ANNALES D'ÉCONOMIE ET DE STATISTIQUE. – N° 57 – 2000

Une nouvelle solution au problème de KLOEK et MENNES

Philippe CASIN *

RÉSUMÉ. – La technique des doubles moindres carrés n'est pas utilisable lorsqu'il y a beaucoup de variables et peu d'observations. Pour résoudre ce problème, KLOEK et MENNES ont proposé de remplacer certaines variables par leurs composantes principales.

Cet article traite de la façon de déterminer des combinaisons linéaires de variables, qui peuvent être obtenues autrement qu'en utilisant l'analyse en composantes principales.

KLOEK and MENNES's Problem: A New Solution

ABSTRACT. – When there are a great number of variables and only a small number of observations, two-stage least-squares are not practicable. To overcome this problem, KLOEK and MENNES proposed to replace some variables by their principal components.

The aim of this paper is to discuss the way to choose linear combinations of the original variables, not necessarily by using principal component analysis.

---

* Ph. CASIN : Centre d'Analyse des Dynamiques Économiques (CADE), UFR Droits, Économie et Administration, Ile du Saulcy, 57045 Metz Cedex.

1 Introduction

Il peut arriver, lorsqu'on estime de grands modèles macro-économétriques, que le nombre de variables prédéterminées soit élevé par rapport au nombre d'observations. L'utilisation des doubles moindres carrés devient alors délicate, la première étape de cette technique consistant à effectuer des régressions dans lesquelles les variables explicatives sont les variables prédéterminées.

Il y a quarante ans, KLOEK et MENNES [1960] (voir aussi JOHNSTON [1972], JOLLIFFE [1986], MALINVAUD [1978]) ont proposé une méthode pour surmonter cette difficulté, consistant à remplacer dans les régressions les variables prédéterminées par un petit nombre de leurs composantes principales. Cette méthode est un cas particulier de la méthode des variables instrumentales, les variables instrumentales étant obtenues par des régressions où les variables explicatives sont les composantes principales.

Les propriétés mais aussi les limites de la régression sur composantes principales sont mieux connues aujourd'hui qu'il y a quarante ans (pour un exposé complet de la régression sur composantes principales : JOLLIFFE [1986]) : l'analyse en composantes principales normée fournit les variables synthétiques les mieux liées linéairement aux variables explicatives, ce qui constitue une réponse aux problèmes de multicolinéarité, mais rien ne garantit que ces composantes principales aient une corrélation élevée avec la variable à expliquer.

Aussi, une autre méthode de régression biaisée, la régression PLS (*Partial Least-Squares*), technique intermédiaire entre l'analyse en composantes principales et la régression multiple a été développée ces dernières années (par exemple, ALMOY [1996], HELLAND [1990], HELLAND et ALMOY [1994], PALM et IEMMA [1995] ; pour un historique de la technique : TENENHAUS [1995] ; pour une extension à plus de deux tableaux de données CASIN [1996]).

La régression PLS détermine la combinaison linéaire des variables explicatives ayant, sous une contrainte de normalisation, la plus forte covariance possible avec la variable à expliquer et constitue donc un compromis entre la régression multiple (la corrélation entre la variable à expliquer et la combinaison linéaire de variables explicatives doit être la plus élevée possible) et l'analyse en composantes principales (la variance de la combinaison linéaire de variables explicatives doit être la plus élevée possible, sous une contrainte de normalisation).

Ces progrès de la statistique permettent un réexamen du problème posé par KLOEK et MENNES ; l'objet de cet article est de discuter, à la lumière des connaissances nouvelles, de la manière de construire des variables instrumentales.

L'organisation de l'article est la suivante : la section 2 expose le problème de KLOEK et MENNES ; dans la section 3, les solutions proposées par ces deux auteurs sont décrites et discutées ce qui permet de dégager des critères pour juger de la pertinence des variables instrumentales ; une nouvelle méthode est alors proposée dans la section 4. Enfin, dans la section 5, le modèle de Klein I (KLEIN, [1950]) est utilisé pour illustrer cette nouvelle technique et comparer ses résultats avec ceux de la méthode de KLOEK et MENNES.

2 Le problème

2.1 Notations

KLOEK et MENNES considèrent une équation structurelle extraite d'un système d'équations simultanées :

$$y = Ya + X_1 b + u$$

où :

– y est le vecteur colonne des T observations de la variable « *à expliquer* »

– Y est la matrice $T \times M$ des observations des M autres variables endogènes de l'équation ;

– X_1 est la matrice $T \times L$ des observations des L variables prédéterminées de l'équation ;

– u est un vecteur colonne dont les T lignes sont les perturbations de l'équation, s^2 désignant la variance des composantes de u ;

– a et b sont respectivement le vecteur-colonne à M lignes des coefficients des variables Y et le vecteur-colonne à L lignes des coefficients des variables X_1.

Soit X_2 la matrice $T \times (N-L)$ des observations des $N-L$ variables prédéterminées du système qui sont exclues de l'équation et soit X la matrice des observations de toutes les variables prédéterminées : $X = [X_1, X_2]$.

Enfin, Y_k, $X_{1,i}$ et $X_{2,h}$ désignent respectivement la k-ième variable Y, la i-ième variable X_1 et la h-ième variable X_2. Par commodité, il est supposé que les colonnes de Y et celles de X sont centrées réduites. $R(z, v)$ et $\mathrm{Cov}(z,v)$ sont respectivement le coefficient de corrélation et la covariance entre deux variables z et v. $\mathrm{Var}(z)$ est la variance de la variable z.

2.2 Le problème

La méthode des doubles moindres carrés comporte deux étapes successives :

(1) Calcul de Y^*, la matrice des projections des variables Y sur l'espace engendré par les colonnes de X. Y_k^* désigne la k-ième colonne de Y^*.

(2) Calcul de y^* (et donc de a^* et b^*, qui sont des estimations de a et b), y^* étant la projection de y sur l'espace engendré par les colonnes de X_1 et par les colonnes de Y^*.

Les doubles moindres carrés constituent un cas particulier de la méthode des variables instrumentales, les variables instrumentales étant alors les projections des variables Y sur l'espace engendré par les colonnes de X.

À la première étape, si N est plus grand que T, il n'est pas possible de calculer l'inverse de $X'X$; comme le notent KLOEK et MENNES [1960] (voir aussi KLEIN [1969]), même si N est inférieur à T, la situation est délicate lorsqu'il y a seulement un petit nombre de degrés de liberté. Remplacer X_2 par un petit nombre de variables z^j, $j = 1, \ldots, J$, combinaisons linéaires des variables X et deux à deux non corrélées permet de surmonter cette difficulté.

La méthode obtenue est, alors, un cas particulier de la méthode des variables instrumentales, les variables instrumentales étant les projections de Y sur l'espace engendré par les colonnes de X_1 et par les variables z^j.

Les variables z^j proposées par KLOEK et MENNES sont obtenues en effectuant une analyse en composantes principales.

3 Les stratégies de KLOEK et MENNES pour sélectionner les composantes principales

3.1 Les solutions de KLOEK et MENNES

Il s'agit donc de construire des variables instrumentales permettant d'estimer les coefficients de l'équation de la section 2.1.

Les premières composantes principales du tableau X_2 expliquent un pourcentage élevé de la variance des variables X_2. Aussi, dans la première méthode proposée par KLOEK et MENNES, X_2 est remplacée par la matrice de ses premières composantes principales. Mais il se peut que ces composantes principales soient très corrélées avec quelques-unes des variables X_1 et par conséquent ne soient pas très utiles : aussi, la deuxième méthode proposée par KLOEK et MENNES utilise les composantes principales des résidus (au sens des moindres carrés) de la régression des colonnes de X_2 par les colonnes de X_1.

Le but des troisième et quatrième méthodes proposées par KLOEK et MENNES est de réduire le temps de calcul ; ce problème est beaucoup moins important aujourd'hui que dans les années 1960 ; aussi, seule la seconde méthode de KLOEK et MENNES sera examinée ici.

3.2 Quelques remarques sur la méthode de KLOEK et MENNES

Remarque 1 : KLOEK et MENNES utilisent les vecteurs propres de la matrice de variances-covariances des résidus des régressions des variables X_2 centrées-réduites sur les variables X_1. Ils notent : «… *« large » residuals are indicators of important parts of* X_2 *(uncorrelated with* X_1*) and so play an important role in the relevant reduced-form equation.* »

Soit F_h^1 le résidu (au sens des moindres carrés) de la régression de la h-ième colonne de X_2 par les colonnes de X_1 et soit F^1 la matrice de dimension $T \times (N - L)$ dont les colonnes sont les F_h^1, pour $h = 1,\ldots, N - L$.

Ces résidus F_h^1 peuvent être instables s'ils ont une variance faible. En effet, soit D_h une « *petite* » perturbation non corrélée avec la variable $X_{2,h}$ et avec les variables X_1 et soit L_h le résidu de la régression de $(X_{2,h} + D_h)$ par les variables X_1.

Puisque D_h est non corrélé avec les variables X_1 et avec $X_{2,h}$: $L_h = F_h^1 + D_h$ et :

$$R^2(F_h^1,\, L_h) = \frac{\mathrm{Cov}^2(F_h^1,\, L_h)}{\mathrm{Var}(F_h^1)\,\mathrm{Var}(L_h)} = \frac{\mathrm{Cov}^2(F_h^1,\, F_h^1 + D_h)}{\mathrm{Var}(F_h^1)\,\mathrm{Var}(F_h^1 + D_h)}$$

$$= \frac{\mathrm{Var}^2(F_h^1)}{\mathrm{Var}(F_h^1)\,(\mathrm{Var}(F_h^1) + \mathrm{Var}(D_h))}$$

$$= \frac{1}{1 + \dfrac{\mathrm{Var}(D_h)}{\mathrm{Var}(F_h^1)}}$$

Si $\mathrm{Var}(F_h^1)$ est petit par rapport à $\mathrm{Var}(D_h)$, alors la valeur de $R^2\,(F_h^1,\, L_h)$ est petite et l'estimation du h-ième résidu est imprécise.

Donc, pour être stable, une variable z^1, combinaison linéaire des variables F^1, doit être fortement corrélée avec les résidus dont la variance est élevée, c'est-à-dire que :

$$\sum_{h=1}^{N-L} R^2(z^1,\, F_h^1)\,\mathrm{Var}(F_h^1)$$

doit avoir une valeur élevée.

Puisque z^1 est non corrélée avec les variables X_1 :

$$\sum_{h=1}^{N-L} R^2(z^1,\, F_h^1)\,\mathrm{Var}(F_h^1) = \sum_{h=1}^{N-L} R^2(z^1,\, X_{2,h})$$

Notons que cette dernière valeur est maximale si z^1 est la première composante principale de KLOEK et MENNES. En effet :

$$\sum_{h=1}^{N-L} R^2(z^1,\, X_{2,h}) = \frac{T \displaystyle\sum_{h=1}^{N-L} \mathrm{Cov}^2(z^1,\, F_h^1)}{(z^1)'\, z^1}$$

$$= \frac{(z^1)' F^1 (F^1)' z^1}{T (z^1)' z^1}$$

est maximum lorsque la variable z^1 est le premier vecteur propre de $\dfrac{1}{T} F^1 (F^1)'$. La composante principale z^j de KLOEK et MENNES est la combinaison des variables F^1, non corrélée à $z^1, \ldots, z^{j-1}$, et rendant maximum

$$\sum_{h=1}^{N-L} R^2(z^j,\, X_{2,h}).$$

Cette variable z^j s'écrit alors $z^j = F^1 a^j$, a^j étant le j-ème vecteur propre normé de $\frac{1}{T}(F^1)'F^1$.

Remarque 2 : Lorsque $(X'X)$ n'est pas inversible, l'emploi d'une inverse généralisée permet de déterminer la projection d'une variable Y sur l'espace engendré par les colonnes de X ; le projecteur sur l'espace engendré par les colonnes de X est égal à la somme du projecteur sur l'espace engendré par les colonnes de X_1 et des $N-L$ projecteurs sur les composantes principales du tableau F^1 associées à des valeurs propres non nulles (les $(N-L)$ premiers vecteurs propres de $F^1(F^1)'$) ; on retrouve pour la projection de Y les problèmes de stabilité décrits dans la remarque précédente, les dernières de ces $N-L$ composantes principales étant des combinaisons linéaires des variables F^1 très faiblement liées linéairement aux variables X_2.

Remarque 3 : Puisque X_2 est remplacée par quelques composantes principales, Y^* est la matrice des projections des variables Y sur l'espace engendré par les colonnes de X_1 et par ces composantes principales. La méthode de KLOEK et MENNES pose le problème suivant : si une variable Y_k est non corrélée avec les composantes principales, alors Y_k^* est une combinaison linéaire des variables X_1 et il n'est pas possible de calculer des estimations des paramètres a et b. Si la corrélation entre Y_k et les composantes principales est faible, alors Y_k^* est proche d'une combinaison linéaire des variables X_1 et l'estimation des paramètres a et b n'est pas très précise.

Comme le notent KLOEK et MENNES, les composantes principales doivent avoir une corrélation nulle avec les variables X_1. Mais, même si une composante principale est non corrélée avec les variables X_1, il n'est pas sûr qu'elle apporte une contribution à la procédure d'estimation.

Remarque 4 : Considérons, maintenant, un cas plus général : Y^* est la projection de Y sur l'espace engendré par les colonnes de X_1 et par les variables z^j, qui sont des combinaisons linéaires des variables F^1, mais plus nécessairement des composantes principales.

Cette procédure est un cas particulier de la méthode des variables instrumentales, les variables instrumentales étant les colonnes de Y^* et fournit, sous les hypothèses usuelles, des estimateurs convergents. Puisque les colonnes de Y^* sont les projections des variables Y sur l'espace engendré par les variables X_1 et par les variables z^j, les variances asymptotiques de a^* et b^* (qui sont des estimateurs de a et b, voir section 2.2) sont les éléments diagonaux de la matrice de variances-covariances suivante (JOHNSTON [1972], MALINVAUD [1978]) :

$$V(a^*,b^*) = s^2 \begin{bmatrix} (Y^*)'(Y^*) & (Y^*)'(X_1) \\ (X_1)'(Y^*) & (X_1)'(X_1) \end{bmatrix}^{-1}$$

Par exemple, calculons la variance du premier paramètre de a^*, a_1^* :

$$V(a^*,b^*) = s^2 \begin{bmatrix} (Y_1^*)'(Y_1^*) & (Y_1^*)'(O) \\ (O)'(Y_1^*) & (O)'(O) \end{bmatrix}^{-1}$$

où O désigne la matrice $T \times (M - 1 + L)$ suivante :

$$O = [Y_2^*, \ldots, Y_M^*, X_{1,1}, \ldots, X_{1,L}]$$

et, par conséquent, la variance de a_1^* est :

$$\text{Var}\,(a_1^*) = s^2[(Y_1^*)'(Y_1^*) - (Y_1^*)'O((O)'O)^{-1}(O)'(Y_1^*)]^{-1}$$

Si $M = 1$, les variables O se confondent avec les variables X_1. Alors, puisque Y_1^* est une combinaison linéaire des variables X_1 et des variables z, et puisque les variables X_1 sont non corrélées avec les variables z :

$$\text{Var}\,(a_1^*) = s^2\,[(Y_1^*)'z(z'z)^{-1}\,(z)'(Y_1^*)]^{-1}$$

z étant la matrice dont les colonnes sont les variables z, et, par conséquent :

$$\text{Var}\,(a_1^*) = s^2\,[(Y_1)'z(z'z)^{-1}(z)'(Y_1)]^{-1} = s^2[R_z(Y_1))^2]^{-1}$$

où $R_z(Y_1)$ est le coefficient de corrélation multiple entre Y_1 et les variables z.

Par conséquent, lorsque $M = 1$, Var(a_1^*) a une valeur « *faible* » si et seulement si $R_z(Y_1)$ a une valeur « *élevée* ».

Lorsque $M > 1$:

$$(Y_1^*)'O((O)'O)^{-1}(O)'(Y_1^*) \geqslant (Y_1^*)'X_1((X_1)'X_1)^{-1}(X_1)'(Y_1^*)$$

et donc :

$$\text{Var}\,(a_1^*) \geqslant s^2[(R_z(Y_1))^2]^{-1}$$

Par conséquent, quand $M > 1$, Var (a_1^*) a une valeur « *faible* » seulement si $R_z\,(Y_1)$ a une valeur « *élevée* ».

Aussi, la conclusion de cette remarque est que $R_z(Y_1)$ doit avoir une valeur élevée. Il est intéressant de noter que la valeur de $R_z\,(Y_1)$ est maximale dans le cas des doubles moindres carrés.

4 Une nouvelle façon de déterminer les variables instrumentales

4.1 Les critères de détermination des pseudo-composantes principales

Les combinaisons linéaires déterminées par la nouvelle méthode proposée dans cet article sont appelées « *pseudo-composantes principales* ». Ces pseudo-composantes principales constituent un ensemble de variables synthétiques non corrélées entre elles, calculées dans le but de remplacer les variables X_2 et donc de construire des variables instrumentales pour estimer les coefficients de l'équation du paragraphe 2.1. D'après les remarques de la section précédente, z^j, la j-ième pseudo-composante principale doit être non corrélée avec les variables X_1.

z^j est une combinaison linéaire des variables X, non corrélée avec les variables X_1 et non corrélée avec les variables $z^1, \ldots, z^{j-1}$.

Par conséquent, z^j est une combinaison linéaire des variables F_h^j, F_h^j étant le résidu de la régression de la h-ième colonne centrée-réduite de X_2 par les colonnes de X_1 et par $z^1, \ldots, z^{j-1}$.

Soit F^j la matrice $T \times (N - L)$ dont les colonnes sont les variables F_j^h. z^j s'écrit :

$$z^j = \sum_{h=1}^{N-L} F_h^j \, a_h^j = F^j a^j$$

où a^j est un vecteur à $N - L$ lignes tel que $(a^j)' a^j = 1$.

– ***D'après la remarque 4 :*** $A^j = R_z^2(Y_k)$ doit avoir une valeur élevée ; comme les variables z^j sont deux à deux non corrélées :

$$R_z^2 \, (Y_k) = \sum_{k=1}^{M} R^2(z^j, \, Y_k)$$

et donc :

$$A^j = \sum_{k=1}^{M} R^2(z^j, \, Y_k)$$

doit avoir une valeur élevée.

Soit E^1 la matrice $T \times M$ des résidus de la régression des variables centrées réduites Y_k, $k = 1, \ldots, M$ par les variables X_1, et soit E_k^1 la k-ième colonne de E^1.

Soit E^j la matrice $T \times M$ des résidus de la régression des variables Y_k, $k = 1, \ldots, M$ par les variables X_1 et par les variables $z^1, \ldots, z^{j-1}$, et soit E_k la k-ième colonne de E^j. Alors :

$$A^j = \sum_{k=1}^{M} R^2(z^j, \, E_k^j) \, \mathrm{Var} \, (E_k^j)$$

Preuve :

$$A^j = \frac{\sum_{k=1}^{M} \mathrm{Cov}^2(z^j, \, Y_k)}{\mathrm{Var}(z^j)}$$

$$= \frac{\sum_{k=1}^{M} (\mathrm{Cov}(z^j, \, E_k^j) + \mathrm{Cov}(z^j, \, -E_k^j + Y_k))^2}{\mathrm{Var} \, (z^j)}$$

Puisque $E_k^j - Y_k$ est une combinaison linéaire des variables X_1 et des variables $z^1, \ldots, z^{j-1}$ et puisque z^j est non corrélée avec les variables X_1 et

avec $z^1, \ldots, z^{j-1}$:

$$A^j = \frac{\sum_{k=1}^{M} \text{Cov}^2(z^j, E_k^j)}{\text{Var}(z^j)} = \sum_{k=1}^{M} R^2(z^j, E_k^j) \text{ Var } (E_k^j)$$

– ***D'après la remarque 1 :***

$$B^j = \sum_{h=1}^{N-L} R^2(z^j, X_{2,h})$$

doit avoir une valeur élevée

z^j étant non corrélé aux variables $z^1, \ldots, z^{j-1}$, B^j s'écrit aussi :

$$B^j = \sum_{h=1}^{N-L} R^2(z^j, F_h^j) \text{ Var } (F_h^j)$$

Puisque (*cf.* Annexe) :

$$\text{Var}^2(z^j) \leqslant \sum_{h=1}^{N-L} \text{Cov}^2 (z^j, F_h^j)$$

et, puisque F_h^j est non corrélé avec les variables X_1 et avec $z^1, \ldots, z^{j-1}$:

$$\sum_{h=1}^{N-L} \text{Cov}^2(z^j, F_h^j) = \sum_{h=1}^{N-L} R^2(z^j, X_{2,h}) \text{ Var } (z^j)$$

d'où : $\text{Var } (z^j) \leqslant \sum_{h=1}^{N-L} R^2(z^j, X_{2,h})$

Si $\text{Var } (z^j)$ a une valeur élevée, alors B^j a une valeur élevée. Aussi, le critère retenu est : $B^{*j} = \text{Var } (z^j)$ doit avoir une valeur élevée.

4.2 Le calcul des pseudo-composantes principales

À l'étape 1, maximiser A^1 ou maximiser B^{*1} sont deux objectifs différents, voire contradictoires. Un compromis consiste à maximiser $A^1 \times B^{*1}$.

Le critère d'optimalité est alors :

$$\max \Big(\sum_{k=1}^{M} R^2(z^1, E_k^1) \text{ Var } (E_k^1) \text{ Var } (z^1)\Big)$$

$$= \max \sum_{k=1}^{M} \text{cov}^2 (z^1, E_k^1) = \max (z^1)' E^1 (E^1)' z^1$$

$$= \max (a^1)'(F^1)' E^1 (E^1)'(F^1)(a^1)$$

sous la contrainte : $(a_1)' a_1 = 1$

La solution, a_1, est le vecteur propre de $(F^1)' E^1 (E^1)' (F^1)$ associé à la valeur propre la plus élevée.

La recherche se poursuit au delà de la première étape ; à l'étape *j*, le compromis consiste à maximiser

$$A^j \times B^{*j} = \sum_{k=1}^{M} \text{cov}^2 (z^j, E_k^j)$$

avec : $z^j = F^j a^j$
sous la contrainte : $(a^j)' a^j = 1$.

Puisque

$$\sum_{k=1}^{M} \text{cov}^2(z^j, E_k^j) = (a^j)'(F^j)' E^j (E^j)' (F^j)(a^j),$$

a^j est le vecteur propre de $(F^j)' E^j (E^j)' (F^j)$ associé à sa valeur propre la plus élevée.

4.3 Combien de pseudo-composantes principales faut-il calculer ?

L'objet de la méthode proposée est d'expliquer la variance des variables E^1 à partir de combinaisons linéaires de variables F^1 possédant une variance suffisante. Aussi, à l'étape *j*, il est utile de tenir compte des indicateurs A^j_{max} (la variance des variables E^1 non expliquée par $z^1, \ldots, z^{j-1}$) et B^j_{max} (la variance des variables F^1 non utilisée par $z^1, \ldots, z^{j-1}$).

Soit $A^j_{\text{max}} = \sum_{k=1}^{M} \text{Var}\,(E_k^j)$

A^j_{max} est donc la variance des variables *Y* qui n'a pas été expliquée par les variables X_1 et par $z^1, \ldots, z^{j-1}$: $R_{X_1}(Y_k)$ étant le coefficient de corrélation multiple entre Y_k et les variables X_1 puisque $z^1, \ldots, z^j$ est un ensemble de variables deux à deux orthogonales et orthogonales aux variables X_1 :

$$\text{Var}\,(E_k^j) = 1 - (R^2_{X_1}(Y_k) + \sum_{g=1}^{j-1} R^2\,(Y_k, z^g))\text{et}$$

$$A^j_{\text{max}} = M - \sum_{k=1}^{M} (R^2_{X_1}(Y_k) + \sum_{g=1}^{j-1} R^2\,(Y_k, z^g))$$

Et comme (voir 4.1) :

$$A^j = \sum_{k=1}^{M} R^2(z^j, E_k^j)\,\text{Var}(E_k^j),$$

alors : $A^j \leqslant A^j_{\text{max}}$.

Si $A^j_{\max}$ a une valeur faible, il n'est pas nécessaire de calculer une pseudo-composante principale supplémentaire.

Soit $B^j_{\max} = \sum_{h=1}^{N-L} \text{Var}\,(F^j_h)$

$B^j_{\max}$ est la variance des variables F^1 qui n'a pas été utilisée par $z^1, \ldots, z^{j-1}$ pour expliquer les variables Y ; $R_{X_1}\,(X_{2,h})$ étant le coefficient de corrélation entre $X_{2,h}$ et les variables X_1, puisque $z^1, \ldots, z^j$ est un ensemble de variables deux à deux orthogonales et orthogonales aux variables X_1 :

$$B^j_{\max} = N - L - \sum_{h=1}^{N-L} (R^2_{X_1}\,(x_{2,h}) + \sum_{g=1}^{j-1} R^2(X_{2,h}, z^g))$$

et : $B^j \leqslant B^j_{\max}$

Les raisons sont les mêmes que celles exposées pour A^j et $A^j_{\max}$.

Si $B^j_{\max}$ a une valeur faible, il n'est pas possible de calculer une pseudo-composante principale supplémentaire.

Remarques :

1. Comme l'ACP, la nouvelle méthode détermine une base orthogonale de l'espace engendré par les $N - L$ résidus de la régression des variables X_2 par les variables X_1 (si toutes les pseudo-composantes principales sont utilisées, la nouvelle méthode se confond avec les doubles moindres carrés). Pour une valeur de h donnée :

$$\sum_{j=1}^{N-L} R^2(z^j,\, X_{2,h}) + R^2_{X_1}\,(X_{2,h}) = 1$$

Alors :

$$\sum_{h=1}^{N-L} \sum_{j=1}^{N-L} R^2(z^j, X_{2,h}) + \sum_{h=1}^{N-L} R^2_{X_1}\,(X_{2,h}) = N - L$$

et :

$$\sum_{j=1}^{N-L} B^j = B^1_{\max}$$

2. Puisque z^j est non corrélé avec $z^1, \ldots, z^{j-1}$ et avec les variables X_1 :

$$E^{j+1}_k = E^j_k - \text{Cov}(z^j,\, E^j_k)\, z^j$$

et donc : $A^j_{\max} - A^{j+1}_{\max} = \sum_{k=1}^{M} R^2\,(Y_k, z^j) = A^j$

d'autre part :

$$F^{j+1}_h = F^j_h - \text{Cov}(z^j, F^j_h)\, z^j$$

et : $B^j_{\max} - B^{j+1}_{\max} = \sum_{h=1}^{N-L} R^2(z^j, X_{2,h}) = B^j$

3. Les pseudo-composantes principales retenues ne sont pas forcément les premières. Supposons, par exemple, que A^1 ait une grande valeur et B^1 une faible valeur : la première combinaison linéaire n'est pas retenue, mais les suivantes peuvent l'être.

4. Puisque le nombre de variables Y est M, il faut au moins M nouvelles variables pour que le calcul de b^* soit possible.

5. Les pseudo-composantes principales ne sont pas forcément des combinaisons linéaires des premières composantes principales.

5 Application au modèle de KLEIN I

5.1 Le modèle

Le modèle utilisé par KLOEK et MENNES pour illustrer les différentes méthodes et comparer leurs performances est le modèle de KLEIN I (KLEIN [1950]). Les équations de ce modèle sont les suivantes :

$$
\begin{aligned}
C &= a_0 + b_{11}\,\Pi + a_{11}\,\Pi_{-1} + a_{12}\,(W_1 + W_2) + u_1 \\
I &= b_0 + a_{21}\,\Pi + b_{21}\,\Pi_{-1} + b_{22}\,K_{-1} + u_2 \\
W_1 &= c_0 + a_{31}(Y + T - W_2) + b_{31}(Y + T - W_2)_{-1} \\
&\quad + b_{32}\,(t - 1931) + u_3 \\
Y + t &= C + I + G \\
Y &= \Pi + W_1 + W_2 \\
K - K_{-1} &= I
\end{aligned}
$$

où C (la consommation), Π (les profits), W_1 (les salaires du secteur privé), W_2 (les salaires du secteur public), I (l'investissement net), K_{-1} (le stock de capital en début de période), Y (le revenu national), T (les impôts indirects), t (l'année), G (la demande des administrations) sont observés durant la période 1921-1941.

Il y a six variables conjointement déterminées : C, I, W_1, Y, Π et K ; les autres variables (Π_{-1}, W_2, T, G, t, $(Y + T - W_2)_{-1}, K_{-1}$) sont des variables prédéterminées.

5.2 Les différentes estimations du modèle de KLEIN I

Les différents estimateurs utilisés sont l'estimateur des doubles moindres carrés (2SLS), l'estimateur de la deuxième méthode de KLOEK et MENNES utilisant une composante principale (KM1) ou deux composantes principales (KM2), l'estimateur de la méthode proposée dans cet article, utilisant une pseudo-composante principale (NM1) ou deux pseudo-composantes principales (NM2).

TABLEAU 1

Résultats de la première équation

	b_{11}^*	a_{11}^*	a_{12}^*	s_1	$R_z(\Pi)$	$R_z(W_1)$
2SLS	0.017 0.131	0.216 0.119	0.810 0.045	1.14	0.484	0.654
KM2	0.010 0.182	0.222 0.160	0.811 0.045	1.14	0.338	0.651
NM2	0.043 0.136	0.189 0.124	0.817 0.044	1.10	0.444	0.629

Commentaire : Pour la méthode de KLOECK et MENNES, les écart-types de b_{11}* et a_{11}* sont plus élevés d'environ 40 pour cent par rapport aux autres méthodes.

TABLEAU 2

Résultats de la deuxième équation

	a_{21}^*	b_{21}^*	b_{22}^*	s_2	$R_z(\Pi)$
2SLS	0.150 0.193	0.615 0.181	– 0.158 0.040	1.31	0.360
KM1	– 0.027 0.458	0.768 0.404	– 0.183 0.074	1.63	0.189
NM1	0.027 0.344	0.721 0.308	– 0.175 0.059	1.52	0.235
KM2	0.155 0.207	0.612 0.192	– 0.157 0.041	1.30	0.346
NM2	0.151 0.200	0.614 0.188	– 0.157 0.041	1.31	0.347

Commentaire : Les résultats de 2SLS, KM2 et NM2 sont presque identiques, car le coefficient de corrélation entre Π et les deux premières composantes principales ou entre Π et les deux premières pseudo-composantes principales est très proche du coefficient de corrélation entre Π et les cinq variables X_2. Les écarts-types sont plus élevés pour KM1 que pour NM1.

TABLEAU 3

Résultats de la troisième équation

	a_{31}^*	b_{31}^*	b_{32}^*	s_3	$R_z\ (Y + T - W_2)$
2SLS	0.150 0.040	0.439 0.043	0.147 0.032	0.130	0.77
KM1	0.357 0.58	0.224 0.061	0.150 0.039	0.90	0.324
NM1	0.388 0.047	0.194 0.050	0.143 0.035	0.83	0.368
KM2	0.395 0.47	0.188 0.050	0.141 0.034	0.81	0.367
NM2	0.415 0.43	0.169 0.046	0.136 0.033	0.78	0.382

Commentaire : Quand on utilise le même nombre de composantes, on obtient avec NM une valeur plus élevée pour $R_z(Y + T - W_2)$ et des écarts-types moins élevés qu'avec KM.

s_i, $i = 1, 2, 3$ désigne l'écart-type des perturbations de l'équation i, a_k^* (resp. b_i^*) est l'estimation ponctuelle du coefficient a_k (resp. b_i) ; le nombre en dessous de cette estimation ponctuelle est son écart-type asymptotique, $R_z(Y_k)$ est la corrélation entre une variable Y_k et la projection de Y_k sur l'espace engendré par les variables X_2 dans le cas des doubles moindres carrés, sur l'espace engendré par les composantes principales dans le cas de la méthode de KLOEK et MENNES, sur l'espace engendré par les pseudo-composantes principales dans le cas de la nouvelle méthode.

Conclusion de la comparaison des méthodes

Pour résumer l'ensemble des résultats numériques, il apparaît que les écart-types sont plus élevés avec la méthode de KLOEK et MENNES qu'avec la nouvelle méthode (le coefficient de corrélation entre Y_k et Y_k^* est plus élevé pour la nouvelle méthode).

6 Conclusion

Quarante ans après, nous avons réexaminé le problème de KLOEK et MENNES avec les connaissances et les outils d'aujourd'hui. L'approche de KLOEK et MENNES est axée sur les problèmes de multicolinéarité entre les variables X ; l'approche développée dans cet article prend en compte non seulement ces problèmes de multicolinéarité, mais aussi, afin d'obtenir des variances asymptomatiques des estimateurs plus faibles, les corrélations entre les variables Y et les variables instrumentales Y^*. L'esprit de la méthode de KLOEK et MENNES est celui de la régression sur composantes principales, alors que la méthode développée ici est proche de la régression PLS.

• Références bibliographiques

ALMOY T. (1996). – « A Simulation Study on Comparison of Prediction Methods When Only a Few Components are Relevant », *Computational Statistics and Data Analysis*, 21, 1, pp. 87-107.

CASIN Ph. (1996). – « Une généralisation de l'analyse en composantes principales », *Revue de Statistique Appliquée*, 44, 3.

HELLAND I.S. (1990). – « Partial Least Square Regression and Statistical Models », *Scandinavian Journal of Statistics*, 17, pp. 97-114.

HELLAND I.S., ALMOY T. (1989). – « Comparison of Prediction Methods When Only a Few Components are Relevant », *Journal of the American Statistical Association*, 89, 1994, pp. 583-591.

JOHNSTON J. (1972). – *Econometric Methods*, McGraw-Hill, New-York.

JOLLIFFE I.T. (1986). – *Principal Components Analysis*, Springer, New-York.

KLEIN L.R. (1950). – *Economic Fluctuations in the United States, 1921-1941*, ed. John Wiley and Sons, New York.

KLEIN L.R. (1969). – « Estimation of Interdependant System in Macroeconometrics », *Econometrica*, 37, 1969, pp. 171-192.

KLOEK T., MENNES L.B.M. (1960). – « Simultaneous Equations Estimation based on Principal Components of Predetermined Variables », *Econometrica*, 28, 1, pp. 45-61.

MALINVAUD E. (1978). – *Méthodes statistiques de l'Économétrie*, Dunod, Paris.

PALM R., IEMMA A. F. (1995). – « Quelques alternatives à la régression classique dans le cas de la colinéarité », *Revue de Statistique Appliquée*, 43, 2, pp. 5-33.

TENENHAUS M. (1995). – « Nouvelles méthodes de régression PLS », *Les Cahiers de Recherche HEC*, 540.

TUCKER L.R. (1958). – « An Inter-Battery Method of Factor Analysis », *Psychometrika*, 23, 2, pp. 111-136.

Montrons que :

Si $\quad z^j = \sum_{h=1}^{N-L} F_h^j \, a_h^j = F^j a^j$

avec $\quad (a^j)' a^j = 1$

alors : $\quad \sum_{h=1}^{N-L} \text{cov}^2(z^j, F_h^j) \geqslant \text{Var}^2(z^j)$

***Preuve* :**

$$\sum_{h=1}^{N-L} \text{Cov}^2(z^j, F_h^j) = (z^j)' F^j (F^j)' z^j = (F^j a^j)' F^j (F^j)' F^j a^j$$

$$= (a^j)' (F^j)' F^j (F^j)' F^j a^j$$

$(F^j)' F = HDH'$ où H est la matrice $(N-L) \times (N-L)$ des vecteurs propres normés de $(F^j)' F^j$ et D la matrice diagonale $(N-L) \times (N-L)$ des valeurs propres de $(F^j)' F^j \cdot H' H = HH' = Id$, où Id désigne la matrice identité.

Donc $\quad \sum_{h=1}^{N-L} \text{Cov}^2(z^j, F_h^j) = (a^j)' HDH' \; HDH' a^j = (a^j)' H \; D^2 \; H' a^j$

Soit $\quad v^j = H' a^j$, alors $(v^j)' \, (v^j) = (a^j)' HH' a^j = 1$

et :

$$\sum_{h=1}^{N-L} \text{Cov}^2(z^j, F_h^j) = (a^j) \; HD^2 \; H' a^j = (v^j)' \; D^2 \; v^j$$

Soit $\quad v^j = (V_1, \ldots, v_{N-L})$ et soit d_h la h-ième valeur propre de $\underline{D}$:

alors :

$$\sum_{h=1}^{N-L} \text{Cov}^2(z^j, F_h^j) = \sum_{h=1}^{N-L} v_h^2 d_h^2$$

Comme : $\quad \sum_{h=1}^{N-L} v_h^2 d_h^2 \geqslant \left(\sum_{h=1}^{N-L} v_h^2 d_h \right)^2$

(La moyenne des d_h^2 est supérieure ou égale au carré de la moyenne des d_h)

Donc $$\sum_{h=1}^{N-L} \text{Cov}^2(z^j, F_h^j) \geqslant \left(\sum_{h=1}^{N-L} v_h^2 d_h \right)^2$$

par conséquent : $\sum_{h=1}^{N-L} \text{Cov}^2(z^j, F_h^j) \geqslant \text{Var}^2(z^j)$

car :

$$\begin{aligned} \text{Var}(z^j) = (z^j)'(z^j) &= (a^j)'(F^j)'(F^j)a^j \\ &= (a^j)' H D H' a^j \\ &= (v^j)' D (v^j) \end{aligned}$$

Remarque : si a^j est un vecteur propre de $(F^j)'(F^j)$:

$$\sum_{h=1}^{N-L} \text{Cov}^2(z^j, F_h^j) = \text{Var}^2(z^j)$$