JOURNAL DE LA SOCIÉTÉ STATISTIQUE DE PARIS

JEAN-JACQUES BOULANGER

Le problème des unités en analyse statistique multidimensionnelle appliquée à la macro-économie

Journal de la société statistique de Paris, tome 121, n° 2 (1980), p. 90-95

<http://www.numdam.org/item?id=JSFS_1980__121_2_90_0>

LE PROBLÈME DES UNITÉS EN ANALYSE STATISTIQUE MULTIDIMENSIONNELLE APPLIQUÉE A LA MACRO-ÉCONOMIE

Jean-Jacques BOULANGER

Dans l'étude qui suit, l'auteur analyse le problème posé par le choix des unités en analyse statistique multidimensionnelle appliquée à la macro-économie et suggère une solution permettant d'en amoindrir les effets sur la détermination des vecteurs propres et des composantes principales.

In the following paper, the author analyzes the problem raised by the choice of the units in multivariate statistical analysis applied to macroeconomics, and suggests a solution enabling to lessen the effects on the determination of the latent vectors and of the principal components.

In der Studie, die folgt, analysiert der Verfasser das Problem, das die multidimensionale statistische Analyse für ein gegebenes Problem stellt und zwar für die Mikroökonomie. Er schlägt eine Lösung vor, die die Wirkung auf die charakteristischen Vektoren vermindert und ebenso auf die hauptäschtlichsten Componenten.

Le choix des unités lors de l'analyse en composantes principales intrigue les statisticiens. Faut-il, par exemple, utiliser les mesures propres à un échantillon, ou bien remplacer les variables x_i par leurs valeurs réduites $\frac{x_i}{\sigma_i}$, c'est-à-dire effectuer un changement d'unité. Or on obtient des valeurs propres et des vecteurs propres dans les deux cas très différents, et il en sera de même pour tout autre changement d'unité.

Nous avons essayé d'analyser le problème en cherchant la première composante principale de la matrice des (variances et) covariances relatives aux 3 variables α_1, α_2 et α_3 définies de la façon suivante :

α_1 = taux d'accroissement annuel en volume du produit national brut (PNB)
α_2 = taux d'accroissement annuel en volume de la formation brute de capital (FBC)
α_3 = taux d'accroissement annuel en volume des importations et revenus versés au reste du monde (Imp + etc.).

Le choix s'est porté sur ces 3 variables parmi les agrégats normalisés de l'ancienne nomenclature SEEF, parce que ce sont celles qui présentent entre elles les corrélations les plus importantes. On est ainsi parti, pour la période 1960-71, de la matrice des covariances suivantes :

$$A = \begin{bmatrix} 0{,}911 & 3{,}706 & 3{,}245 \\ 3{,}706 & 21{,}083 & 16{,}880 \\ 3{,}245 & 16{,}880 & 22{,}469 \end{bmatrix}$$

et l'on obtient comme première valeur propre λ_1 et premier vecteur propre u_1 les résultats suivants :

$$\lambda_1 = 39{,}30 \qquad u_1 = \begin{bmatrix} 0{,}1268 \\ 0{,}6880 \\ 0{,}7146 \end{bmatrix}$$

On en déduit la première composante principale :

$$y_1 = 0{,}1268\,(\alpha_1 - \bar{\alpha}_1) + 0{,}6880\,(\alpha_2 - \bar{\alpha}_2) + 0{,}7146\,(\alpha_3 - \bar{\alpha}_3)$$

qui peut être considérée comme un facteur général de croissance de l'économie.

Mais au lieu de considérer les variables centrées $x_i = \alpha_i - \bar{\alpha}_i$ définies plus haut, on peut s'intéresser aux variables centrées $x_i = k_i\,(\alpha_i - \bar{\alpha}_i)$, k_i étant le rapport au PNB du même PNB, de la FBC et des (Imp+ etc.), ceux-ci étant mesurés en milliards de francs par exemple, ce qui conduit à estimer les k_i par $k_1 = 1$, $k_2 = 0{,}27$, $k_3 = 0{,}15$.

Dans le premier cas on cherche à déterminer un facteur général de croissance à partir des taux de croissance α_1, α_2, α_3 des 3 agrégats définis plus haut, donc à partir d'accroissements relatifs. Dans le second cas on essaie de déterminer un facteur général de croissance à partir des accroissements absolus (à un facteur près) du PNB, de la FBC et des (Imp+ etc). Ces deux choix sont tous deux extrêmement défendables. Cependant le deuxième choix consiste à majorer l'importance du PNB par rapport aux 2 autres agrégats, le premier choix consiste à la diminuer.

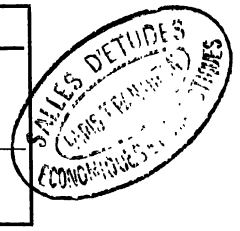
Le deuxième choix conduit à la matrice des covariances suivantes :

$$A = \begin{bmatrix} 0{,}9111 & 1{,}0006 & 0{,}4868 \\ 1{,}0006 & 1{,}5369 & 0{,}6836 \\ 0{,}4868 & 0{,}6836 & 0{,}5056 \end{bmatrix}$$

et l'on obtient comme première valeur propre λ_1 et premier vecteur propre u_1 les résultats suivants :

$$\lambda_1 = 2{,}607 \qquad u_1 = \begin{bmatrix} 0{,}5485 \\ 0{,}7494 \\ 0{,}3708 \end{bmatrix}$$

Tableau 1	Cas 1	Cas 2	Cas 3	Cas 4
$u_1 = \begin{bmatrix} u_{11} \\ u_{12} \\ u_{13} \end{bmatrix}$	0,127 0,688 0,715	0,280 0,783 0,556	0,548 0,749 0,371	0,579 0,592 0,560
$\frac{\lambda_1}{\text{Trace}}$	0,884	0,881	0,883	0,853



On voit que les composantes du premier vecteur propre sont dans les deux cas nettement différentes et conduisent à deux axes factoriels sans rapport entre eux. Plus généralement on a analysé les 4 cas suivants caractérisés par les choix :

Cas 1	$x_i = \alpha_i - \bar{\alpha}_i$
Cas 2	$x_i = \sqrt{k_i}\,(\alpha_i - \bar{\alpha}_i)$
Cas 3	$x_i = k_i\,(\alpha_i - \bar{\alpha}_i)$
Cas 4	$x_i = \dfrac{\alpha_i - \bar{\alpha}_i}{\sigma_i}$ (tableau 1)

La première composante principale s'exprime donc dans les 4 cas par :

$$1^{er}\ \text{cas} : \quad y_1 = u_{11}(\alpha_1 - \bar{\alpha}_1) + u_{12}(\alpha_2 - \bar{\alpha}_2) + u_{13}(\alpha_3 - \bar{\alpha}_3)$$

$$2^{e}\ \text{cas} : \quad y_1 = u_{11}(\alpha_1 - \bar{\alpha}_1) + u_{12}\sqrt{k_2}(\alpha_2 - \bar{\alpha}_2) + u_{13}\sqrt{k_3}(\alpha_3 - \bar{\alpha}_3)$$

$$3^{e}\ \text{cas} : \quad y_1 = u_{11}(\alpha_1 - \bar{\alpha}_1) + u_{12}k_2(\alpha_2 - \bar{\alpha}_2) + u_{13}k_3(\alpha_3 - \bar{\alpha}_3)$$

$$4^{e}\ \text{cas} : \quad y_1 = u_{11}\frac{\alpha_1 - \bar{\alpha}_1}{\sigma_1} + u_{12}\frac{\alpha_2 - \bar{\alpha}_2}{\sigma_2} + u_{13}\frac{\alpha_3 - \bar{\alpha}_3}{\sigma_3}$$

Pour rendre les résultats plus comparables on a entrepris de présenter dans les 4 cas y_1 sous la forme :

$y_1 = l_1(\alpha_1 - \bar{\alpha}_1) + l_2(\alpha_2 - \bar{\alpha}_2) + l_3(\alpha_3 - \bar{\alpha}_3)$, l_1, l_2, l_3 étant les composantes d'un vecteur unitaire, ce qui conduit à multiplier dans chacun des cas les 3 u_{11}, u_{12} et u_{13} par un même nombre. On obtient ainsi pour les 4 cas les résultats suivants :

$$(1)\quad \left|\begin{array}{l} \text{Cas 1} : y_1 = 0{,}127(\alpha_1 - \bar{\alpha}_1) + 0{,}688(\alpha_2 - \bar{\alpha}_2) + 0{,}715(\alpha_3 - \bar{\alpha}_3) \\ \text{Cas 2} : y_1 = 0{,}520(\alpha_1 - \bar{\alpha}_1) + 0{,}755(\alpha_2 - \bar{\alpha}_2) + 0{,}400(\alpha_3 - \bar{\alpha}_3) \\ \text{Cas 3} : y_1 = 0{,}934(\alpha_1 - \bar{\alpha}_1) + 0{,}344(\alpha_2 - \bar{\alpha}_2) + 0{,}095(\alpha_3 - \bar{\alpha}_3) \\ \text{Cas 4} : y_1 = 0{,}961(\alpha_1 - \bar{\alpha}_1) + 0{,}204(\alpha_2 - \bar{\alpha}_2) + 0{,}187(\alpha_3 - \bar{\alpha}_3) \end{array}\right.$$

On voit que l'on obtient comme expressions de la première composante principale des résultats encore plus différents que pour les u_{11}, u_{12}, u_{13}, ce qui conduit à douter de l'intérêt de la méthode. Par contre, si l'on rapporte λ_1 à la trace, on obtient des valeurs extrêmement voisines, qui confirment dans les 4 cas l'importance de la première composante comme facteur général de croissance. Mais ceci ne suffit pas pour exprimer les variables x_1, x_2, x_3 en fonction de facteurs généraux et spécifiques, tels que le suggère l'analyse factorielle.

Si nous revenons à l'expression de y_1 en fonction des variables centrées $\alpha_1 - \bar{\alpha}_1$, $\alpha_2 - \bar{\alpha}_2$, $\alpha_3 - \bar{\alpha}_3$ nous avons vu que le premier cas défavorise la variable PNB au profit des 2 autres; au contraire les 3^e et 4^e cas favorisent la variable PNB et le bon sens conduit à penser que la meilleure estimation de y_1 doit être aussi éloignée des cas 3 et 4 que du cas 1, donc peu éloignée de celle du cas 2. Mais quelle expression préférer? En réalité il est impossible d'effectuer un choix basé sur la mathématique.

Considérons une loi de Gauss à p variables $x_1, x_2, \ldots, x_p$. Sa densité de probabilité est égale $Ke^{-\frac{Q}{2}}$ avec :

$$Q = \text{forme quadratique d'ordre } p = x'Ax,\ K = (2\pi)^{-\frac{n}{2}} \mid A \mid^{\frac{1}{2}},$$

$$A = \begin{bmatrix} a_{11} \ldots\ldots a_{1p} \\ \ldots\ldots\ldots\ldots \\ a_{p1} \ldots\ldots a_{pp} \end{bmatrix},\quad x = \begin{bmatrix} x_1 \\ \vdots \\ x_p \end{bmatrix},\ x' = \text{vecteur ligne transposé de } x \text{ et } \mid A \mid = \text{déter-}$$

minant formé des éléments de la matrice A.

On a d'autre part $A = V^{-1}$ (V étant la matrice des covariances).

L'équation $Q = 1$ représente un hyperellipsoïde dont les axes principaux ont comme directions les p vecteurs propres de la matrice A, alors que les p demi-axes principaux sont égaux aux inverses des racines carrées des p valeurs propres $\lambda_1, \lambda_2, \ldots, \lambda_p$ de la matrice A. Les valeurs propres et vecteurs propres de la matrice A permettent donc d'estimer les 1/2 axes principaux de l'hyperellipsoïde $Q = 1$. Mais ceci suppose que les variables $x_1, x_2, \ldots, x_p$ suivent une loi de Gauss à p variables. En effet pour obtenir à partir d'une matrice symétrique d'ordre p, p valeurs propres et p vecteurs propres, il suffit que la matrice soit définie positive.

D'autre part il existe toujours une loi de Gauss à p variables ayant pour matrice des covariances la matrice des covariances expérimentalement observées et ce sont les 1/2 axes principaux de l'hyperellipsoïde correspondant ($Q = x'V^{-1}x = 1$)* que les valeurs propres et vecteurs propres permettent d'estimer. Or on peut imaginer un très grand nombre de fonctions ayant pour matrice des covariances la matrice des covariances observées. Dans tous les cas les vecteurs propres sont les mêmes. Et si la loi des $x_1, x_2, \ldots, x_p$ est trop éloignée d'une loi de Gauss à p variables, les estimations n'ont plus qu'une valeur réduite.

Considérons maintenant une loi de Gauss à 3 variables x_1, x_2, x_3. A cette loi de Gauss correspond un ellipsoïde dont les directions principales sont les vecteurs propres de la matrice $A = V^{-1}$. On peut imaginer cet ellipsoïde tracé dans l'espace ainsi que ses directions principales. Supposons que sans toucher à x_1 et x_2 on choisisse comme unité de comptage sur l'axe Ox_3 une unité 10 fois plus petite. Tout se passera comme si l'on avait effectué une affinité parallèlement à Ox_3 de rapport 1/10. On obtiendra un ellipsoïde très aplati selon le plan $x_1 Ox_2$, et les axes de symétrie de l'ellipsoïde seront très différents de ce qu'ils étaient avant l'affinité. Or garder la variable x_3 et diviser l'unité de mesure sur Ox_3 par 10, ou garder l'unité de mesure sur Ox_3 et diviser x_3 par 10 conduit au même résultat : Les axes de symétrie de l'ellipsoïde sont grandement transformés et il en est de même des vecteurs propres de la matrice A.

Reprenons maintenant notre recherche de la meilleure première composante principale. Nous avons vu que, du point de vue de la mathématique, il est impossible de fixer un choix. On peut cependant s'appuyer sur certaines remarques. Par exemple on pourra penser que plus une variable a une corrélation multiple importante avec les 2 autres, plus sa contribution sera importante dans la détermination de la première composante principale. Le coefficient de corrélation multiple entre α_1 d'une part, α_2 et α_3 d'autre part est égal à $R_{1.23} = 0{,}851$. De même on a : $R_{2.13} = 0{,}880$ et $R_{3.12} = 0{,}784$. On est donc conduit à penser que c'est d'abord α_2 qui contribue le plus à la formation de la première composante principale y_1, puis α_1 et enfin seulement α_3, ce qui est conforme à ce qui est communément admis.

Posons

$$x_1 = \alpha_1 - \bar{\alpha}_1, \quad x_2 = (k_2)^z (\alpha_2 - \bar{\alpha}_2), \quad x_3 = (k_3)^z (\alpha_3 - \bar{\alpha}_3).$$

Nous allons maintenant analyser l'évolution du vecteur unitaire l (l_1, l_2, l_3) en fonction de z. Laissons de côté le cas 4 et considérons le cas 5 défini par $z = -1/2$, qui conduit à :

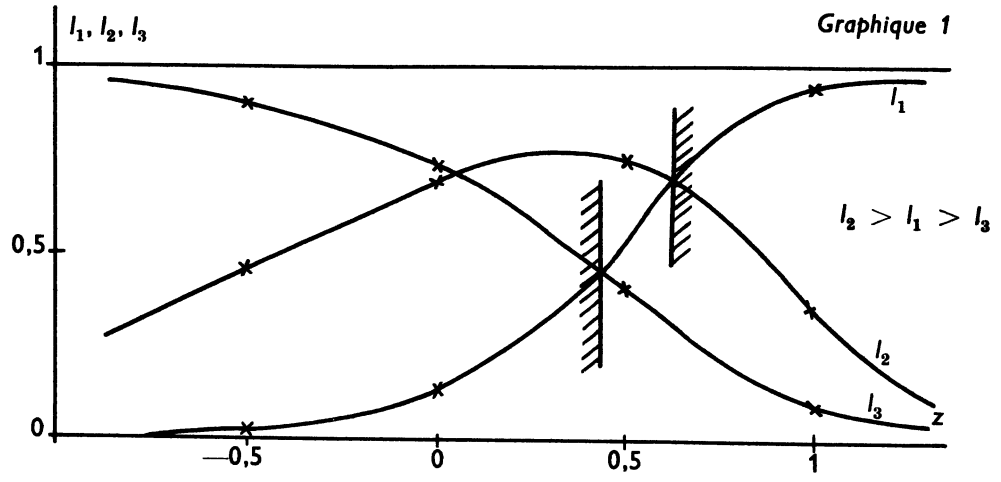
Cas 5

$$y_1 = 0{,}022\ (\alpha_1 - \bar{\alpha}_1) + 0{,}442\ (\alpha_2 - \bar{\alpha}_2) + 0{,}897\ (\alpha_3 - \bar{\alpha}_3)$$

Nous avons porté sur le graphique 1 les valeurs de l_1, l_2, l_3 en fonction des valeurs de z : —0,5, 0, 0,5, 1 et nous avons ajusté par les courbes l_1, l_2, l_3 en fonction de z. Nous voyons que si nous imposons comme conditions $l_2 > l_1 > l_3$ il reste pour le choix de z un domaine assez étroit comprenant $z = 0{,}5$, ce qui conduit à choisir z voisin de 0,5 et plus simplement $z = 0{,}5$.

Mais on peut apporter une autre limitation aux variations du vecteur l. En effet si k_2 et k_3 étaient égaux à k_1 (=1), $(k_2)^z$ et $(k_3)^z$ seraient constants quel que soit z et le vecteur l demeurerait fixe. On est donc amené à choisir des agrégats, des branches, des sous-branches, etc., qui ont à peu près la même valeur monétaire, ce qui n'est pas toujours possible, car il n'est pas toujours possible de scinder une branche (ou de regrouper deux sous-branches) en des parties économiquement valables. Mais cette réserve faite, on doit chercher à déterminer des unités économiques ayant la valeur monétaire la plus voisine possible. La combi-

* On peut aussi considérer l'hyperellipsoïde inverse correspondant à $Q' = x'\,V\,x = 1$, qui a mêmes axes principaux et des valeurs de demi-axes inverses des précédents.



naison de ce choix et de celui de $z = 0{,}5$ doit conduire à des composantes du vecteur unitaire l acceptables.

Signalons sans tarder que la méthode qui vient d'être exposée n'est pas d'un usage général. Elle suppose que tous les agrégats étidiés ont une valeur qui s'exprime en monnaie et devient inadéquate, si l'on considère en plus des effectifs de salariés, des heures-ouvrées, des productivités... Autre exemple : en biométrie humaine on peut décomposer une taille et une envergure en des composantes dont les mesures sont aussi voisines que possible, mais il est plus difficile de décomposer un périmètre thoracique, bien qu'il se détermine avec la même unité. Et nous ne parlons pas du poids dont l'unité (kg ou livre) n'a aucune commune mesure avec le centimètre. Dans ces cas le problème des unités reste posé et sans solution, bien qu'il concerne l'ensemble des méthodes d'analyse statistique multidimensionnelle autres que l'analyse des corrélations canoniques (et son cas particulier la corrélation multiple et la régression linéaire multiple) d'une part, l'analyse des correspondances d'autre part, puisque celle-ci n'utilise que des pourcentages.

Cependant, si l'on a obtenu des composantes principales satisfaisantes, alors l'analyse en composantes principales débouche tout naturellement sur une analyse factorielle généralisée (généralisée parce qu'on commence par supposer que les p facteurs sont tous des facteurs généraux).

Soient x le vecteur des p variables centrées, y le vecteur des p composantes principales et U la matrice des vecteurs propres, on a :

$x = Uy$ d'où :

$$\left|\begin{array}{l} x_1 = u_{11}y_1 + u_{21}\,y_2 + \ldots\ldots + u_{p1}y_p \\ x_2 = u_{12}y_1 + u_{22}y_2 + \ldots\ldots + u_{p2}y_p \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ x_p = u_{1p}y_1 + u_{2p}y_2 + \ldots\ldots + u_{pp}y_p \end{array}\right.$$

Posons $y_j = \sqrt{\lambda_j}g_j$ et $w_{ij} = \sqrt{\lambda_j}u_{ji}$,

on a :

$$(2) \qquad \left|\begin{array}{l} x_1 = w_{11}g_1 + w_{12}g_2 + \ldots\ldots + w_{1p}g_p \\ x_2 = w_{21}g_1 + w_{22}g_2 + \ldots\ldots + w_{2p}g_p \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ x_p = w_{p1}g_1 + w_{p2}g_2 + \ldots\ldots + w_{pp}g_p \end{array}\right.$$

avec $E[g_i] = 0$, $E[g_i^2] = 1$, $E[g_i g_j] = 0^*$, i et j variant de l à p.

D'où :

$$\left|\begin{array}{l} E[x_1^2] = w_{11}^2 + w_{12}^2 + \ldots\ldots + w_{1p}^2 \\ E[x_2^2] = w_{21}^2 + w_{22}^2 + \ldots\ldots + w_{2p}^2 \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ E[x_p^2] = w_{p1}^2 + w_{p2}^2 + \ldots\ldots + w_{pp}^2 \end{array}\right.$$

On passe de l'expression (2) à l'expression de l'analyse factorielle classique en éliminant tous les w_{ij} dont le carré est inférieur à $kE[x_i^2]$, k étant fixé a priori (égal à 0,05 ou 0,01 ou 0,001 par exemple); ce qui permet de dégager les facteurs généraux, les facteurs de groupe et les facteurs spécifiques. En effet g_j sera un facteur général si tous les w_{ij} ($i = 1, 2, \ldots, p$) sont différents de 0, un facteur spécifique si tous les w_{ij} sont nuls sauf un, un facteur de groupe dans les autres cas.

On peut déterminer les w_{ij} en utilisant la déflation classique :

$$\overset{1}{A} = A - \lambda_1 u_1 u_1'$$

A étant la matrice de départ, λ_1 la première valeur propre, u_1 le premier vecteur propre, u_1' son transposé.

De même on a : $\overset{2}{A} = \overset{1}{A} - \lambda_2 u_2 u_2'$

$\ldots\ldots\ldots\ldots$

jusqu'à $\overset{p}{A} = \overset{p-1}{A} - \lambda_p u_p u_p' = 0.$

D'où : $A = A_1 u_1 u_1' + \lambda_2 u_2 u_2' + \ldots\ldots + \lambda_p u_p u_p'$

avec $u_1 = \begin{bmatrix} u_{11} \\ u_{12} \\ \cdot \\ \cdot \\ \cdot \\ u_{1p} \end{bmatrix}$, $u_2 = \begin{bmatrix} u_{21} \\ u_{22} \\ \cdot \\ \cdot \\ \cdot \\ u_{2p} \end{bmatrix}$, $\ldots\ldots$, $u_p = \begin{bmatrix} u_{p1} \\ u_{p2} \\ \cdot \\ \cdot \\ \cdot \\ u_{pp} \end{bmatrix}$,

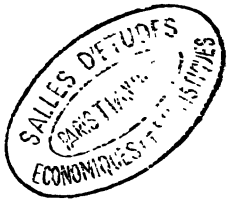
les u' étant les transposés des u; d'où les équations (2) en posant encore $w_{ij} = \sqrt{\lambda_j} u_{ji}$. Mais si l'on a pris soin de conserver les diverses matrices $\overset{1}{A}$, $\overset{2}{A}$, $\ldots\ldots$, il est encore plus simple, partant de $\overset{1}{A} = A - \lambda_1 u_1 u_1'$, d'écrire : $\overset{1}{E}[x_i^2] = E[x_i^2] - \lambda_1 u_{1i}^2$

d'où : $w_{i1}^2 = E[x_i^2] - \overset{1}{E}[x_i^2]$

puis $w_{i2}^2 = \overset{1}{E}[x_i^2] - \overset{2}{E}[x_i^2]$

$\ldots\ldots\ldots\ldots\ldots\ldots$

$\overset{1}{E}[x_i x_j]$ étant le terme général de $\overset{1}{A}$, $\overset{2}{E}[x_i x_j]$ celui de $\overset{2}{A}$, etc.



En résumé une analyse factorielle sans hypothèses préalables découle naturellement d'une analyse en composantes principales, si l'on a déterminé des composantes principales suffisamment valables. Car ce n'est que dans certains cas, souvent un peu particuliers, grâce aussi à un choix réfléchi des variables utilisées, qu'on obtient des composantes principales auxquelles on peut raisonnablement accorder sa confiance.

* On a : $x = Uy$ d'où $x' = y'U'$. La forme quadratique $Q = x'Ax$ devient $Q = y'U'AUy = y'D_\lambda y$, D_λ étant la matrice diagonale des valeurs propres λ_i. La matrice des covariances des y_i est donc égale à D_λ^{-1}. C'est une matrice diagonale, puisque D_λ est diagonale. On en déduit : $E[y_i y_j] = 0$ d'où : $E[g_i g_j] = 0$.