

REVUE DE STATISTIQUE APPLIQUÉE

J. BOULARAN

H. CAUSSINUS

Le modèle fonctionnel pour la régression sur variables entachées d'erreurs : étude générale et exemple

Revue de statistique appliquée, tome 42, n° 3 (1994), p. 63-88

http://www.numdam.org/item?id=RSA_1994__42_3_63_0

© Société française de statistique, 1994, tous droits réservés.

L'accès aux archives de la revue « Revue de statistique appliquée » (<http://www.sfds.asso.fr/publicat/rsa.htm>) implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques

<http://www.numdam.org/>

LE MODÈLE FONCTIONNEL POUR LA RÉGRESSION SUR VARIABLES ENTACHÉES D'ERREURS : ETUDE GÉNÉRALE ET EXEMPLE

J. BOULARAN (1 2), H. CAUSSINUS (2)

(1) *Gaz de France, Direction de la recherche,
Groupe Mathématiques et Statistiques,*

361 Av. du P. Wilson, 93211 La Plaine Saint Denis cedex, France.

(2) *Laboratoire de Statistique et Probabilités,
U.R.A. CNRS 745, Université Paul Sabatier,*

118 route de Narbonne, 31062 Toulouse cedex.

RÉSUMÉ

Cet article donne une présentation synthétique du modèle de régression sur variables entachées d'erreurs à partir du modèle à effets fixes de l'Analyse en Composantes Principales. Les problèmes d'estimation par moindres carrés sont discutés dans un cadre général, incluant la prise en compte de plusieurs variables dépendantes, l'existence de certaines liaisons linéaires entre les paramètres, le cas « mixte » où certaines variables sont mesurées sans erreurs. Un exemple industriel motive et illustre ces diverses inflexions du modèle initial. Enfin, les questions de stabilité des estimateurs sont évoquées et deux méthodes de protection sont suggérées.

Mots-clés : *Analyse en composantes principales, choix de métrique, contraintes linéaires, effets fixes, modèle fonctionnel linéaire, moindres carrés généralisés, régression avec erreurs sur les variables, régression linéaire, ridge régression.*

SUMMARY

A synthetic presentation of the error in variables regression is provided within the framework of the fixed effect model for Principal Component Analysis. Generalized least squares estimates are discussed for the case of several dependent variables and possibly modified models where linear constraints are taken into account and/or some of the independent variables are measured without errors. An industrial example motivates and illustrates the basic model as well as its modified versions. Finally, the problems arising from the instability of the least squares estimates are considered and two approaches are suggested to prevent them.

Key-words : *Errors in variables regression, fixed effects, generalized least squares, linear constraints, linear functional model, linear regression, metric choice, principal component analysis, ridge regression.*

1. Introduction

Dans le modèle de régression linéaire classique, l'espérance mathématique de la variable Y à expliquer est une fonction linéaire d'une variable explicative X , en général vectorielle, supposée connue même si X peut être l'observation d'une variable aléatoire (c'est alors l'espérance conditionnelle de Y à X qui est fonction linéaire de X). Cependant, dans bon nombre de problèmes pratiques, il est raisonnable de considérer que la relation linéaire est entre la moyenne inconnue de Y , soit y , et une inconnue x , espérance mathématique de l'observable X . Le modèle linéaire «fonctionnel» est alors entre les inconnues x et y , «mesurées» respectivement par les éléments aléatoires X et Y . On dit souvent qu'il s'agit de régression sur variables entachées d'erreurs. Contrairement à la situation de la régression classique, X et Y ont alors des statuts tout à fait comparables au plan mathématique.

Précisons le modèle dans le cas de n observations indicées par i ($i = 1, \dots, n$). Si les X_i sont des vecteurs aléatoires de $\mathbb{R}^q$ et les Y_i des vecteurs aléatoires de $\mathbb{R}^p$, notons $Z_i = \begin{bmatrix} Y_i \\ X_i \end{bmatrix}$ le vecteur correspondant de $\mathbb{R}^{p+q}$ et $z_i = E(Z_i) = \begin{bmatrix} y_i \\ x_i \end{bmatrix}$. Le modèle fonctionnel linéaire stipule l'existence d'une matrice F (inconnue), de dimension $p \times q$, telle que

$$y_i = Fx_i \quad \text{pour tout } i = 1, \dots, n \quad (1.1)$$

c'est-à-dire l'existence d'une matrice T (non unique), de dimension $(p+q) \times p$, telle que

$$T'z_i = 0 \quad \text{pour tout } i = 1, \dots, n \quad (1.2)$$

Cette dernière relation (1.2) écrit que les z_i appartiennent à un certain sous-espace de dimension q de $\mathbb{R}^{p+q}$ dont la recherche (en fait l'estimation) relève de l'Analyse en Composantes Principales (A.C.P.) : voir par exemple Lebart *et al.* (1979, IV.6.2). Les similitudes de la régression sur variables entachées d'erreurs et de l'A.C.P. sont particulièrement claires lorsque cette dernière est introduite à partir du modèle à effets fixes (voir Caussinus (1984) ou Dreesbeke *et al.* (1992, chapitre 3)). Un premier objectif du présent article est de préciser ces liens en présentant l'estimation par moindres carrés des paramètres T ou F (et incidemment des z_i) de la façon qui nous paraît la plus simple tout en étant très générale; nous nous plaçons en effet dans le cas d'un nombre arbitraire p de variables à expliquer, alors que la littérature considère essentiellement le cas $p = 1$, nous considérons une matrice de variances et covariances des erreurs quelconque et discutons en conséquence le choix de la métrique convenable pour l'A.C.P.. Notons que, si l'estimation du modèle implicite (1.2) relève strictement d'opérations usuelles en A.C.P., l'estimation du modèle explicite (1.1) est propre au problème de régression (et, sans présenter de difficulté théorique, demande toutefois quelques précautions pour le cas $p > 1$). Notre présentation rend aussi d'un abord naturel les études d'efficacité asymptotique, en particulier les comparaisons entre techniques d'estimation (choix de métrique, comparaison des algorithmes liés à l'A.C.P. et de la régression ordinaire); de telles études théoriques ne seront pas abordées ici, mais on pourra trouver certains

développements dans Boularan (1993), comme suite aux travaux de Gleser (1981), Anderson (1984), Fuller (1987), entre autres. Ici, nous nous contenterons de donner quelques comparaisons empiriques à partir de simulations (paragraphe 6).

Le type de modélisation ci-dessus est approprié dans un grand nombre de problèmes physiques ainsi qu'il a été souligné par plusieurs auteurs, comme Ragot et Aubrun (1982). Pour notre part, un problème industriel servira d'exemple et de fil conducteur. Il s'agit d'une réaction chimique (pyrolyse du méthane) dans laquelle un mélange de plusieurs composants gazeux en entrée d'un four produit un autre mélange en sortie. Les débits de sortie ($p=9$ composants) sont autant de variables à expliquer par les débits des cinq composants d'entrée, par la température du four et le temps de séjour (soit au total $q=7$). Ces diverses quantités ne sont connues qu'à des erreurs de mesure près; on doit donc envisager un modèle de régression sur variables entachées d'erreurs. Le modèle ci-dessus s'avère néanmoins un peu simpliste pour au moins deux raisons :

- Il ne tient pas compte de relations linéaires connues entre les composants en entrée et en sortie du four, relations exprimant la conservation de la matière; nous voyons comment le modifier pour les prendre en compte (paragraphe 4). En fait, il s'agit là de l'un des trois types de dépendances entre les paramètres étudiés par Box *et al.* (1973). Mentionnons que des contraintes différentes mais de nature voisine ont été envisagées par Ragot et Aubrun (1982).

- Contrairement aux autres, la variable température est mesurée pratiquement sans erreurs, si bien que l'on est en présence d'une situation mixte dans laquelle certaines variables sont connues avec certitude (comme en régression ordinaire) et d'autres ne sont connues qu'à une erreur de mesure non négligeable près : nous voyons comment traiter ce cas au paragraphe 5.

Le champ d'application des questions étudiées ici ne se limite pas aux problèmes de nature «physique» qui viennent d'être mentionnés. On peut noter par exemple, comme Anderson (1984), que l'approche des modèles à équations simultanées utilisés en économétrie (voir Malinvaud (1981)) conduit au modèle que nous envisageons.

Cazes (1975) a montré que la régression sur variables entachées d'erreurs (avec $p = 1$) pouvait être envisagée dans certains cas pour «protéger» une régression (dans le même esprit que la ridge régression). Toutefois, cela ne doit pas faire oublier que les estimations ainsi obtenues peuvent être aussi très instables et que la régression sur variables entachées d'erreurs devrait être parfois elle-même «protégée». Nous proposons pour cela quelques techniques heuristiques (paragraphe 3), en poursuivant les travaux de Legendre (1977).

2. Le modèle

Les vecteurs étant assimilés à des matrices colonnes et X' désignant la transposée de la matrice X , on considère n vecteurs aléatoires indépendants de même loi $Z_i = \begin{bmatrix} Y_i \\ X_i \end{bmatrix}$ ($i = 1, \dots, n$) avec Y_i à valeurs dans $\mathbb{R}^p$, X_i à valeurs dans $\mathbb{R}^m$, Z_i à valeurs dans $\mathbb{R}^m$ ($m = p + q$). On suppose que les moments d'ordre deux existent et

l'on pose :

$$\begin{aligned} E(Y_i) &= y_i \in \mathbb{R}^p, E(X_i) = x_i \in \mathbb{R}^q \\ \text{Var}(Y_i) &= \sigma^2 \Gamma_Y, \text{Var}(X_i) = \sigma^2 \Gamma_X, \text{Cov}(Y_i, X_i) = \sigma^2 \Gamma_{YX} \end{aligned}$$

où les y_i et x_i sont des paramètres inconnus, Γ_Y et Γ_X sont des matrices symétriques définies positives connues, respectivement d'ordre p et q , Γ_{YX} est une matrice $p \times q$ connue, et σ est un réel strictement positif éventuellement inconnu. Dans le modèle fonctionnel linéaire on suppose de plus qu'il existe F , matrice $p \times q$ de paramètres inconnus, telle que :

$$y_i = Fx_i \quad \text{pour tout } i = 1, \dots, n \quad (2.1)$$

Le premier problème qui se pose est l'estimation des paramètres $y_i, x_i (i = 1, \dots, n)$ et, surtout, F . Ce modèle peut s'écrire comme un modèle à effets fixes en Analyse en Composantes Principales. En effet, l'équation (2.1) peut se mettre sous la forme :

$$T' z_i = 0 \quad (2.2)$$

avec $z_i = (y'_i, x'_i)'$ et, par exemple, $T' = [I_p] - F$ où I_p désigne la matrice identité d'ordre p . On peut noter que d'autres choix de T sont possibles, si bien que ce paramètre matriciel ($m \times p$) n'est pas identifiable (contrairement à F). Le point important est que les colonnes de T engendrent un sous-espace vectoriel de dimension p orthogonal aux z_i . Cela revient à dire qu'il existe une variété linéaire E_q de $\mathbb{R}^m$, de dimension q , telle que :

$$z_i \in E_q \quad \text{pour tout } i = 1, \dots, n. \quad (2.3)$$

Puisque $z_i = E(Z_i)$ où les Z_i sont des vecteurs aléatoires indépendants avec $\text{var}(Z_i) = \sigma^2 \Gamma$ en posant $\Gamma = \begin{bmatrix} \Gamma_Y & \Gamma_{YX} \\ \Gamma'_{YX} & \Gamma_X \end{bmatrix}$, qu'on suppose définie positive, il s'agit bien du modèle à effets fixes en Analyse en Composantes Principales, tel qu'il est présenté, par exemple, dans Caussinus (1985 a,b) ou Seber (1984).

3. Estimation des paramètres

3.1. Estimation par moindres carrés

Si nous supposons que les n vecteurs Z_i suivent indépendamment une loi gaussienne de dimension $m = p + q$, d'espérance z_i et de même matrice de variance-covariance $\sigma^2 \Gamma$, la fonction de vraisemblance est :

$$(2\pi\sigma^2)^{-mn/2} |\Gamma|^{-n/2} \exp \left(-\left(\frac{1}{2\sigma^2}\right) \times \sum_{i=1}^n \|Z_i - z_i\|_{\Gamma^{-1}}^2 \right)$$

L'estimation du maximum de vraisemblance conduit à chercher l'espace vectoriel E_q de dimension q et les vecteurs $z_i (i = 1, \dots, n)$ vérifiant (2.3) qui minimisent

$\sum_{i=1}^n \| Z_i - z_i \|_{\Gamma^{-1}}^2$. Sans faire l'hypothèse de normalité, en munissant $\mathbb{R}^m$ d'une

métrique M , l'estimateur des moindres carrés est celui qui minimise $\sum_{i=1}^n \| Z_i - z_i \|_M^2$.

Dans ce dernier cas, les estimateurs de E_q et des z_i sont donnés par la proposition suivante :

Proposition 1 :

L'estimateur des moindres carrés de E_q , pour la métrique M , est le sous-espace $\hat{E}_q$ engendré par les q vecteurs propres associés aux q plus grandes valeurs

propres de VM , avec $V = \frac{1}{n} \sum_{i=1}^n Z_i Z_i'$, et l'estimateur de z_i est le vecteur :

$$\hat{z}_i = \Pi_{\hat{E}_q}^M(Z_i) \quad (3.1)$$

où $\Pi_{\hat{E}_q}^M$ est le projecteur M -orthogonal sur $\hat{E}_q$.

En effet, puisque z_i appartient à E_q , on a :

$$\| Z_i - z_i \|_M^2 = \| Z_i - \Pi_{E_q}^M Z_i \|_M^2 + \| \Pi_{E_q}^M Z_i - z_i \|_M^2 \quad \text{pour } i = 1, \dots, n.$$

D'autre part, par définition de la norme au sens de M et par la M -symétrie d'un projecteur M -orthogonal :

$$\begin{aligned} \sum_{i=1}^n \| Z_i - \Pi_{E_q}^M Z_i \|_M^2 &= \sum_{i=1}^n Z_i' (I - \Pi_{E_q}^M)' M (I - \Pi_{E_q}^M) Z_i \\ &= \text{Tr} \left[\sum_{i=1}^n Z_i Z_i' M (I - \Pi_{E_q}^M) \right] = n \text{Tr} \left[VM (I - \Pi_{E_q}^M) \right] \end{aligned}$$

$$\text{Donc } \sum_{i=1}^n \| Z_i - z_i \|_M^2 = n \text{Tr} \left[VM (I - \Pi_{E_q}^M) \right] + \sum_{i=1}^n \| \Pi_{E_q}^M Z_i - z_i \|_M^2.$$

Le minimum est obtenu lorsque le deuxième terme est nul, ce qui donne les $\hat{z}_i$ définis par (3.1), et lorsque le premier terme est minimum, ce qui donne bien pour $\hat{E}_q$ l'espace engendré par les vecteurs propres M -orthonormés associés aux q plus grandes valeurs propres de VM . $\square$

Les p vecteurs propres associés aux p plus petites valeurs propres de VM sont M -orthogonaux à $\hat{E}_q$; en les prémultipliant par M , on obtient donc les colonnes de

$\hat{T}$, estimant une matrice T particulière (telle que $T'M^{-1}T = I_p$) parmi celles qui réalisent (2.2). En changeant de base dans l'orthogonal (canonique) de $\hat{E}_q$, c'est-à-dire par identification des termes de l'équation :

$$\hat{T}A = \begin{bmatrix} \hat{T}_1 \\ \hat{T}_2 \end{bmatrix} \quad A = \begin{bmatrix} I_p \\ -\hat{F}' \end{bmatrix}$$

où $\hat{T}_1$ est $p \times p$, $\hat{T}_2$ est $q \times p$ et A est $p \times p$, on obtient :

$$\hat{F}' = -\hat{T}_2\hat{T}_1^{-1} \quad (3.2)$$

sous l'hypothèse que $A = \hat{T}_1^{-1}$ existe, ce qui est vrai presque sûrement (Gleser, 1981).

Proposition 2

Soit $\hat{U} = \begin{bmatrix} \hat{U}_1 \\ \hat{U}_2 \end{bmatrix}$ la matrice dont les colonnes sont les q vecteurs propres associés

aux q plus grandes valeurs propres de VM , où $\hat{U}_1$ et $\hat{U}_2$ sont respectivement de dimension $p \times q$ et $q \times q$. On a :

$$\hat{F} = \hat{U}_1\hat{U}_2^{-1} \quad (3.3)$$

Il suffit d'utiliser la relation (3.1) et la relation $\hat{U}'\hat{T} = 0$. □

Dans le cas gaussien, il est clair que les estimateurs ci-dessus sont ceux du maximum de vraisemblance si l'on prend $M = \Gamma^{-1}$. De façon plus générale, un choix optimal de M proposé dans Besse *et al.* (1987, 1988) est obtenu de même pour $M = \Gamma^{-1}$ selon une propriété de type Gauss-Markov. En fait cette propriété est établie pour σ^2 petit en développant les calculs au moyen de la théorie des perturbations. Fine et Pousse (1992) montrent que ce choix est valide asymptotiquement (pour n grand) quelle que soit la valeur de σ^2 sous l'hypothèse que les erreurs aient les mêmes moments d'ordre 4. En effet l'étude de la convergence de ces estimateurs montre que ceux-ci sont consistants si et seulement si $M = \Gamma^{-1}$.

En écrivant autrement (3.1), les estimations de y_i et x_i sont fournies par :

$$\begin{bmatrix} \hat{y}_i \\ \hat{x}_i \end{bmatrix} = \Pi_{\hat{E}_q}^M \begin{bmatrix} Y_i \\ X_i \end{bmatrix} \quad (3.4)$$

Dans le cas gaussien, l'estimation du maximum de vraisemblance de σ^2 est :

$$\hat{\sigma}^2 = \frac{1}{m} \sum_{i=1}^n \|Z_i - \hat{z}_i\|_{\Gamma^{-1}}^2 = \frac{1}{m} \sum_{j=q+1}^m \hat{\lambda}_j \quad (3.5)$$

où $\hat{\lambda}_j$ est la j^{me} valeur propre de $V\Gamma^{-1}$, lorsque celles-ci sont rangées par valeurs décroissantes. Cependant cet estimateur est très biaisé. Gleser (1981) montre que l'on

a : $\lim_{n \rightarrow \infty} p^{-1} m \hat{\sigma}^2 = \sigma^2$ avec probabilité 1. Cela conduit à estimer σ^2 par : $\frac{1}{p} \sum_{j=q+1}^m \hat{\lambda}_j$

qui converge presque sûrement vers σ^2 . Cette propriété est encore vraie dans le cas non gaussien sous des hypothèses peu restrictives. On peut aussi améliorer cet estimateur

en prenant $\frac{n-q}{np} \sum_{j=q+1}^m \hat{\lambda}_j$ (voir par exemple Legendre (1977) pour le cas $p = 1$).

3.2. Problèmes d'instabilité; autres estimateurs

Les formules (3.2) ou (3.3) font intervenir l'inverse d'une matrice. Des difficultés peuvent donc apparaître si cette matrice est mal conditionnée. Cette situation n'est pas exceptionnelle en régression classique dès que les variables explicatives sont corrélées entre elles, c'est-à-dire lorsque les x_i sont dans un sous-espace de $\mathbb{R}^m$ de dimension strictement inférieure à q , ou du moins proche d'un tel sous-espace. Les techniques de ridge-régression (par exemple) ont été proposées pour faire face à de telles situations.

Pour tourner ce type de difficulté dans le cadre de la régression sur variables entachées d'erreurs, nous proposons deux types d'approches.

3.2.1. Recherche de relations linéaires supplémentaires

Une première approche des problèmes d'instabilité consiste à supposer qu'il existe entre les variables des relations linéaires autres que celles que l'on cherche à estimer dans le modèle du paragraphe 2, c'est-à-dire qu'il existe des liaisons linéaires entre les seules variables explicatives x à côté des p relations cherchées entre x et y . L'idée est d'essayer d'extraire l'ensemble de ces relations car, si les premières sont plus «fortes» que les secondes, elles risquent de prévaloir et d'apparaître ainsi, dans l'analyse du paragraphe 3.1, à la place de certaines des p relations d'intérêt. On suppose donc comme dans le paragraphe 2 qu'il existe p relations entre les z_i exprimées par (2.2), mais que, en outre, il existe r relations concernant les variables explicatives seules ($r < q$) exprimées par :

$$T^{*'} z_i = 0$$

où T^* est une matrice $m \times r$ dont les éléments des p premières lignes sont nuls.

Au total on suppose donc l'existence d'une matrice W de dimension $m \times (p+r)$ et de rang $(p+r)$ telle que :

$$W' z_i = 0 \quad \text{pour tout } i = 1, \dots, n \quad (3.6)$$

où la matrice $W = [T|T^*]$ peut se décomposer sous la forme :

$$W = \left[\begin{array}{c|c} I_p & 0 \\ \hline -F' & C \end{array} \right] \quad (3.7)$$

Cependant, à cause de la liaison entre les x_i , cette écriture de W n'est pas ici unique; mais on peut écrire de façon unique :

$$W = \left[\begin{array}{c} I_{p+r} \\ \hline -F'_1 \end{array} \right] \quad (3.8)$$

En fait, faire un choix parmi les matrices W satisfaisant (3.6) est équivalent à choisir une base de l'orthogonal canonique du sous-espace contenant les z_i , en postmultipliant W par une matrice régulière $(p+r) \times (p+r)$.

Le modèle est analogue à celui du paragraphe 2, si ce n'est que la dimension du sous-espace contenant les z_i est maintenant $q-r$. Comme plus haut, ce sous-espace E_{q-r} est estimé par le sous-espace $\widehat{E}_{q-r}$ engendré par les $q-r$ vecteurs propres de VM associés aux $q-r$ plus grandes valeurs propres. Pour estimer la matrice W , considérons les $p+r$ vecteurs propres de VM associés aux $p+r$ plus petites valeurs propres. Ils engendrent le sous-espace M -orthogonal à $\widehat{E}_{q-r}$; en les prémultipliant par M on obtient les colonnes d'une matrice $\widehat{W}$.

Ecrivons $\widehat{W} = \left[\begin{array}{c} \widehat{W}_1 \\ \hline \widehat{W}_2 \end{array} \right]$ où $\widehat{W}_1$ et $\widehat{W}_2$ sont respectivement de dimensions $(p+r) \times (p+r)$ et $(q-r) \times (p+r)$. On estime F_1 par un raisonnement semblable à celui qui conduit à (3.2). Si $\widehat{W}_1$ est régulière, ce qui est presque sûrement vrai, on pose :

$$\widehat{F}'_1 = -\widehat{W}_2 \widehat{W}_1^{-1}$$

En notant U^* la matrice $(q-r) \times m$ dont les colonnes sont les $q-r$ vecteurs propres associés aux $q-r$ plus grandes valeurs propres de VM , et écrivant $U^* = \left[\begin{array}{c} U_1^* \\ \hline U_2^* \end{array} \right]$ où U_1^* et U_2^* sont respectivement de dimensions $(p+r) \times (q-r)$ et $(q-r) \times (q-r)$, on a $U^{*'} \widehat{W} = 0$, d'où :

$$\widehat{F}_1 = U_1^* U_2^{*-1}$$

Commentaires :

L'existence de liaisons entre les x_i implique que la matrice F n'est pas définie de façon unique. La méthode ci-dessus correspond à estimer l'une de ces matrices. En comparant (3.7) et (3.8), on voit qu'il suffit pour cela de prendre les p premières

lignes de $\widehat{F}_1$ (ce qui donne la matrice $p \times (q - r) : \widehat{F}_1^*$) et de la faire « précéder » d'une matrice $p \times r$ de zéros, c'est-à-dire d'estimer F par $[0_{p,r} | \widehat{F}_1^*]$. On notera que cela revient à « sortir » les r premières coordonnées de x_i de la relation entre y_i et x_i .

En pratique, on peut penser que l'inversion de U_2^* est plus « stable » que celle de $\widehat{U}_2$ puisque U_2^* est une sous-matrice de $\widehat{U}_2$. Mais, d'un autre côté, supposer à tort des liaisons entre les x_i comporte des inconvénients évidents. La méthode reste donc heuristique, en particulier sur le choix de r (nous pensons qu'une petite valeur, un ou deux, doit être raisonnable dans la plupart des cas) et nécessitera une étude théorique et pratique plus approfondie avant toute application systématique.

Si Γ est connue, on aura intérêt à prendre $M = \Gamma^{-1}$ comme plus haut.

3.2.2. Une approche bayésienne.

La deuxième approche que nous préconisons est d'inspiration bayésienne. Le modèle est toujours celui du paragraphe 2, que l'on précise en supposant la loi de Z_i normale, conditionnellement aux paramètres $z_i (i = 1, \dots, n)$ et E_q . Par ailleurs, ces derniers sont maintenant considérés comme des variables aléatoires dont la loi *a priori* est définie comme suit :

- (i) Conditionnellement à E_q , les z_i suivent chacun, indépendamment les uns des autres, une loi (ou plutôt une pseudo-loi) uniforme sur E_q ,
- (ii) La loi *a priori* de E_q est donnée de la façon suivante : le sous-espace Γ^{-1} — orthogonal à E_q — est engendré par un système de vecteurs aléatoires $(v_1, \dots, v_p)$ de $\mathbb{R}^m$, indépendants tels que v_j est de loi gaussienne d'espérance μ_j et de matrice de variance-covariance $\sigma^2 \Sigma$ indépendante de j .

Cela généralise à $p \geq 1$ le modèle proposé par Caussinus (1984, 4.4) pour le cas $p = 1$. Noter que l'espace $E_q^\perp$ ainsi engendré est presque sûrement de dimension p , donc E_q est de dimension $q = m - p$, dès lors que $p \leq \text{rang}(\Sigma)$ ce que l'on supposera. On supposera en fait Σ régulière et, de plus, $\Sigma = \frac{1}{a} \Gamma (a > 0)$ pour simplifier les calculs.

Dans le cadre ci-dessus, l'estimation des paramètres sera fournie par la valeur modale de leur densité *a posteriori* sachant les $Z_i (i = 1, \dots, n)$. Celle-ci est donnée par la proposition suivante dont on trouvera la démonstration en Appendice.

Proposition 3 :

Sous les hypothèses énoncées ci-dessus et si $\Sigma = \frac{1}{a} \Gamma$:

$\widehat{E}_q$ est engendré par les vecteurs propres normés associés aux q plus grandes valeurs propres de $(V - \frac{a}{n} W) \Gamma^{-1}$ où $V = (1/n) \sum Z_i Z_i'$ et $W = \sum \mu_j \mu_j'$,
 $\widehat{z}_i = \Pi_{\widehat{E}_q}^{\Gamma^{-1}}(Z_i)$ pour tout $i = 1, \dots, n$.

Commentaire :

La méthode bayésienne présentée conduit donc à remplacer la diagonalisation de $V\Gamma^{-1}$ par celle de $(V - \frac{a}{n}W)\Gamma^{-1}$, ce qui la fait très proche parente de la Ridge-Régression. En fait notre méthode généralise celle de Legendre (1977), proposée pour $p = 1$ à partir d'arguments très voisins de ceux de Hoerl and Kennard (1970) pour la Ridge-Régression classique (bornes pour la norme des paramètres de régression). Comme en Ridge-Régression, le choix optimal de a peut se faire en étudiant la variation des estimateurs en fonction de ce paramètre et en adoptant une valeur de a au voisinage de laquelle cette variation est faible. Quand au choix de W , il peut résulter d'un choix *a priori* des μ_j si quelques éléments pour cela sont disponibles. Sinon, on

peut envisager de prendre $W = \left[\begin{array}{c|c} I_p & 0 \\ \hline 0 & 0 \end{array} \right]$ généralisant ainsi les propositions de Legendre (1977). Cependant, la question reste très largement à approfondir.

4. Présence de contraintes linéaires

Dans l'exemple de la pyrolyse du méthane, les vraies valeurs des débits doivent vérifier les contraintes de conservation de la matière. En particulier, le nombre de moles de carbone (resp. hydrogène) en «entrée» doit être égal au nombre de moles de carbone (resp. hydrogène) en «sortie», ce qui conduit à deux contraintes linéaires sur les vecteurs x_i et y_i . De façon générale, ce type de contraintes amène à modifier le problème initial en supposant l'existence de deux matrices $A(d \times p)$ et $B(d \times q)$ connues telles que :

$$Ay_i + Bx_i = 0 \quad \text{pour tout } i = 1, \dots, n \quad (4.1)$$

La contrainte (4.1) s'écrit encore comme l'existence d'une matrice $C(d \times m)$ connue telle que :

$$Cz_i = 0 \quad \text{pour tout } i = 1, \dots, n \quad (4.2)$$

en notant $C = [A|B]$. Sans réduire la généralité, on supposera les lignes de C linéairement indépendantes. Notons H le sous-espace de $\mathbb{R}^m$ canoniquement orthogonal aux lignes de C . On a $\dim(H) = m - d$ et (4.2) signifie que z_i appartient à H pour tout $i = 1, \dots, n$. Dans le cadre du modèle à effets fixes du paragraphe 2, cela revient à ajouter la condition :

$$E_q \subset H \quad (4.3)$$

où le sous-espace H est donné, de dimension $m - d$. Bien entendu, puisque E_q est de dimension q , il faut $q < m - d$ soit $d < p$. Quand les paramètres sont exprimés sous cette contrainte complémentaire la Proposition 1 devient la Proposition 4 suivante dont la démonstration est donnée en Appendice :

Proposition 4 :

Sous les hypothèses du modèle du paragraphe 2, et l'hypothèse additionnelle

(4.3), l'estimateur des moindres carrés de E_q pour la métrique M est $\hat{E}_q$, sous-espace engendré par les q vecteurs propres associés aux q plus grandes valeurs propres de $\tilde{V}M$, où

$$\tilde{V} = \frac{1}{n} \sum_{i=1}^h \Pi_H^M Z_i (\Pi_H^M Z_i)'$$

L'estimateur de z_i est $\hat{z}_i = \Pi_{E_q}^M(Z_i)$.

On notera que cette estimation vérifie bien $\hat{E}_q \subset H$ car les vecteurs propres de $\tilde{V}M$ associés à des valeurs propres non nulles appartiennent à H . L'estimation de F sera obtenue à partir de l'estimation de E_q en utilisant (3.2) ou (3.3), où $\hat{T}$ et $\hat{U}$ sont simplement calculés à partir de $\tilde{V}$ au lieu de V .

Les résultats de Besse *et al.* (1987, 1988) peuvent s'étendre au cas précédent, si bien que la métrique $M = \Gamma^{-1}$ doit être ici encore préconisée.

5. Matrice des covariances non régulière

5.1. Modèle sans contrainte

Le modèle est toujours le même qu'au paragraphe 2, sauf que la matrice des variances-covariances $\sigma^2 \Gamma$ d'un Z_i n'est plus supposée définie positive. Nous nous limitons à la situation où la matrice Γ est de la forme :

$$\Gamma = \begin{bmatrix} \Gamma_1 & 0 \\ 0 & 0 \end{bmatrix} \quad \text{avec } \Gamma_1 \text{ matrice } m_1 \times m_1 \text{ de rang } m_1 \quad (5.1)$$

qui correspond au cas où certaines variables sont considérées sans erreurs comme en régression classique (dans l'exemple de la pyrolyse du méthane, il en est ainsi de la température qui est connue pour chaque expérience avec une précision extrême par rapport aux autres variables). On peut noter cependant que, grâce à des transformations linéaires convenables, la situation ainsi traitée est, de fait, beaucoup plus générale. On supposera que toutes les variables «à expliquer» (y_i) sont mesurées avec erreur, de sorte que l'on a $m_1 \geq p$. On prendra $m_1 > p$ pour éliminer le cas trivial de la régression ordinaire. On notera

$$Z = (Z_1, Z_2, \dots, Z_n) = \begin{bmatrix} R \\ S \end{bmatrix}$$

où R est une matrice $m_1 \times n$ telle que la ligne j de R est la j^{me} variable avec erreur mesurée successivement sur les n individus, et S est une matrice $(m - m_1) \times n$, les

lignes de S correspondant aux variables sans erreur. On a donc :

$$\left. \begin{array}{l} Z_i^j = R_i^j \\ Z_i^{m_1+h} = S_i^h \end{array} \quad \begin{array}{l} j = 1, \dots, m_1 \\ h = 1, \dots, (m - m_1) \end{array} \right\} \quad \text{pour } i = 1, \dots, n$$

Les colonnes de R sont n vecteurs aléatoires R_i indépendants tels que $E(R_i) = r_i$, avec $r_i^j = z_i^j$ pour tout $i = 1, \dots, n$ et $j = 1, \dots, m_1$ et $\text{Var}(R_i) = \sigma^2 \Gamma_1$. On posera $r = E(R)$ et on utilisera s ou S pour désigner la matrice connue S , puisque $s = E(S) = S$.

L'équation (2.2) s'écrit ici en posant $T = \begin{bmatrix} T_R \\ T_S \end{bmatrix}$ avec une décomposition en blocs évidente :

$$T_R' r_i + T_S' s_i = 0 \quad \text{pour tout } i = 1, \dots, n \quad (5.2)$$

Une estimation des moindres carrés demande ici le choix d'une métrique M_1 sur $\mathbb{R}^{m_1}$. On peut encore considérer M_1 comme le bloc diagonal supérieur $m_1 \times m_1$ de la métrique M définie sur $\mathbb{R}^m$. Les estimateurs correspondants sont donnés par la proposition suivante dont on trouvera la démonstration dans l'Appendice.

On note $P_s = s'(ss')^{-1}s$ le projecteur orthogonal canonique de $\mathbb{R}^n$ sur le sous-espace engendré par les lignes de la matrice S , c'est-à-dire les vecteurs s^k ($k = 1, \dots, m - m_1$), et on note $\hat{\beta} = Rs'(ss')^{-1}$ la matrice $m_1 \times (m - m_1)$ des estimateurs des moindres carrés des paramètres de la régression des lignes de R sur celles de s , soit $R = \beta s + \varepsilon$. Enfin $q_1 = m_1 - p$ désigne le nombre de variables explicatives mesurées avec erreurs.

Proposition 5 :

Sous le modèle du paragraphe 2, modifié par l'hypothèse (5.1), les estimateurs des moindres carrés sont :

- (i) $\hat{z} = \begin{bmatrix} \hat{r} \\ \hat{s} \end{bmatrix}$ où $\hat{r} = \Pi_{\hat{E}_{q_1}}^{M_1} R + \Pi_{\hat{E}_{q_1}^\perp}^{M_1} R P_s$ et $\hat{E}_{q_1}$ est le sous-espace engendré par les vecteurs propres associés aux q_1 plus grandes valeurs propres de $V_1 M_1$, avec $V_1 = (1/n) R(I - P_s) R'$.
- (ii) L'estimation d'une matrice T est $\hat{T} = \begin{bmatrix} \hat{T}_R \\ \hat{T}_s \end{bmatrix}$ où les colonnes de $\hat{T}_R$ sont, prémultipliés par $\hat{M}$, les vecteurs propres associés aux p plus petites valeurs propres de $V_1 M_1$ et $\hat{T}_s = -\hat{\beta}' \hat{T}_R$.
- (iii) En posant ici encore $\hat{T} = \begin{bmatrix} \hat{T}_1 \\ \hat{T}_2 \end{bmatrix}$ où $\hat{T}_1$ est $p \times p$, l'estimation de F est donnée par $\hat{F}' = -\hat{T}_2 \hat{T}_1^{-1}$.

Remarque :

Un exemple d'application banal est celui du modèle affine dans lequel la condition $T'z_i = 0$ est remplacée par $T'z_i = b$ pour tout $i = 1, \dots, n$, où b est un vecteur inconnu de $\mathbb{R}^p$. Il suffit en effet de remplacer T' par $[T' | -b]$ et d'ajouter une coordonnée connue égale à 1 à tous les z_i pour se ramener à un cas particulier de la situation ci-dessus où m est remplacé par $m+1$ et m_1 par m . On vérifie facilement que le projecteur $I - P_s$ consiste à centrer : $R(I - P_s)$ est ici $[Z_1 - \bar{Z}] \dots [Z_n - \bar{Z}]$ avec $\bar{Z} = \frac{1}{n} \sum_{i=1}^n Z_i = \hat{\beta}$; la matrice V_1 est la matrice des variances et covariances empiriques des Z_i (alors que V est la matrice des moments d'ordre deux).

5.2. Modèle avec contraintes linéaires

Nous supposons ici que le modèle doit prendre en compte à la fois le problème des contraintes sur les espérances et celui de certaines variables explicatives sans erreur. C'est le cas dans l'exemple de la pyrolyse du méthane : les vraies valeurs des débits doivent satisfaire les contraintes de stoechiométrie (bilans matières) et la température est supposée mesurée sans erreur. Pour cela, nous devons supposer dans notre modèle que le rang m_1 de la matrice des variances et covariances est inférieur à m comme dans le cas précédent et imposer que le sous-espace dans lequel sont les vraies valeurs z_i est contenu dans un sous-espace de contrainte H , comme dans le paragraphe 4.

Le modèle est celui du paragraphe 2 plus les hypothèses additionnelles (4.2) ou (4.3) et (5.1). On suppose que les contraintes n'affectent pas les variables connues sans erreurs, c'est-à-dire que C peut s'écrire $C = [C_R | 0]$ où C_R est de dimension $d \times m_1$. Avec encore $T' = [T'_R | T'_S]$ nous avons donc pour tout $i = 1, \dots, n$:

$$\begin{aligned} C_R r_i &= 0 \\ T'_R r_i + T'_S s_i &= 0 \end{aligned}$$

Les estimateurs des moindres carrés sont donnés par la proposition suivante.

Proposition 6 :

Sous le modèle 2 modifié par les hypothèses exprimées ci-dessus, $\hat{T}' = [\hat{T}'_R | \hat{T}'_S]$ où les colonnes de $\hat{T}_R$ sont les vecteurs propres, prémultipliés par M_1 , associés aux p plus petites valeurs propres de $\tilde{V}_1 M_1$ avec :

$$\tilde{V}_1 = \frac{1}{n} \Pi_{H_R}^{M_1} R(I - P_s)(\Pi_{H_R}^{M_1} R(I - P_s))',$$

H_R étant le sous espace de $\mathbb{R}^{m_1}$ canoniquement orthogonal aux lignes de C_R , et $\hat{T}_S = -\hat{\beta}' \hat{T}_R$ où $\hat{\beta} = \hat{r}s'(ss')^{-1}$.

$\hat{F}$ se déduit de $\hat{T}$ comme plus haut.

6. Comparaison des techniques sur données simulées

Nous donnons ici des résultats sur données simulées pour comparer les estimateurs obtenus en supposant le modèle à erreurs et ceux de la régression classique. En particulier, nous essayons de montrer comment la prise en compte d'une incertitude sur les variables explicatives, hypothèse qui paraît bien souvent mieux représenter la réalité, conduit alors à des estimations meilleures. A cet effet, plusieurs échantillons ont été générés au moyen de l'algorithme suivant :

(a.) Pour chaque $i = 1, \dots, n$, x_i est la réalisation d'une loi uniforme sur le pavé $[x_0, x_0 + ec_0]^q$ de $\mathbb{R}^q$ avec x_0 (resp. ec_0) réalisation d'une loi normale $N_q(0, 10000)$ (resp. $N_q(0, 400)$). Notons que les x_i sont considérés comme fixés malgré ce choix aléatoire décidé seulement pour des raisons de commodité et d'objectivité. Puis, pour tout i , y_i a été calculé tel que : $y_i = Fx_i$ avec F donné $p \times q$.

(b.) Les données entachées d'erreurs sont obtenues en générant des vecteurs aléatoires ε_i indépendants, chacun de loi $N_m(0, I_m)$, et en posant :

$$Z_i = z_i + \sigma \Gamma^{\frac{1}{2}} \varepsilon_i \quad \text{pour } i = 1, \dots, n \quad (6.1)$$

où Γ et $\sigma > 0$ sont donnés, et $z_i = \begin{bmatrix} y_i \\ x_i \end{bmatrix}$, $Z_i = \begin{bmatrix} Y_i \\ X_i \end{bmatrix}$.

Plusieurs situations ont été étudiées. Nous présentons l'une de ces études, qui est assez bien représentative de l'ensemble. Dans cet exemple, on a simulé 200 cas. Pour chacun, les échantillons sont de taille $n = 50$, avec $p = 2$ variables à expliquer et $q = 3$ variables explicatives. On a pris

$$F = \begin{bmatrix} 1 & 2 & 3 \\ 5 & 3 & 2 \end{bmatrix}, \quad \sigma^2 = 1, \quad \text{et} \quad \Gamma = \text{Diag}(100, 75, 75, 25, 100).$$

A titre indicatif, le Tableau 1 donne la moyenne et l'écart-type des 5 coordonnées pour les 50 individus générés pour fixer les vraies valeurs z_i^j ($i = 1, \dots, 50$; $j = 1, \dots, 5$).

TABLEAU 1

(moyenne, écart-type) des z_i^j simulés $j = 1, \dots, 5$.

j	1	2	3	4	5
z^j	(48, 152)	(-242, 505)	(-89, 99)	(67, 16)	(1, 40)

Pour mesurer la qualité de l'estimation des paramètres incidents z_i , nous avons examiné la somme des écarts quadratiques entre les vraies valeurs z_i et les estimations $\hat{z}_i$ pour les différentes techniques d'estimation. Le tableau 2 représente les moyennes

et les écarts-types de $\sum_{i=1}^n (\hat{z}_i^j - z_i^j)^2$ pour chaque j de 1 à $m = 5$ sur les 200 échantillons simulés comme indiqué ci-dessus. Les écarts entre les vraies valeurs à expliquer y_i et leur estimation $\hat{y}_i$ (deux premières coordonnées de z_i et $\hat{z}_i$) sont nettement plus faibles en moyenne quadratique quand les erreurs sur les variables explicatives x sont prises en compte dans la méthode d'estimation.

TABLEAU 2
(Moyenne ($\times 10^{-2}$), écart-type ($\times 10^{-2}$)) de $\sum_{i=1}^n (\hat{z}_i^j - z_i^j)^2$ pour $j = 1, \dots, 5$

j	1	2	3	4	5
régression classique	(489, 98)	(1161, 226)	(37, 7)	(12, 2)	(50, 10)
Modèle à erreurs M=I	(42, 8)	(36, 7)	(7, 2)	(15, 3)	(13, 3)
Modèle à erreurs M= Γ^{-1}	(42, 8)	(36, 7)	(6, 2)	(11, 2)	(12, 3)

Les erreurs d'estimation des vraies valeurs x^j pour $j = 3, 4, 5$, sous les modèles à erreurs, sont de même plus faibles que celles commises en supposant les variables explicatives X^j ($j = 3, 4, 5$) sans erreur dans le modèle de régression classique. On note enfin que le «bon choix» de la métrique dans le modèle à erreurs ($M = \Gamma^{-1}$ au lieu de $M = I$) n'est pas d'une importance primordiale dans cet exemple, même si les erreurs obtenues sont, dans l'ensemble, sensiblement moindres.

La figure 1 représente les «boîtes à moustaches» des 200 valeurs de $\sum_{i=1}^n (\hat{z}_i^j - z_i^j)^2$ pour $j = 1, \dots, 5$ et les trois méthodes d'estimation. Les extrémités des moustaches représentent les centiles 5 et 95 des 200 échantillons.

L'examen de cette figure confirme et précise le gain obtenu par la prise en compte des erreurs dans les modèles, que M soit la métrique identité ou que M soit égal à Γ^{-1} .

Concernant l'estimation des éléments de F , la figure 2 représente les boîtes à moustaches des $\hat{F}^{jk}$ calculés pour les 200 échantillons simulés, dans le cas des modèles à erreur ($M = \text{Identité}$ et $M = \Gamma^{-1}$) et pour la régression classique. Les lignes pointillées rappellent les vraies valeurs de F^{jk} .

L'estimation de F par régression usuelle, est toujours très biaisée. On peut remarquer que si l'on ne choisit pas $M = \Gamma^{-1}$, le modèle à erreur donne cependant des estimations moins biaisées que le modèle sans erreurs, mais que la meilleure estimation est obtenue quand la métrique choisie est $M = \Gamma^{-1}$.

A : modèle à erreur avec $M = \text{Identité}$.

B : modèle à erreur avec $M = \Gamma^{-1}$.

C : régression classique.

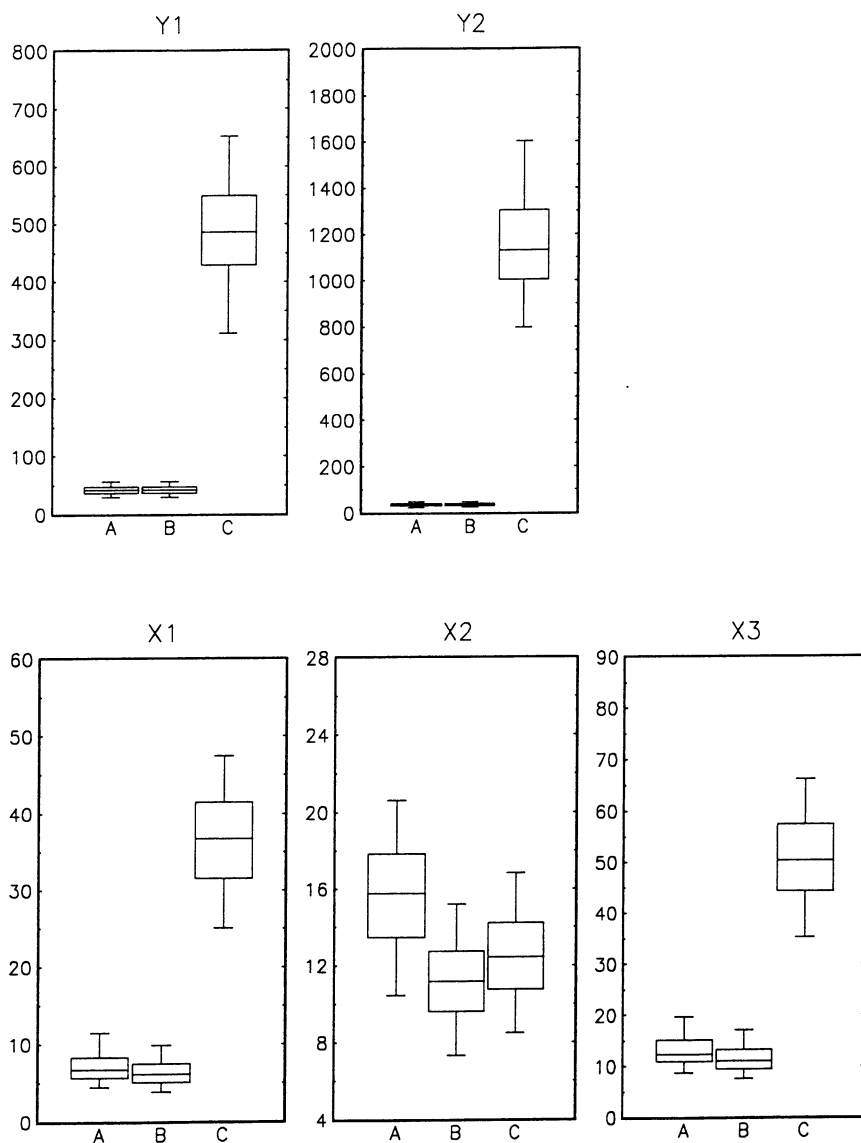


Figure 1 : boîtes à moustaches de $\sum_{i=1}^n (\hat{z}_i^j - z_i^j)^2$ pour $j = 1, \dots, 5$.

A : modèle à erreur avec $M = \text{Identité}$.

B : modèle à erreur avec $M = \Gamma^{-1}$.

C : régression classique.

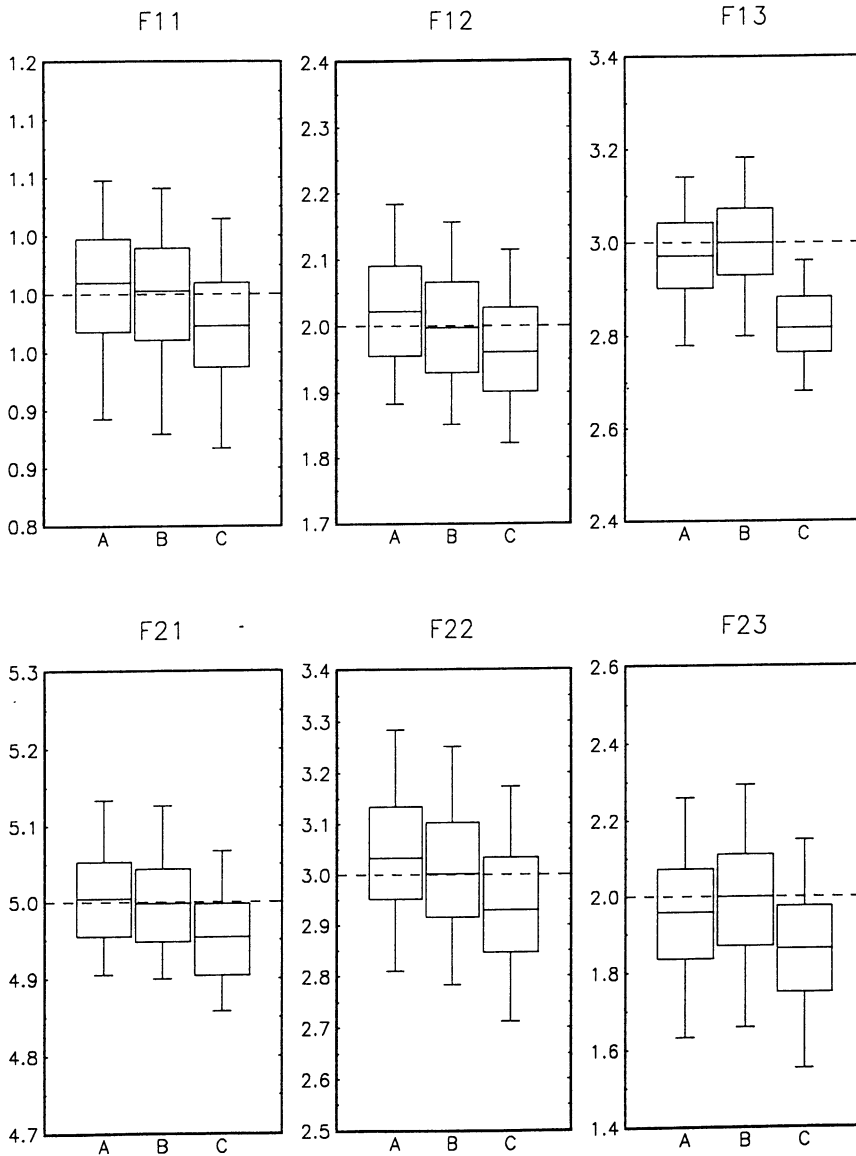


Figure 2 : boîtes à moustaches de $\hat{F}^{jk}$ pour $j=1, 2$ et $k=1, 2, 3$.

7. Exemple réel

Cet exemple concerne l'étude de la pyrolyse du méthane. Il s'agit d'un projet de recherche concernant les réactions chimiques à hautes températures, peu connues pour l'instant. Ce projet doit déboucher sur le développement d'un procédé industriel de production de gaz industriels à hautes valeurs ajoutées, à partir du gaz naturel. Nous nous occupons ici de modéliser la transformation chimique du gaz naturel et de l'hydrogène, obtenue par chauffage dans un four à haute température, et d'en estimer les paramètres. Cette modélisation sera introduite ensuite dans un simulateur, en vue d'optimiser le coût du processus de production de l'acétylène et de l'éthylène notamment; il est donc intéressant de l'estimer au mieux. Les données à notre disposition sont, pour chaque expérience, les mesures des débits en moles/heure des constituants en entrée du four, la température et le temps de séjour, ainsi que les mesures des débits des constituants du mélange produit en sortie. Ces mesures sont obtenues par chromatographie en phase gazeuse et sont entachées d'erreurs qui peuvent être assez importantes. Nous pouvons considérer que les mesures sont indépendantes d'une expérience à l'autre (nettoyage du four, étalonnage du chromatographe avant chaque expérience). Un jeu de $n = 164$ expériences nous a été fourni. Ici, $m = 16$, $q = 7$ et $p = 9$. Les sept variables «entrée» (X) sont les suivantes :

H_2 , CH_4 , C_3H_8 , C_4H_{10} , C_2H_6 , Température, Temps de séjour

tandis que les neuf variables en sortie du four (Y) sont :

H_2 , CH_4 , C_2H_4 , C_2H_2 , C_2H_6 , C_6H_6 , C_3C_5 , C_7^+ , Coke.

Dans la suite, pour éviter les confusions, certains produits se retrouvant aussi bien en entrée qu'en sortie, le nom des variables de sortie est précédé par un s (par exemple : sH_2 pour le débit d'hydrogène en sortie). La théorie chimique nous dit qu'il faut respecter les bilans matières. Ici, il s'agit des bilans carbone et hydrogène. Le nombre de moles de carbone (resp. hydrogène) en sortie doit être égal au nombre de moles de carbone (resp. hydrogène) en entrée.

En notant : $A = \begin{pmatrix} A_1 \\ A_2 \end{pmatrix}$ et $B = \begin{pmatrix} B_1 \\ B_2 \end{pmatrix}$

avec $A_1 = [0, 0, 0, 1, 2, 3, 4]$, $A_2 = [0, 0, 1, 2, 3, 4, 5]$,

$B_1 = -[0, 1, 2, 2, 2, 6, 2.9, 3.9, 1]$, $B_2 = -[1, 2, 2, 1, 3, 3, 0, 0, 0]$

et $X'_i = [\text{Température Temps de séjour } H_2 \text{ } CH_4 \text{ } C_2H_6 \text{ } C_3H_8 \text{ } C_4H_{10}]$

$Y'_i = [sH_2 \text{ } sCH_4 \text{ } sC_2H_4 \text{ } sC_2H_2 \text{ } sC_2H_6 \text{ } sC_6H_6 \text{ } sC_3C_5 \text{ } sC_7^+ \text{ } sCoke]$

nous avons sur les vraies valeurs des débits (inconnus) la contrainte (4.1) :

$Ax_i + By_i = 0$, qui écrit les relations de conservation :

$sCH_4 + 2(sC_2H_4 + sC_2H_2 + sC_2H_6) + 6sC_6H_6 + 2,9sC_3C_5 + 3,9sC_7^+ + sCoke$
 $= CH_4 + 2C_2H_6 + 3C_3H_8 + 4C_4H_{10}$

et $sH_2 + 2(sCH_4 + sC_2H_4) + sC_2H_2 + 3(sC_2H_6 + sC_6H_6)$
 $= H_2 + 2CH_4 + 3C_2H_6 + 4C_3H_8 + 5C_4H_{10}$

Pour des raisons de simplicité et du fait qu'on ne dispose pas *a priori* d'information sur la forme du phénomène, un modèle linéaire a été choisi entre les variables «d'entrée» et de «sortie».

Soit $\sigma^2\Gamma$ la matrice des variances et covariances des erreurs. Quand les paramètres sont estimés par moindres carrés, il est préférable de choisir la métrique $M = \Gamma^{-1}$ en accord avec la discussion du paragraphe 3.1 et les simulations du paragraphe 6. Cependant Γ est inconnue, ce qui a conduit à envisager (et comparer) plusieurs choix de M .

Le premier choix de M a été simplement $M = I$, mais on sait que celui-ci n'est optimal que si les erreurs sur les diverses variables sont non corrélées et de même variance. Cette métrique ne sera donc bien adaptée que si les mesures de chaque constituant sont du même ordre de grandeur et faites indépendamment sur le même instrument, ce qui n'est pas le cas ici. Dans un deuxième temps, nous avons estimé l'erreur de mesure commise sur chaque variable j , à partir de 10 essais effectués dans les mêmes conditions expérimentales, indépendamment de l'expérience proprement dite décrite ici. Les estimations des variances des erreurs de mesure sont très différentes d'une variable à l'autre, comme on peut le voir dans le tableau 3. En particulier, la température est mesurée sans erreur, ou du moins celle-ci est-elle négligeable. Nous sommes donc dans le cas d'une matrice non régulière de la forme (5.1) de rang $m_1 = 15$. Nous n'avons pas tenu compte des corrélations possibles entre les erreurs à cause du peu de données à notre disposition (10 pour chaque variable), et nous avons donc supposé que Γ était diagonale, d'éléments diagonaux estimés par les valeurs données au tableau 3. Cela nous a conduit à utiliser la Proposition 6, avec $m_1 = 15$ et la métrique diagonale des inverses des variances d'erreurs estimées pour les variables autre que la température.

TABLEAU 3

Estimation de la variance de l'erreur pour les 16 variables.

variable	sH ₂	sCH ₄	sC ₂ H ₄	sC ₂ H ₂	sC ₂ H ₆	sC ₃ C ₅	sC ₆ H ₆	sC ₇	sCOKE
variance	6.7	1.35	.05	0.04	0.002	0.00067	0.0055	0.01	0.95
variable	H ₂	CH ₄	C ₃ H ₆	C ₃ H ₈	C ₄ H ₁₀	Tempséj	TEMPE		
variance	10.6	5.11	0.002	1.03 10 ⁻⁵	9.9 10 ⁻⁶	98.7	0		

Le tableau 4 donne l'estimation des éléments de F avec $M = I$ en tenant compte des contraintes (cf. Proposition 4). Le tableau 5 donne les estimations des éléments de F en supposant Γ singulière de rang 15, avec M_1^* définie ci-dessus et en tenant compte des contraintes (cf. Proposition 6). On peut remarquer que l'estimation de σ^2 , soit 1.01, est proche de 1 dans le cas où l'on utilise M_1^* , ce qui est cohérent, alors qu'elle vaut 0.06 pour $M = I$, ce qui correspond à une «moyenne» des valeurs du tableau 3.

Le tableau 6 représente l'estimation des éléments de F obtenue par moindres carrés ordinaires sous un modèle de régression sans erreurs sur les variables explicatives.

TABLEAU 4

Estimation de F avec $M=I$.

F	H ₂	CH ₄	C ₃ H ₆	C ₃ H ₈	C ₄ H ₁₀	tempsej	TEMPE
sH ₂	0.51438	-0.45408	32.855	173.51	-187.59	0.031646	0.026152
sCH ₄	0.28799	1.2226	-16.548	-109.50	120.98	-0.017532	-0.015915
sC ₂ H ₄	-0.0076082	0.012710	0.10212	5.4913	-5.2315	0.00084170	0.00043870
sC ₂ H ₂	-0.040298	-0.016208	1.8946	22.186	-21.929	0.00095052	0.0033901
sC ₂ H ₆	0.00076940	0.00024189	0.018044	-0.42616	-0.32218	0.00016689	-1.8265E-005
sC ₃ C ₅	-0.0051709	0.0046599	0.036505	2.2544	-1.2984	-0.0011161	0.00022376
sC ₆ H ₆	-0.012387	-0.00033786	0.36245	5.8647	-5.3376	9.4664E-005	0.00048855
sC ₇	-0.0051385	0.0032057	0.062577	3.1494	-1.8682	-0.00078968	0.00013650
sCOKE	-0.084362	-0.24005	12.993	3.9865	-18.939	0.019362	0.0041815

TABLEAU 5

Estimation de F avec $M_1 = M_1^$.*

F	H ₂	CH ₄	C ₃ H ₆	C ₃ H ₈	C ₄ H ₁₀	tempsej	TEMPE
sH ₂	-7.2550	-8.1275	-70.456	345.43	-376.49	-12.452	2.7628
sCH ₄	5.0802	5.9904	44.718	-212.51	237.51	7.6618	-1.7005
sC ₂ H ₄	-0.19192	-0.17904	-1.6385	8.8055	-9.8132	-0.28903	0.064305
sC ₂ H ₂	-0.87635	-0.86829	-7.5182	38.462	-42.118	-1.3297	0.29585
sC ₂ H ₆	0.022375	0.022658	0.19601	-0.95432	0.37437	0.033592	-0.0073888
sC ₃ C ₅	-0.10894	-0.10349	-0.82787	4.4457	-4.3150	-0.16220	0.035757
sC ₆ H ₆	-0.23747	-0.23166	-1.9243	10.128	-10.965	-0.35477	0.078614
sC ₇	-0.12875	-0.12588	-0.99638	5.4561	-5.1623	-0.19342	0.042636
sCOKE	-0.74555	-0.76011	-5.9642	27.947	-31.954	-1.1382	0.25329

TABLEAU 6

Estimation de F par régression ordinaire.

F	H ₂	CH ₄	C ₃ H ₆	C ₃ H ₈	C ₄ H ₁₀	tempsej	TEMPE
sSH ₂	0.693	0.104	2.68	32.1	-37.0	-0.0297	0.0371
sCH ₄	0.144	0.870	0.867	-16.0	19.4	-0.0153	-0.0123
sC ₂ H ₄	-0.00556	0.0132	0.0470	1.39	-1.42	0.00128	0.000564
sC ₂ H ₂	-0.0224	0.0167	0.160	4.77	-5.59	0.00112	0.00350
sC ₂ H ₆	-0.000191	0.000570	0.0121	-0.0328	0.285	-7.06e-05	2.21e-05
sC ₃ C ₅	-0.00163	0.00220	0.0148	-0.199	0.0365	-0.000121	0.000129
sC ₆ H ₆	-0.00621	0.00276	0.0106	0.432	0.459	0.00109	0.000419
sC ₇	-0.000874	0.000476	-0.00565	-0.190	1.20	0.000393	4.28e-05
sCOKE	-0.0218	0.00340	0.440	-0.378	0.0906	-0.00250	0.00367

Quelques commentaires.

Un des objectifs du procédé est de produire de l'acétylène à partir du gaz naturel. Examinons donc plus particulièrement les estimations des paramètres concernant le débit d'acétylène (C_2H_2) en sortie du four. On peut noter que C_3H_8 (respectivement C_4H_{10}) est le constituant qui a un coefficient positif (respectivement négatif) le plus élevé en valeur absolue quelle que soit la méthode d'estimation. Cela provient essentiellement des faibles débits de ces constituants dans le mélange en entrée. Pour obtenir plus d'acétylène en sortie, les estimations données dans les tableaux 4 et 6 semblent indiquer qu'il faut augmenter la température et le temps de séjour car les paramètres correspondants sont positifs (bien que proches de zéro, mais cela est dû au fait que ces variables ont des valeurs très élevées relativement aux autres variables). Or, les simulations du paragraphe 6 ont montré qu'il faut prendre les résultats de la régression classique avec précaution. En particulier, la valeur du coefficient du temps de séjour diminue quand on l'estime en supposant le modèle à erreur avec la métrique identité, et même cette estimation est négative quand on cherche à ce rapprocher de la métrique optimale (tableau 5), ce qui signifierait au contraire qu'il faut plutôt laisser le mélange moins longtemps dans le four pour augmenter la production d'acétylène. Cela est cohérent avec le fait que le coefficient de corrélation entre C_2H_2 et le temps de séjour est aussi négatif ($-0,4$). On peut remarquer par ailleurs que les paramètres sont bien plus grands quand il sont estimés en supposant le modèle à erreurs (tableaux 4 et 5). Cela peut faire craindre une certaine instabilité de ces estimations et suggère d'utiliser des techniques de protection, par exemple celles présentées au paragraphe 3, dès que leurs propriétés auront été mieux établies.

Appendice

Nous donnons ici les démonstrations des propositions 3, 4, 5 et 6.

On notera $\mathcal{E}_q$ l'ensemble des sous-espaces vectoriels de $\mathbb{R}^m$ de dimension q .

Démonstration de la proposition 3 :

La densité jointe de $(Z_i \ i = 1, \dots, n; v_j \ j = 1, \dots, p; z_i \ i = 1, \dots, n)$ est donnée par :

$$L(Z, z, v) = (2\pi\sigma^2)^{-(1/2)m(n+p)} |\Gamma|^{-n/2} |\Sigma|^{-p/2} \exp \left(-\frac{1}{2\sigma^2} \left(\sum_{i=1}^n \|Z_i - z_i\|_{\Gamma^{-1}}^2 + \sum_{j=1}^p \|v_j - \mu_j\|_{\Sigma^{-1}}^2 \right) \right)$$

La densité de (v, z) conditionnelle à Z est proportionnelle à $L(Z, z, v)$, la constante de proportionnalité ne dépendant pas de (v, z) . Maximiser cette densité en (v, z) revient donc à minimiser, en tenant compte de $\Sigma^{-1} = a\Gamma^{-1}$:

$$\sum_{i=1}^n \|Z_i - z_i\|_{\Gamma^{-1}}^2 + a \sum_{j=1}^p \|v_j - \mu_j\|_{\Gamma^{-1}}^2 \quad (a)$$

Notons $F(v)$ le sous-espace engendré par $v_1, \dots, v_p$; rappelons que $E_q = E(v)$ est le Γ^{-1} -orthogonal complémentaire de $F(v)$. Puisque tout z_i appartient à E_q , il vient :

$$\|Z_i - z_i\|_{\Gamma^{-1}}^2 = \|\Pi_{F(v)}^{\Gamma^{-1}} Z_i\|_{\Gamma^{-1}}^2 + \|\Pi_{E(v)}^{\Gamma^{-1}} Z_i - z_i\|_{\Gamma^{-1}}^2 \quad \text{pour tout } i = 1, \dots, n.$$

Puisque tout v_j appartient à $F(v)$, il vient :

$$\|v_j - \mu_j\|_{\Gamma^{-1}}^2 = \|v_j - \Pi_{F(v)}^{\Gamma^{-1}} \mu_j\|_{\Gamma^{-1}}^2 + \|\Pi_{E(v)}^{\Gamma^{-1}} \mu_j\|_{\Gamma^{-1}}^2 \quad \text{pour tout } j = 1, \dots, p.$$

Pour $E(v)$ (et $F(v)$) donnés, on minimise donc l'expression (a) en prenant : $\hat{z}_i = \Pi_{E(v)}^{\Gamma^{-1}}(Z_i)$ et $\hat{v}_j = \Pi_{F(v)}^{\Gamma^{-1}}(\mu_j)$, pour tout i et tout j .

Il reste donc à déterminer le sous-espace $E_q = E(v)$ qui minimise :

$$\begin{aligned} \sum_{i=1}^n \|\Pi_{E_q^\perp}^{\Gamma^{-1}} Z_i\|_{\Gamma^{-1}}^2 + a \sum_{j=1}^p \|\Pi_{E_q}^{\Gamma^{-1}} \mu_j\|_{\Gamma^{-1}}^2 \\ = \sum_{i=1}^n \|\Pi_{E_q^\perp}^{\Gamma^{-1}} Z_i\|_{\Gamma^{-1}}^2 + a \sum_{j=1}^p \|\mu_j\|_{\Gamma^{-1}}^2 - a \sum_{j=1}^p \|\Pi_{E_q^\perp}^{\Gamma^{-1}} \mu_j\|_{\Gamma^{-1}}^2 \end{aligned}$$

Le deuxième terme du second membre de l'expression précédente est constant et on a :

$$\begin{aligned} \sum_{i=1}^n \|\Pi_{E_q^\perp}^{\Gamma^{-1}} Z_i\|_{\Gamma^{-1}}^2 &= \text{Tr} \left[\sum_{i=1}^n Z_i Z_i' \Gamma^{-1} \Pi_{E_q^\perp}^{\Gamma^{-1}} \right] \\ \text{et } \sum_{j=1}^p \|\Pi_{E_q^\perp}^{\Gamma^{-1}} \mu_j\|_{\Gamma^{-1}}^2 &= \text{Tr} \left[\sum_{j=1}^p \mu_j \mu_j' \Gamma^{-1} \Pi_{E_q^\perp}^{\Gamma^{-1}} \right] \end{aligned}$$

En notant $V = (1/n) \sum_{i=1}^n Z_i Z_i'$ et $W = \sum_{j=1}^p \mu_j \mu_j'$, on cherche donc :

$$\text{ArgMin}_{E_q \in \mathcal{E}_q} \text{Tr} \left[\left(V - \frac{a}{n} W \right) \Gamma^{-1} \Pi_{E_q^\perp}^{\Gamma^{-1}} \right]$$

d'où le résultat.

Démonstration de la proposition 4 :

On cherche $\text{ArgMin} \sum_{i=1}^n \|Z_i - z_i\|_M^2$ sous les contraintes $z_i \in E_q, E_q \in \mathcal{E}_q, E_q \subset H$.

On a :

$$\begin{aligned} \|Z_i - z_i\|_M^2 &= \|(I - \Pi_H^M)Z_i\|_M^2 + \|\Pi_H^M Z_i - z_i\|_M^2 \quad \text{pour tout } i = 1, \dots, n. \\ \|Z_i - \Pi_H^M Z_i\|_M^2 &= \|\Pi_{E_q}^M (I - \Pi_H^M)Z_i\|_M^2 + \|\Pi_{E_q^\perp}^M (I - \Pi_H^M)Z_i\|_M^2 \quad \text{ne dépend} \end{aligned}$$

des paramètres, car $E_q \subset H$ entraîne :

$$\Pi_{E_q}^M \Pi_H^M = \Pi_H^M \Pi_{E_q}^M = \Pi_{E_q}^M, \Pi_{E_q}^M (I - \Pi_H^M) = 0 \text{ et } \Pi_{E_q^\perp}^M (I - \Pi_H^M) = I - \Pi_H^M$$

Par ailleurs :

$$\|\Pi_H^M Z_i - z_i\|_M^2 = \|\Pi_{E_q}^M \Pi_H^M Z_i - z_i\|_M^2 + \|(I - \Pi_{E_q}^M) \Pi_H^M Z_i\|_M^2.$$

On annule le premier terme en prenant $z_i = \Pi_{E_q}^M \Pi_H^M Z_i = \Pi_{E_q}^M Z_i$.

Il reste donc à minimiser :

$$\sum_{i=1}^n \|(I - \Pi_{E_q}^M) \Pi_H^M Z_i\|_M^2 = \text{Tr} \left[\left(\sum_{i=1}^n \Pi_H^M Z_i (\Pi_H^M Z_i)' M \Pi_{E_q^\perp}^M \right) \right]$$

d'où le résultat annoncé.

Démonstration de la proposition 5 :

On a ici :

$$\begin{aligned} \|Z_i - z_i\|_M^2 &= (Z_i - z_i)' M (Z_i - z_i) \\ &= [(R_i - r_i)' | 0] \begin{bmatrix} M_1 & M_{12} \\ M_{21} & M_2 \end{bmatrix} \begin{bmatrix} R_i - r_i \\ 0 \end{bmatrix} = \|R_i - r_i\|_{M_1}^2 \end{aligned}$$

On cherche donc le minimum de $\sum_{i=1}^n \|R_i - r_i\|_{M_1}^2$ en $r_i (i = 1, \dots, n)$, T_R et T_S tels que :

$$T_R' r_i + T_S' s_i = 0 \quad \text{pour tout } i = 1, \dots, n. \quad (\text{contrainte (5.2)})$$

En écrivant $(R - r) = (R - r)P_s + (R - r)(I - P_s)$, on a :

$$\sum_{i=1}^n \| R_i - r_i \|_{M_1}^2 = \text{Tr}[(R - r)'M_1(R - r)] = A + B + C \quad \text{avec}$$

$$A = \text{Tr} \left[((R - r)P_s)'M_1(R - r)P_s \right]$$

$$B = \text{Tr} \left[((R - r)(I - P_s))'M_1(R - r)(I - P_s) \right]$$

$$C = 2\text{Tr} \left[((R - r)(I - P_s))'M_1(R - r)P_s \right]$$

Or P_s est un projecteur orthogonal d'où $P_s(I - P_s) = 0$ et $C = 0$.

Remarquons que A (resp. B) ne fait intervenir que rP_s (resp. $r(I - P_s)$) et cherchons la matrice r optimale, soit $\hat{r}$, en obtenant successivement $\hat{r}P_s$ et $\hat{r}(I - P_s)$.

A est nul, donc minimum, pour $\hat{r}P_s = RP_s$.

Posons $R^* = R(I - P_s)$ et $r^* = r(I - P_s)$. On a

$$B = \text{Tr} \left[(R^* - r^*)M_1(R^* - r^*) \right]$$

et (5.2) entraîne, en multipliant à droite par $I - P_s$:

$$T'_R r^* = 0 \quad (*)$$

Sous cette contrainte, le minimum de B s'obtient par la proposition 1 :

on a $\hat{r}^* = \Pi_{\hat{E}_{q_1}}^{M_1} R^* = \Pi_{\hat{E}_{q_1}}^{M_1} R(I - P_s)$ où $\hat{E}_{q_1}$ est engendré par les vecteurs propres associés aux q_1 plus grandes valeurs propres de $V_1 M_1$ avec $V_1 = \frac{1}{n} R^* R'^* = \frac{1}{n} R(I - P_s)R'$.

Rassemblant les deux résultats de minimisation, il vient :

$$\hat{r} = \hat{r}^* + RP_s = \Pi_{\hat{E}_{q_1}}^{M_1} R(I - P_s) + RP_s = \Pi_{\hat{E}_{q_1}}^{M_1} R + \Pi_{\hat{E}_{q_1}^\perp}^{M_1} RP_s.$$

Cependant, $\hat{r}$ minimise $A + B$ sous la contrainte (*) plus faible que (5.2). Il reste donc à montrer que (5.2) peut être satisfaite. Pour cela, vérifions qu'il suffit de prendre $\hat{T}_S = -(ss')^{-1} sR'T_R$. On a alors, en effet, puisque $\hat{T}'_R r^* = 0$:

$$\hat{T}'_R \hat{r} + \hat{T}'_S s = \hat{T}'_R RP_s - \hat{T}'_R R s'(ss')^{-1} s = 0.$$

A partir de $\hat{T} = \begin{bmatrix} \hat{T}_R \\ \hat{T}_S \end{bmatrix}$, qu'on peut encore écrire $\hat{T} = \begin{bmatrix} \hat{T}_1 \\ \hat{T}_2 \end{bmatrix}$ avec $\hat{T}_1$ de

dimension $p \times p$ et $\hat{T}_2$ de dimension $q \times p$, on obtient $\hat{F}$ comme dans le modèle initial

(Cf. (3.2)) :

$$\hat{F}' = -\hat{T}_2 \hat{T}_1^{-1}$$

Références

- ANDERSON, T.W. (1984). Estimating Linear Statistical Relationships. *Ann. Statist.*, 12, 1-45.
- BESSE, P., CAUSSINUS, H., FERRE, L., FINE, J. (1987). Sur l'utilisation optimale de l'Analyse en Composantes Principales. *C. R. Ac. Sc. Paris*, 291, A, 319-322.
- BESSE, P., CAUSSINUS, H., FERRE, L., FINE, J. (1988). Principal Components Analysis and optimization of graphical displays. *Statistics*, 19, 301-312.
- BOULARAN, J. (1993). Deux modèles de régression : Etudes théoriques et exemples. *Thèse, Université Paul Sabatier, Toulouse.*
- BOX, G.E.P., HUNTER, W.G., MACGREGOR, J.F. and ERJAVEC, J., (1973). Some problems associated with the analysis of multiple response data. *Technometrics*, 15, 33-51.
- CAUSSINUS, H. (1984). Analyse en Composantes Principales. Quelques réflexions sur la part des modèles probabilistes en analyse des données. *Publ. du Lab. de Stat. et Proba.*, 01-84, Université Paul Sabatier, Toulouse.
- CAUSSINUS, H. (1985 a). Models and uses of Principal Component Analysis (with discussion). *Multidimensional Data Analysis*, J. de Leeuw *et al.* (eds), 149-178, DSWO Press, Leiden.
- CAUSSINUS, H. (1985 b). Quelques réflexions sur la part des modèles probabilistes en Analyse des Données. *Data Analysis and Informatics IV*, E.Diday *et al.* (eds), 151-165, North- Holland, Amsterdam.
- CAZES, P. (1975). Protection de la régression par utilisation de contraintes linéaires et non linéaires. *Revue de Statistique Appliquée*, XXIII, 3, 37-57.
- DROESBEKE, J. J., FICHET, B. et TASSI, Ph. (1992). *Modèles pour l'analyse des données multidimensionnelles*. Economica, Paris.
- FINE, J. and POUSSE, A. (1992). Asymptotic study of the multivariate functional model. Application to the metric choice in Principal Components Analysis. *Statistics*, 23, 63-83.
- FULLER, W.A. (1987). *Measurement Error Models*. John Wiley, New York.
- GLESER, L.J. (1981). Estimation in a multivariate errors in variables regression model : Large sample results. *Ann. Statist.* 9, 24-44.
- HOERL, A.E. and KENNARD, R.W. (1970). Ridge Regression : Bounded estimation for non orthogonal problems, *Technometrics*, 12,3, 591-612.
- LEBART, L., MORINEAU, A. & FÉNELON, J.P. (1979). *Traitement des données statistiques*. Dunod, Paris.

- LEGENDRE, P. (1977). Régression sur variables entachées d'erreurs. *Thèse de 3ème cycle. Université Paul Sabatier, Toulouse.*
- MALINVAUD, E. (1981). *Méthodes statistiques de l'économétrie*. Dunod, Paris.
- RAGOT, J. et AUBRUN, M. (1982). Application de la régression orthogonale sous contrainte linéaire à un problème d'équilibrage du bilan matière. *Revue de Statistique Appliquée*, XXX, 2, 45-56.
- SEBER, G.A.F. (1984). *Multivariate observations*. John Wiley, New York.