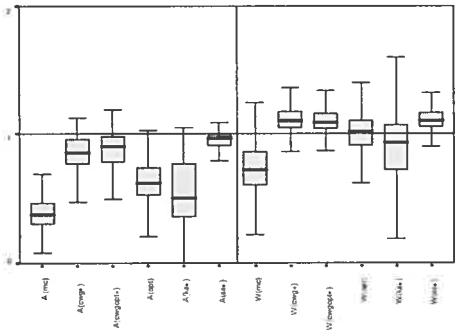
p=50 q=45



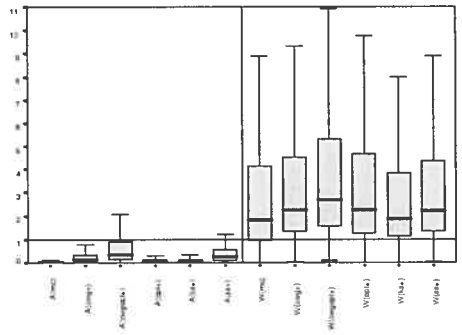
p=50 q=45



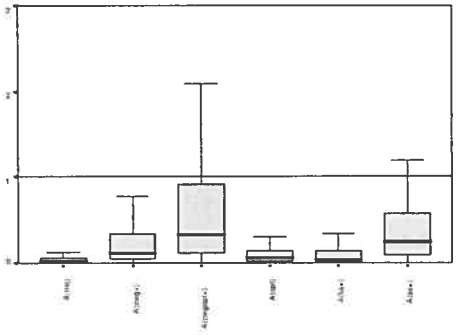
p=50 Index=1



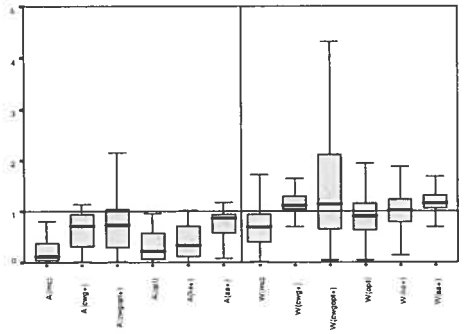
p=50 Index=2



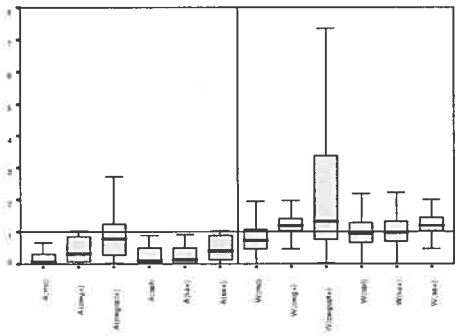
p=50 Index=2



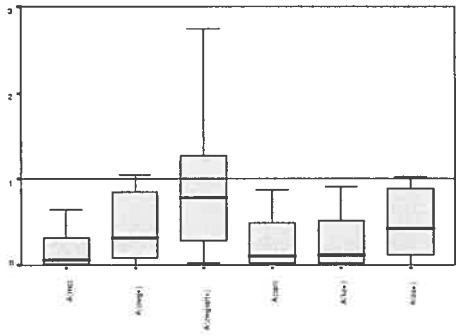
p=50 q=10 N=100



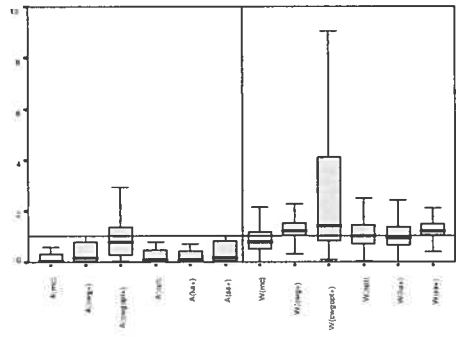
p=50 q=20 N=100



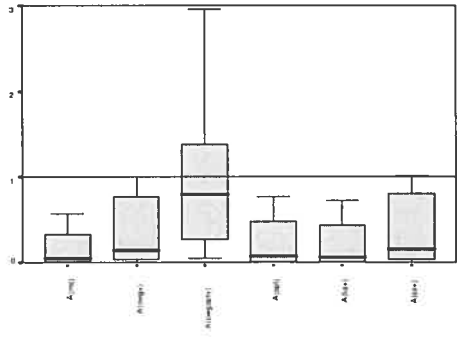
p=50 q=20 N=100



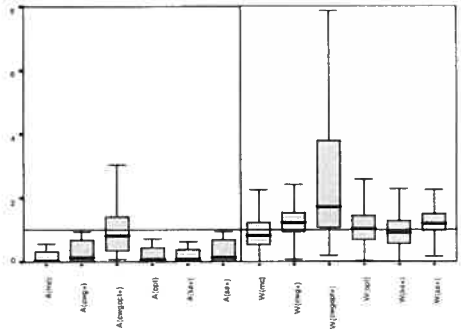
p=50 q=30 N=100



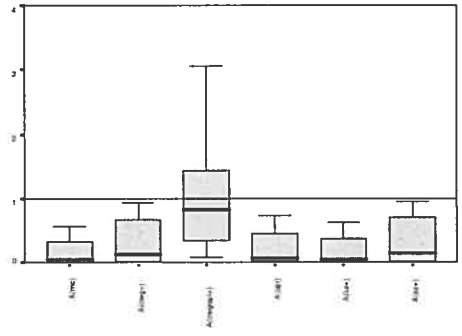
p=50 q=30 N=100



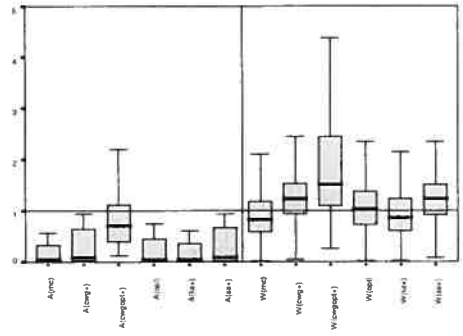
p=50 q=40 N=100



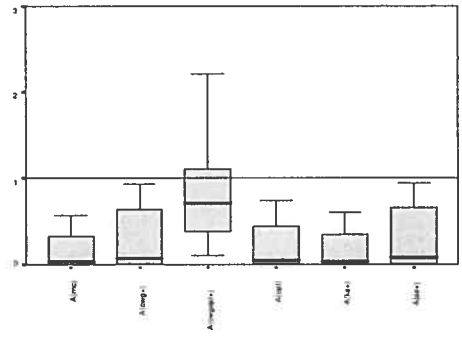
p=50 q=40 N=100



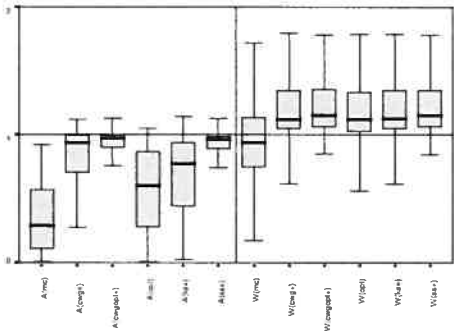
p=50 q=45 N=100



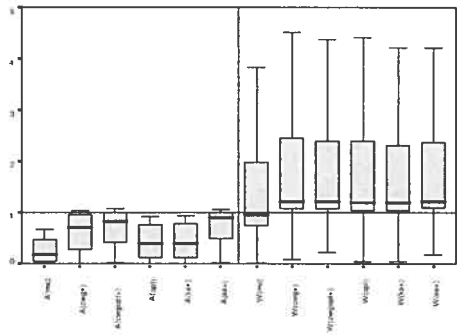
p=50 q=45 N=100



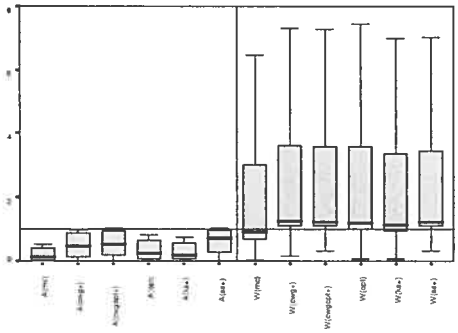
p=50 q=10 N=200,300,400,500



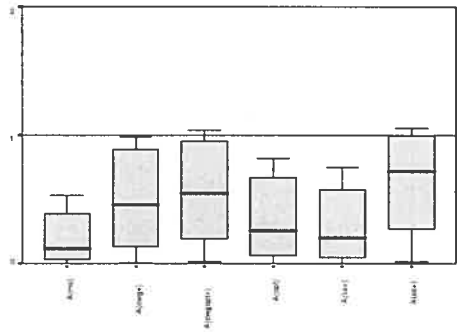
p=50 q=20 N=200,300,400,500



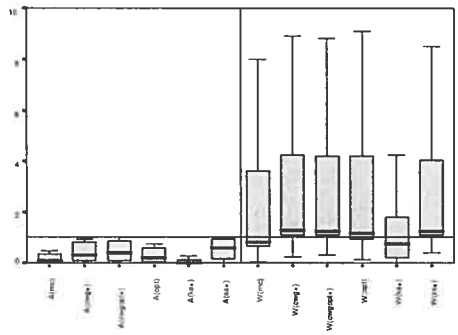
p=50 q=30 N=200,300,400,500



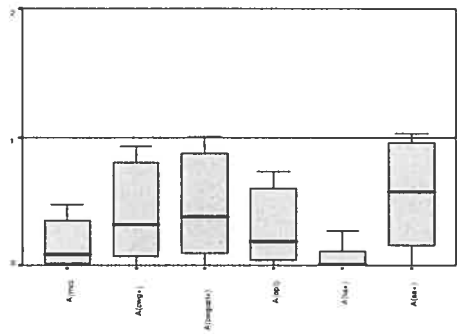
p=50 q=30 N=200,300,400,500



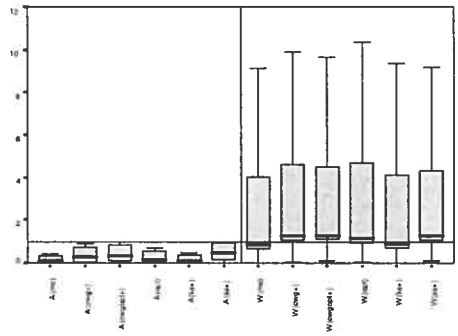
p=50 q=40 N=200,300,400,500



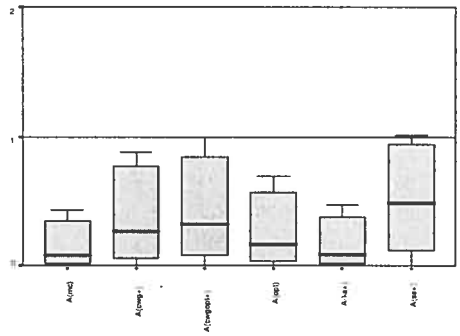
p=50 q=40 N=200,300,400,500



p=50 q=45 N=200,300,400,500

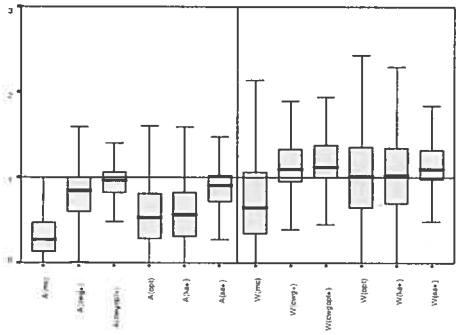


p=50 q=45 N=200,300,400,500

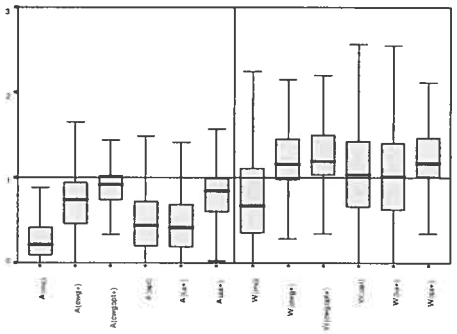


## .1.2. Boxplots de la simulation $t_{p,2}$

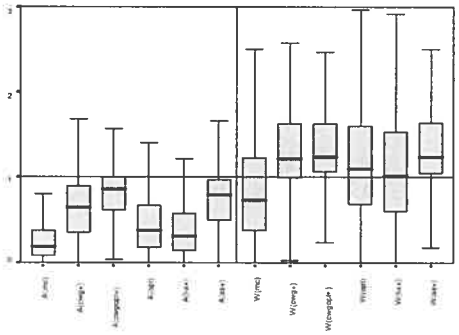
p=10 q=3



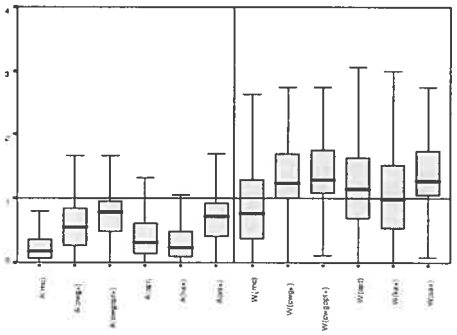
p=10 q=4



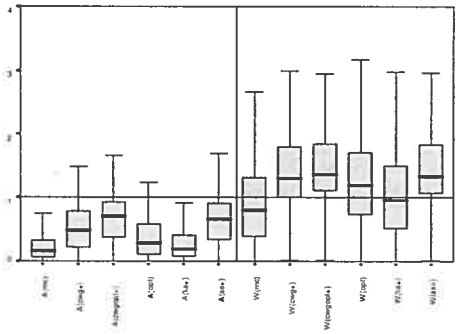
p=10 q=5



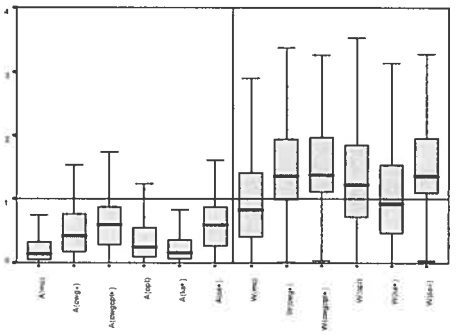
p=10 q=6



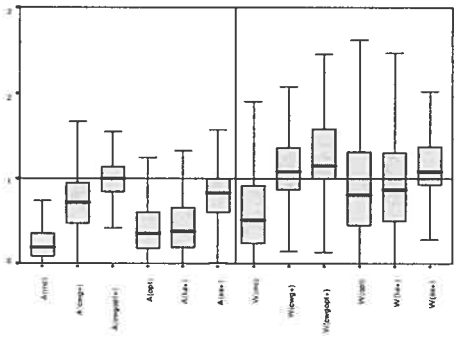
p=10 q=7



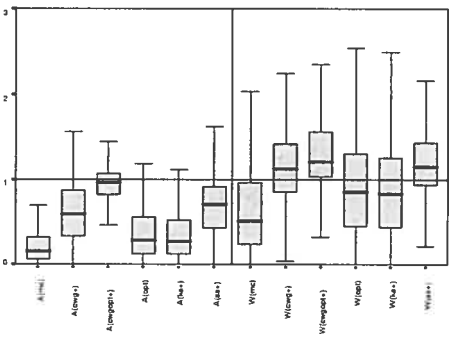
p=10 q=8



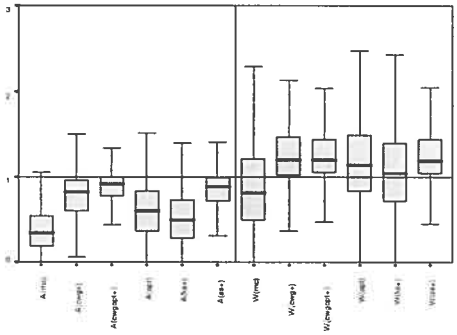
p=10 q=3 N=30



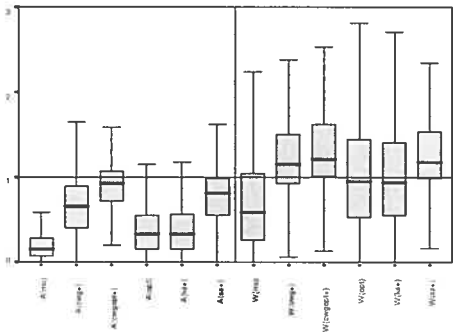
p=10 q=4 N=30



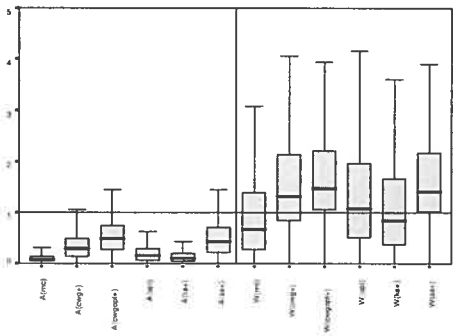
p=10 Index=1



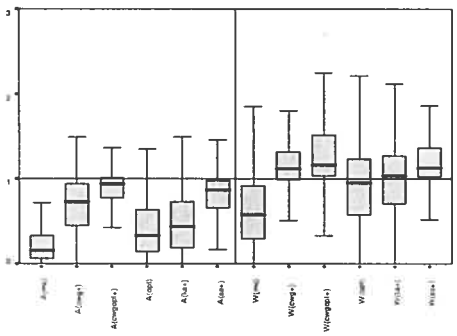
p=10 q=3,4 Index=2



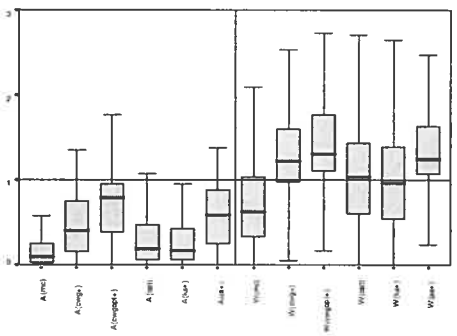
p=10 q=5,6,7,8 Index=2



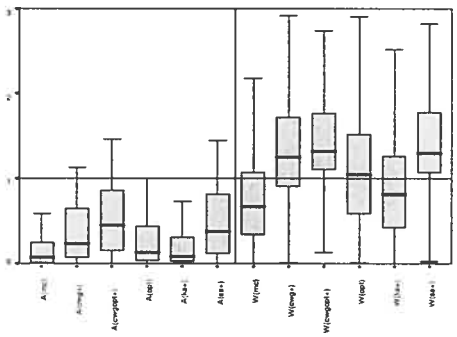
p=20 q=5



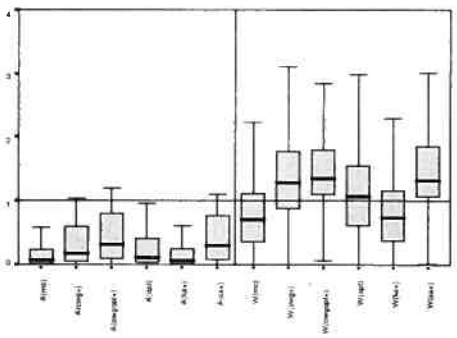
p=20 q=10



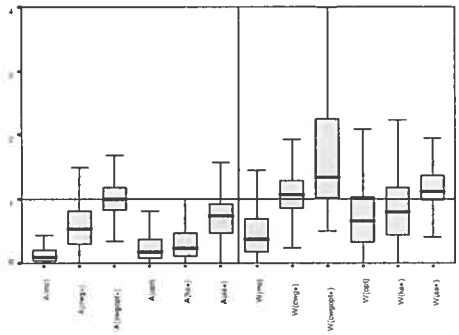
p=20 q=15



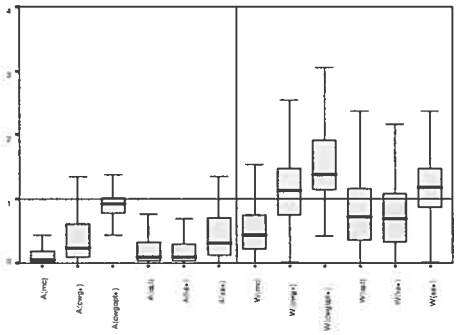
p=20 q=18



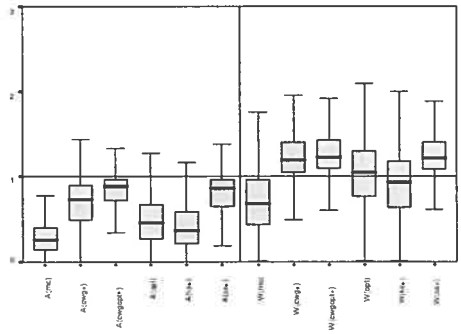
p=20 q=5 N=50



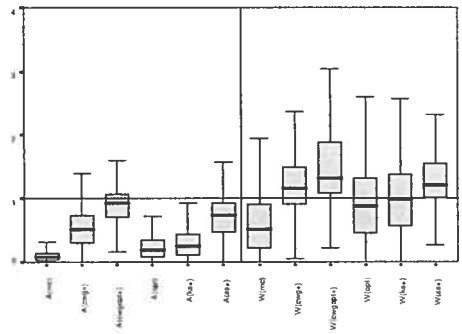
p=20 q=10 N=50



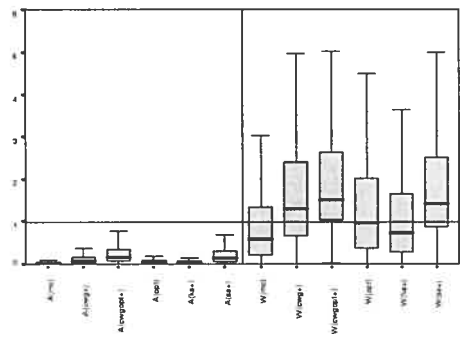
p=20 Index=1



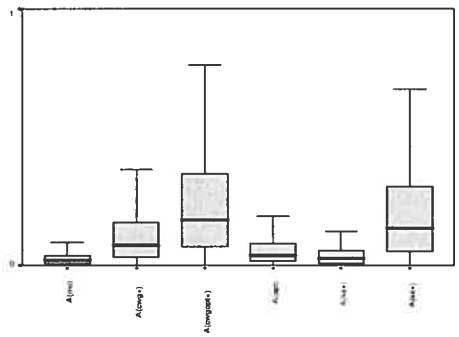
p=20 q=5 Index=2



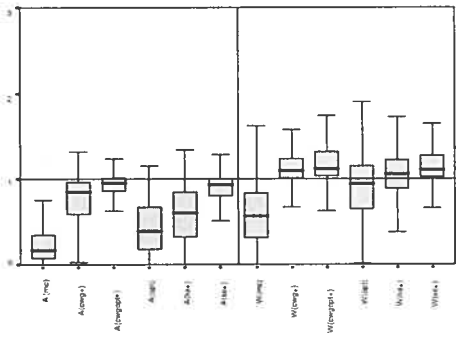
p=20 q=10,15,18 Index=2



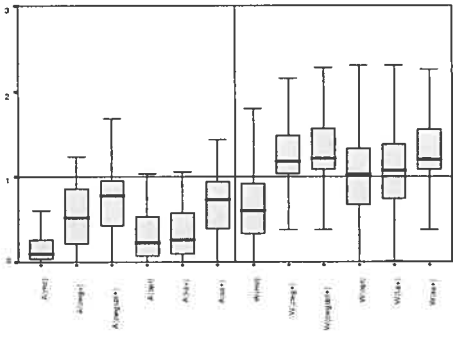
p=20 q=10,15,18 Index=2



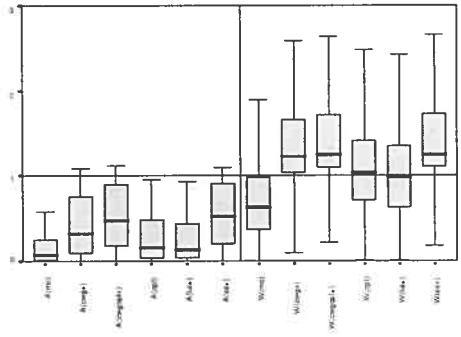
p=30 q=5



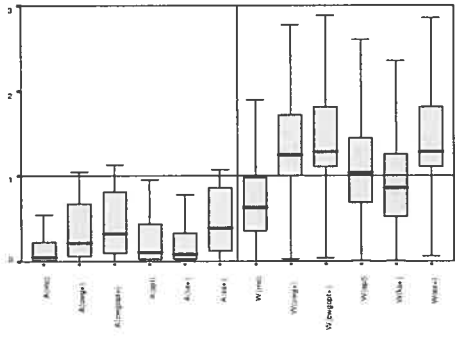
p=30 q=10



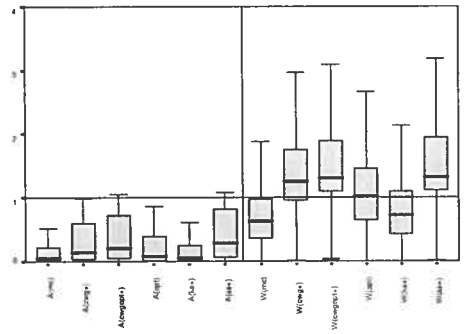
p=30 q=15



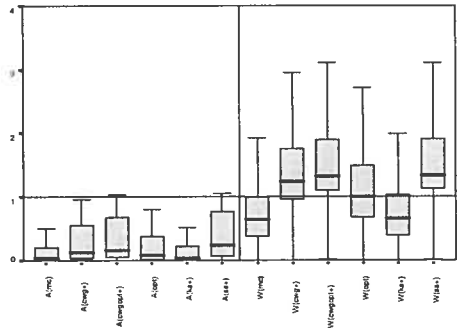
p=30 q=20



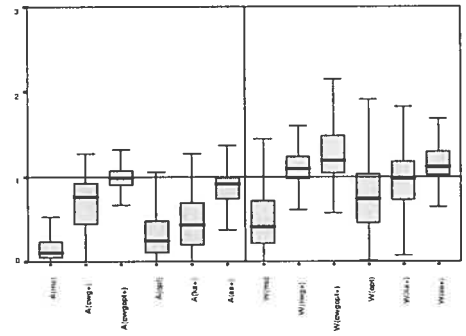
p=30 q=25



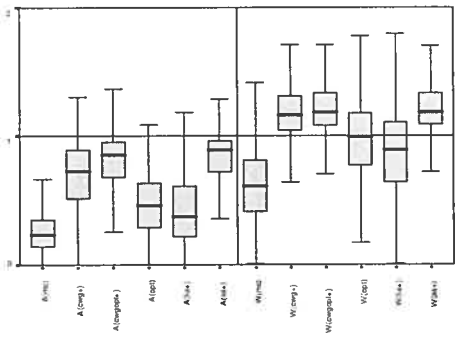
p=30 q=28



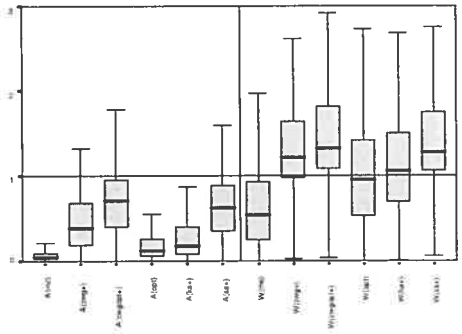
p=30 q=5 N=100



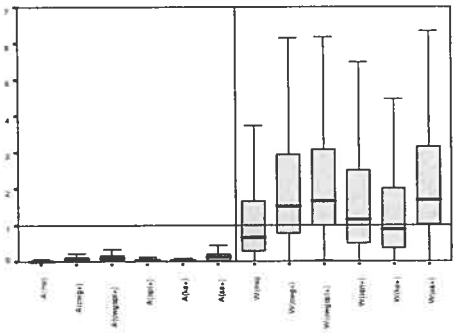
p=30 Index=1



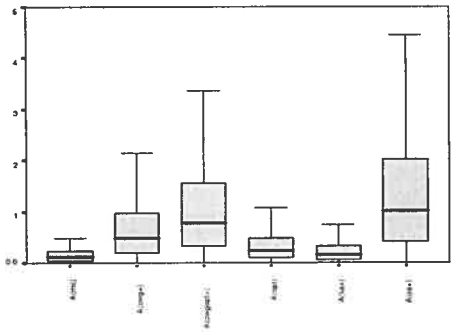
p=30 q=5,10 Index=2



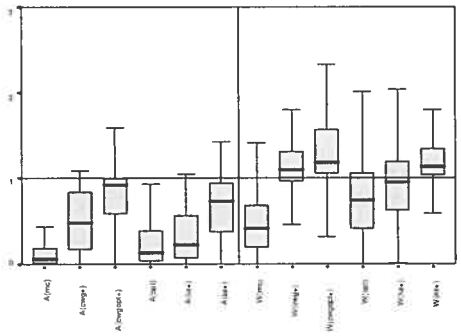
p=30 q=15,20,25,28 Index=2



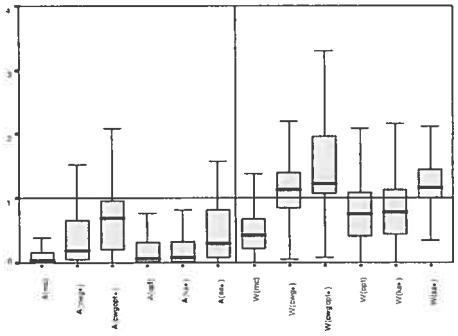
p=30 q=15,20,25,28 Index=2



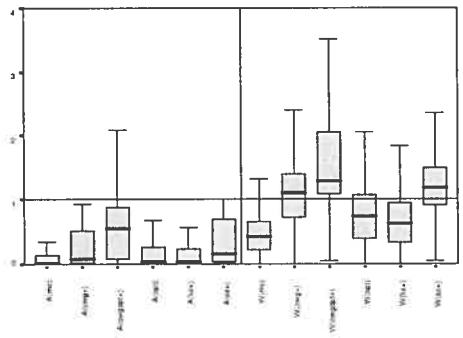
p=50 q=10



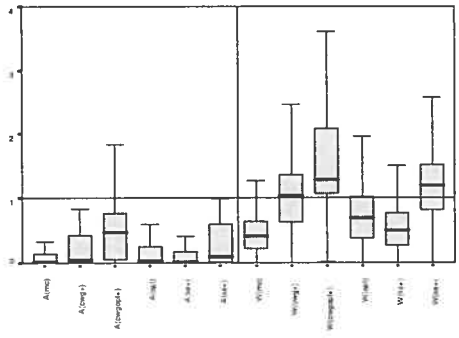
p=50 q=20



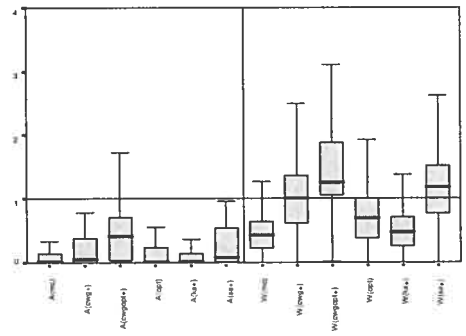
p=50 q=30



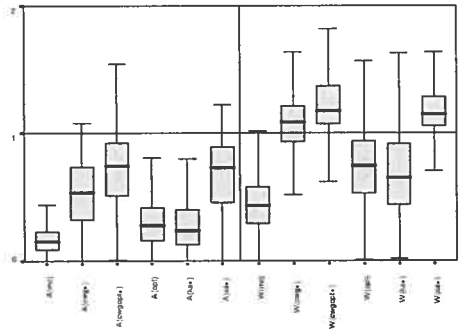
p=50 q=40



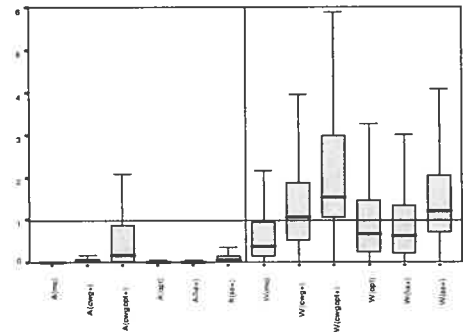
p=50 q=45



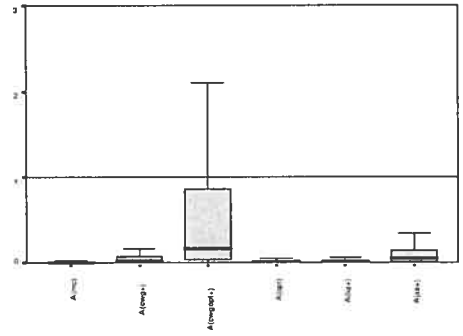
p=50 Index=1



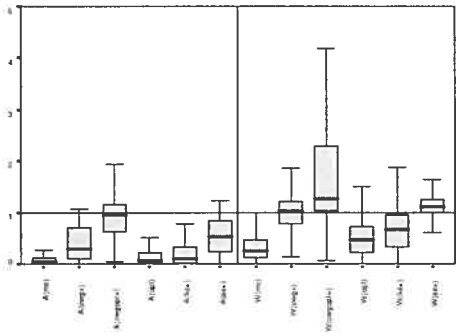
p=50 Index=2



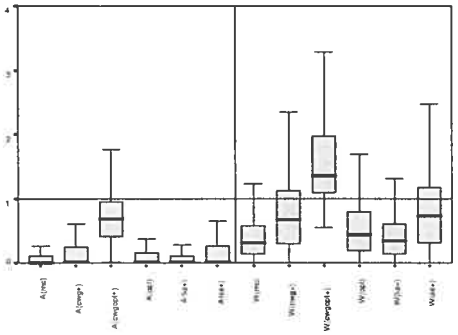
p=50 Index=2



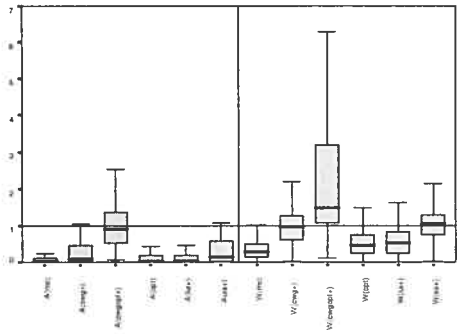
p=50 q=10 N=100



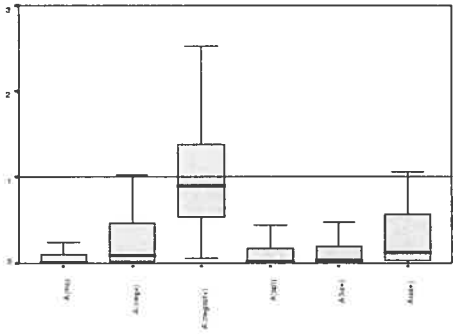
p=50 q=45 N=100



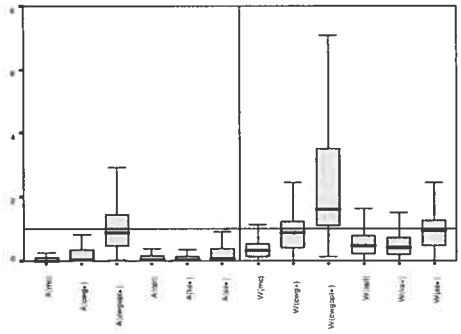
p=50 q=20 N=100



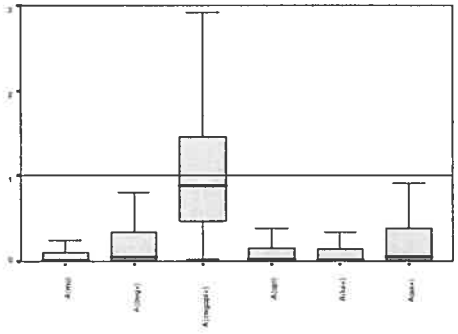
p=50 q=20 N=100



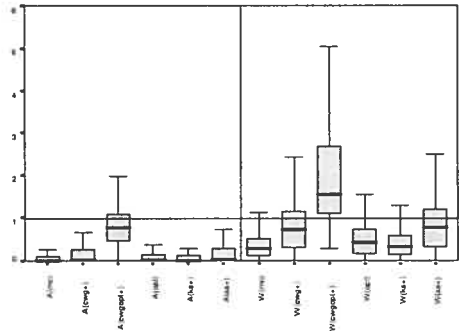
p=50 q=30 N=100



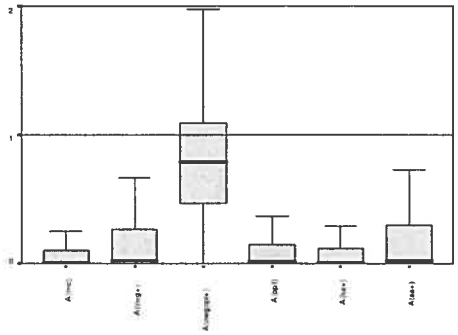
p=50 q=30 N=100



p=50 q=40 N=100



p=50 q=40 N=100



# .2. ANNEXE 2

## .2.1. Chimiométrie

**Les erreurs de prédiction par CV pour $(Y, X)$**

| | $y_1$ | $y_2$ | $y_3$ | $y_4$ | $y_5$ | $y_6$ | moyenne |
|---|---|---|---|---|---|---|---|
| MC | 0.5510 | 1.0991 | 0.2464 | 0.1187 | 0.2700 | 0.1812 | 0.4111 |
| CWG+ | 0.5646 | 0.7573 | 0.2199 | 0.1022 | 0.2315 | 0.1602 | 0.3393 |
| CWGopt+ | 0.5184 | 0.6437 | 0.2544 | 0.0985 | 0.2291 | 0.1578 | 0.3170 |
| OPT | 0.5562 | 0.8960 | 0.2241 | 0.1071 | 0.2393 | 0.1620 | 0.3641 |
| KS+ | 0.5589 | 0.8710 | 0.2261 | 0.1066 | 0.2378 | 0.1632 | 0.3606 |
| SS+ | 0.5666 | 0.7328 | 0.2205 | 0.0994 | 0.2293 | 0.1587 | 0.3345 |
| FCV | 0.4111 | 0.3393 | 0.3170 | 0.3641 | 0.3606 | 0.3345 | 0.3544 |

**Les erreurs de prédiction par CV pour $(Y, ln(X + 0.03))$**

| | $y_1$ | $y_2$ | $y_3$ | $y_4$ | $y_5$ | $y_6$ | moyenne |
|---|---|---|---|---|---|---|---|
| MC | 0.3077 | 0.4056 | 0.2007 | 0.1190 | 0.2287 | 0.2061 | 0.2447 |
| CWG+ | 0.3504 | 0.3953 | 0.1922 | 0.1222 | 0.2066 | 0.1696 | 0.2394 |
| CWGopt+ | 0.3580 | 0.3870 | 0.2202 | 0.1237 | 0.2050 | 0.1641 | 0.2430 |
| OPT | 0.3133 | 0.3956 | 0.1928 | 0.1186 | 0.2103 | 0.1797 | 0.2350 |
| KS+ | 0.3472 | 0.4038 | 0.1975 | 0.1226 | 0.2116 | 0.1804 | 0.2438 |
| SS+ | 0.3571 | 0.3990 | 0.1923 | 0.1224 | 0.2060 | 0.1676 | 0.2407 |
| FCV | 0.2447 | 0.2394 | 0.2430 | 0.2350 | 0.2439 | 0.2407 | 0.2411 |

**Les erreurs de prédiction par CV pour $(Y, \sqrt{X + 0.03})$**

| | $y_1$ | $y_2$ | $y_3$ | $y_4$ | $y_5$ | $y_6$ | moyenne |
|---|---|---|---|---|---|---|---|
| MC | 0.0980 | 0.3334 | 0.1974 | 0.09667 | 0.2290 | 0.1714 | 0.1876 |
| CWG+ | 0.1489 | 0.1898 | 0.1923 | 0.09647 | 0.2113 | 0.1576 | 0.1661 |
| CWGopt+ | 0.1524 | 0.1844 | 0.2292 | 0.09853 | 0.2119 | 0.1551 | 0.1719 |
| OPT | 0.1208 | 0.2423 | 0.1885 | 0.09467 | 0.2092 | 0.1581 | 0.1689 |
| KS+ | 0.1331 | 0.2202 | 0.1932 | 0.09647 | 0.2093 | 0.1597 | 0.1687 |
| SS+ | 0.1490 | 0.1882 | 0.1924 | 0.09686 | 0.2104 | 0.1562 | 0.1655 |
| FCV | 0.1876 | 0.1661 | 0.1719 | 0.16894 | 0.1687 | 0.1655 | 0.1715 |

### .2.2. Pulp-Fiber

Les variables explicatives prennent des valeurs négatives et la transformation Box-Cox est appliquée à (X+0.7). Les valeurs du paramètre de la transformation sont les $\lambda$ qui minimisent les fonctions de vraisemblance respectives pour chaque variable :

| var | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $y_1$ | $y_2$ | $y_3$ | $y_4$ |
|---|---|---|---|---|---|---|---|---|
| $\lambda_{max}$ | 1.2 | 1.4 | 0.6 | -1.3 | 1.0 | 1.4 | 0.9 | 1.0 |

Les paramètres pour les variables réponse sont proches de 1, on ne les transforme pas. On applique les transformations respectives sur les variables explicatives. Les erreurs de prédiction calculées par la méthode CV sur les données transformées sont :

**Les erreurs de prédiction par CV après transformation Box-Cox**

| | $y_1$ | $y_2$ | $y_3$ | $y_4$ | moyenne |
|---|---|---|---|---|---|
| MC | 2.2320 | 0.1392 | 0.4305 | 0.1266 | 0.7321 |
| CWG+ | 2.2144 | 0.1402 | 0.4256 | 0.1271 | 0.7268 |
| CWGopt+ | 2.2120 | 0.1401 | 0.4253 | 0.1269 | 0.7261 |
| OPT | 2.2083 | 0.1392 | 0.4266 | 0.1260 | 0.7250 |
| KS+ | 2.2179 | 0.1382 | 0.4289 | 0.1256 | 0.7277 |
| SS+ | 2.1975 | 0.1398 | 0.4231 | 0.1260 | 0.7216 |
| FCV | 2.2179 | 0.1400 | 0.4274 | 0.1273 | 0.7282 |