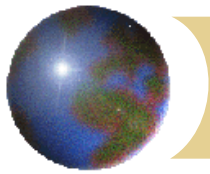


CHAPITRE 6

Analyse Factorielle des Correspondances (AFC)



L'analyse (factorielle) des correspondances

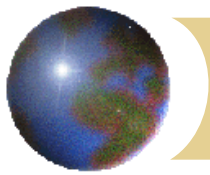
But On cherche à décrire la liaison entre deux variables qualitatives.

Exemple on peut regarder la ventilation des séjours de vacance (location, résidence secondaire, camping,...) en fonction de la catégorie socio-professionnelle.

Différence avec l'ACP l'ACP se fait dans un cadre différent ; les variables sont quantitatives et donc

- il est possible de faire des opérations mathématiques sur les valeurs des variables ;
- par contre, il n'est en général pas possible de compter les individus qui ont une caractéristique donnée (taille=1,83m)

Pourquoi deux variables ? le cas de plus de deux variables est l'analyse de correspondance multiples, traité plus tard dans le cours.



Variables qualitatives

Soit $\mathcal{X}$ une variable qualitative. On dispose d'un échantillon de n individus sur lesquels la variable est mesurée.

Modalités (ou catégories) les valeurs que peut prendre une variable qualitative; si la variable a r modalités (valeurs possibles), on note x_i , $1 \leq i \leq r$, ces modalités.

Effectif le nombre d'occurrence de la modalité x_i dans l'échantillon; on le note n_i , et on a $\sum_{i=1}^r n_i = n$.

Fréquence c'est la grandeur $f_i = n_i/n$; la somme des fréquences sur les modalités est 1. On utilise souvent le pourcentage $100f_i$.

Représentation on peut utiliser un tableau avec r lignes de la forme

$\vdots$	$\vdots$	$\vdots$
x_i	n_i	f_i
$\vdots$	$\vdots$	$\vdots$

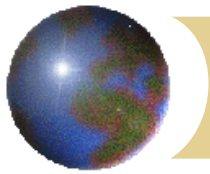


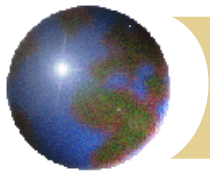
Tableau de contingence

Soient $\mathcal{X}$ et $\mathcal{Y}$ deux variables qualitatives à r et s modalités respectivement décrivant un ensemble de n individus.

Définition le tableau de contingence est une matrice à r lignes et s colonnes renfermant les effectifs n_{ij} d'individus tels que $\mathcal{X} = x_i$ et $\mathcal{Y} = y_j$.

$$\mathbf{N} = \begin{pmatrix} n_{11} & n_{12} & \cdots & n_{1s} \\ n_{21} & n_{22} & & \\ & & \ddots & \\ \vdots & & n_{ij} & \vdots \\ & & & \ddots & \\ n_{r1} & & & & n_{rs} \end{pmatrix}$$

La constitution de ce tableau est ce que les praticiens des enquêtes appellent un « tri croisé ».



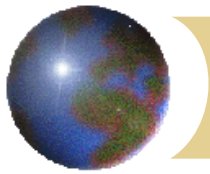
Marges et profils

Marge en ligne c'est la somme $n_{i.} = \sum_{j=1}^s n_{ij}$, c'est-à-dire l'effectif total de la modalité x_i de $\mathcal{X}$.

Marge en colonnes c'est la somme $n_{.j} = \sum_{i=1}^r n_{ij}$, c'est-à-dire l'effectif total de la modalité y_j de $\mathcal{Y}$.

Deux lectures possibles selon la variable que l'on privilégie, on peut définir

- le tableau des *profils-lignes* $n_{ij}/n_{i.}$, qui représente la fréquence de la modalité y_j conditionnellement à $\mathcal{X} = x_i$; la somme de chaque ligne est ramenée à 100%.
- le tableau des *profils-colonnes* $n_{ij}/n_{.j}$, qui représente la fréquence de la modalité x_i conditionnellement à $\mathcal{Y} = y_j$; la somme de chaque colonne est ramenée à 100%.



Propriétés des profils

Moyenne la moyenne des profils-lignes (avec poids correspondant aux effectifs marginaux des lignes) est le profil marginal des colonnes :

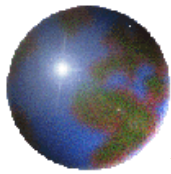
$$\sum_{i=1}^r \frac{n_{i.}}{n} \times \frac{n_{ij}}{n_{i.}} = \frac{n_{.j}}{n},$$

et de même pour les colonnes $\sum_{j=1}^s \frac{n_{.j}}{n} \times \frac{n_{ij}}{n_{.j}} = \frac{n_{i.}}{n}.$

Indépendance empirique lorsque tous les profils lignes sont identiques, il y a indépendance entre $\mathcal{X}$ et $\mathcal{Y}$, puisque la connaissance de $\mathcal{X}$ ne change pas la répartition de $\mathcal{Y}$. On a pour tout j

$$\frac{n_{1j}}{n_{1.}} = \frac{n_{2j}}{n_{2.}} = \dots = \frac{n_{rj}}{n_{r.}} = \frac{n_{1j} + \dots + n_{rj}}{n_{1.} + \dots + n_{r.}} = \frac{n_{.j}}{n}$$

et donc $n_{ij} = \frac{n_{i.} \cdot n_{.j}}{n}.$



Le χ^2 d'écart à l'indépendance

Définition c'est la grandeur suivante, aussi notée χ^2 ou X^2

$$d^2 = \sum_{i=1}^r \sum_{j=1}^s \frac{\left(n_{ij} - \frac{n_{i.}n_{.j}}{n}\right)^2}{\frac{n_{i.}n_{.j}}{n}} = n \left[\sum_{i=1}^r \sum_{j=1}^s \frac{n_{ij}^2}{n_{i.}n_{.j}} - 1 \right].$$

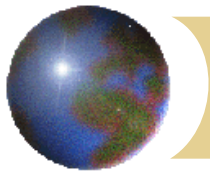
Si les variables sont indépendantes, $d^2 = 0$.

Borne supérieure comme $n_{ij} \leq n_{i.}$, on a

$$\sum_{i=1}^r \sum_{j=1}^s \frac{n_{ij}^2}{n_{i.}n_{.j}} \leq \sum_{i=1}^r \sum_{j=1}^s \frac{n_{ij}}{n_{.j}} = \sum_{j=1}^s \frac{\sum_{i=1}^r n_{ij}}{n_{.j}} = \sum_{j=1}^s \frac{n_{.j}}{n_{.j}} = s,$$

et donc $d^2 \leq n(s-1)$. On fait de même pour r et

$$\varphi^2 = \frac{d^2}{n} \leq \min(s-1, r-1).$$



Le χ^2 d'écart à l'indépendance (suite)

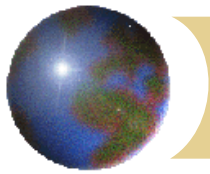
Dépendance fonctionnelle si $\varphi^2 = s - 1$, alors pour chaque i soit $n_{ij} = n_{i.}$, soit $n_{ij} = 0$: il existe une unique case non nulle par ligne. $\mathcal{Y}$ est donc fonctionnellement liée à $\mathcal{X}$.

Dépendance inverse cette relation ne signifie pas que $\mathcal{X}$ est fonctionnellement liée à $\mathcal{Y}$, sauf si $r = s$. On peut alors représenter le tableau comme une matrice diagonale.

Contribution au χ^2 c'est le terme

$$\frac{\left(n_{ij} - \frac{n_{i.}n_{.j}}{n}\right)^2}{\frac{n_{i.}n_{.j}}{n}}$$

qui permet de mettre en évidence les associations significatives entre catégories de deux variables.



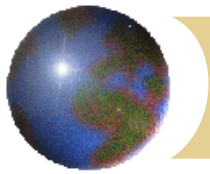
Caractère significatif du χ^2

Problème à partir de quelle valeur de d^2 doit-on considérer que les variables $\mathcal{X}$ et $\mathcal{Y}$ sont indépendantes ?

Méthode on suppose que $\mathcal{X}$ et $\mathcal{Y}$ sont issus de tirages de deux variables aléatoires indépendantes. On peut alors montrer que d^2 est une réalisation d'une variable aléatoire D^2 qui suit une loi $\chi^2_{(r-1)(s-1)}$.

Définition Loi du khi-deux à p degrés de libertés χ_p^2 est la loi de la variable $\sum_{i=1}^p U_i^2$, où les U_i sont des variables gaussiennes réduites indépendantes.

Le test du χ^2 on se fixe un risque d'erreur α (0.01 ou 0.05 en général) et on calcule la valeur d_c^2 telle que $P\left(\chi^2_{(r-1)(s-1)} > d_c^2\right) = \alpha$. Si $d^2 > d_c^2$ on considère que l'événement est trop improbable et que donc que l'hypothèse originale d'indépendance doit être rejetée. On trouvera en général ces valeurs dans une table précalculée.

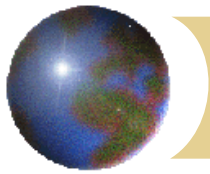


Analyse des correspondances de deux variables : les données

Effectifs on a un tableau de contingence $\mathbf{N}$ à m_1 lignes et m_2 colonnes résultant du croisement de deux variables qualitatives $\mathcal{X}_1$ et $\mathcal{X}_2$ à m_1 et m_2 modalités respectivement. On note $\mathbf{D}_1$ et $\mathbf{D}_2$ les matrices diagonales des effectifs marginaux

$$\mathbf{D}_1 = \begin{pmatrix} n_{1.} & & & 0 \\ & n_{2.} & & \\ & & \ddots & \\ 0 & & & n_{m_1.} \end{pmatrix} \quad \mathbf{D}_2 = \begin{pmatrix} n_{.1} & & & 0 \\ & n_{.2} & & \\ & & \ddots & \\ 0 & & & n_{.m_2} \end{pmatrix}$$

Profils le tableau des profils des lignes $n_{ij}/n_{i.}$ est donné par $\mathbf{D}_1^{-1}\mathbf{N}$ et celui des profils des colonnes $n_{ij}/n_{.j}$ par $\mathbf{N}\mathbf{D}_2^{-1}$.



Représentation géométrique des profils

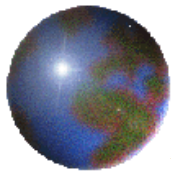
Nuage de points les profils-lignes forment un nuage de m_1 points de $\mathbb{R}^{m_2}$. Chaque point est affecté d'un poids égal à sa fréquence marginale $n_{i.}/n$, et la matrice des poids est donc $\frac{1}{n}\mathbf{D}_1$.

Centre de gravité c'est le profil marginal car

$$\mathbf{g}_\ell = \frac{1}{n}(\mathbf{D}_1^{-1}\mathbf{N})'\mathbf{D}_1\mathbf{1} = \left(\frac{n_{.1}}{n}, \dots, \frac{n_{.m_2}}{n}\right)'$$

Profils-colonnes les lignes du tableau $\mathbf{D}_2^{-1}\mathbf{N}'$ forment un nuage de m_2 points de $\mathbb{R}^{m_1}$, avec matrice de poids $\frac{1}{n}\mathbf{D}_2$ et centre de gravité

$$\mathbf{g}_c = \left(\frac{n_{1.}}{n}, \dots, \frac{n_{m_1.}}{n}\right)'$$



Comment étudier ces données

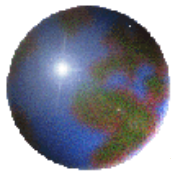
Cas indépendant en cas d'indépendance empirique, on aura

$$\frac{n_{ij}}{n_{i.}} = \frac{n_{.j}}{n} \text{ et } \frac{n_{ij}}{n_{.j}} = \frac{n_{i.}}{n}$$

et les deux nuages sont alors réduits à leurs centres de gravité respectifs.

Dimension des nuages comme les profils somment à 1, les m_1 profils-lignes sont situés dans le sous-espace W_1 de dimension $m_2 - 1$ défini par $\sum_{j=1}^{m_2} x_j = 1$ et $x_j \geq 0$.

ACP l'étude de la forme des nuages au moyen de l'Analyse en Composantes Principales permettra de rendre compte de la structure des écarts à l'indépendance.



La métrique du χ^2

Profils-lignes la distance entre deux profils-lignes i et i' est

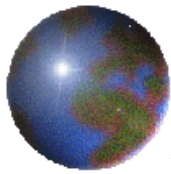
$$d_{\chi^2}^2(i, i') = \sum_{j=1}^{m_2} \frac{n}{n_{\cdot j}} \left(\frac{n_{ij}}{n_{i\cdot}} - \frac{n_{i'j}}{n_{i'\cdot}} \right)^2,$$

ce qui revient à utiliser la métrique diagonale $n\mathbf{D}_2^{-1}$.

Inertie l'inertie totale du nuage des profils-lignes par rapport à $\mathbf{g}_\ell$ est

$$\begin{aligned} I_{\mathbf{g}_\ell} &= \sum_{i=1}^{m_1} \frac{n_{i\cdot}}{n} d_{\chi^2}^2(i, \mathbf{g}_\ell) = \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \frac{n_{i\cdot}}{n_{\cdot j}} \left(\frac{n_{ij}}{n_{i\cdot}} - \frac{n_{\cdot j}}{n} \right)^2 \\ &= \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \frac{1}{n_{i\cdot} n_{\cdot j}} \left(n_{ij} - \frac{n_{i\cdot} n_{\cdot j}}{n} \right)^2 = \varphi^2 \end{aligned}$$

Cette inertie mesure donc l'écart à l'indépendance.



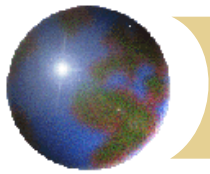
Pourquoi la métrique du χ^2

Pondération la pondération $n/n_{.j}$ permet de donner des importances comparables aux différentes « variables ».

Équivalence distributionnelle si deux colonnes j et j' de $\mathbf{N}$ ont le même profil, il est logique de les regrouper en une seule d'effectif $n_{ij} + n_{ij'}$; on a alors quand $n_{ij}/n_{.j} = n_{ij'}/n_{.j'}$

$$\begin{aligned} \frac{n}{n_{.j}} \left(\frac{n_{ij}}{n_{i.}} - \frac{n_{i'j}}{n_{i'.}} \right)^2 + \frac{n}{n_{.j'}} \left(\frac{n_{ij'}}{n_{i.}} - \frac{n_{i'j'}}{n_{i'.}} \right)^2 \\ = \frac{n}{n_{.j} + n_{.j'}} \left(\frac{n_{ij} + n_{ij'}}{n_{i.}} - \frac{n_{i'j} + n_{i'j'}}{n_{i'.}} \right)^2 \end{aligned}$$

La distance entre les profils-ligne est donc inchangée.



Autres propriétés de la métrique du χ^2

Propriétés de $\mathbf{g}_\ell$ le vecteur $\mathbf{Og}_\ell$ est orthogonal à W_1 au sens de la métrique du χ^2 car

$$\langle \mathbf{g}_\ell \mathbf{x}, \mathbf{Og}_\ell \rangle_{\chi^2} = (\mathbf{x} - \mathbf{g}_\ell)' n \mathbf{D}_2^{-1} \mathbf{g}_\ell = (\mathbf{x} - \mathbf{g}_\ell)' \mathbf{1} = 0.$$

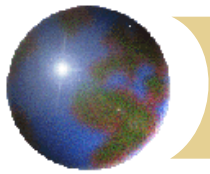
et la norme de $\mathbf{g}_\ell$ est $\|\mathbf{g}_\ell\|_{\chi^2}^2 = \mathbf{g}_\ell' n \mathbf{D}_2^{-1} \mathbf{g}_\ell = \mathbf{g}_\ell' \mathbf{1} = 1.$

Tous les vecteurs centrés du nuages sont donc orthogonaux à $\mathbf{g}_\ell$

Profils-colonnes on définit la distance entre deux profils-colonnes j et j' comme

$$d_{\chi^2}^2(j, j') = \sum_{i=1}^{m_1} \frac{n}{n_{i.}} \left(\frac{n_{ij}}{n_{.j}} - \frac{n_{ij'}}{n_{.j'}} \right)^2,$$

ce qui correspond à une métrique de matrice $n \mathbf{D}_1^{-1}$. Ses propriétés sont similaires à celles sur les profils-lignes.



ACP des deux nuages de profils

Il y a deux possibilités qui sont en dualité exacte

Profils-lignes

- tableau de données $\mathbf{X} = \mathbf{D}_1^{-1}\mathbf{N}$;
- métrique $\mathbf{M} = n\mathbf{D}_2^{-1}$;
- poids $\mathbf{D} = \frac{\mathbf{D}_1}{n}$.

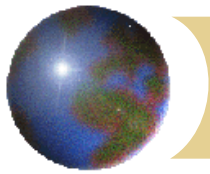
Profils-colonnes

- tableau de données $\mathbf{X} = \mathbf{D}_2^{-1}\mathbf{N}'$;
- métrique $\mathbf{M} = n\mathbf{D}_1^{-1}$;
- poids $\mathbf{D} = \frac{\mathbf{D}_2}{n}$.

Autres données

- Centre de gravité $\mathbf{g} = \mathbf{X}'\mathbf{D}\mathbf{1}$.
- Matrice de variance-covariance

$$\mathbf{V} = \mathbf{X}'\mathbf{D}\mathbf{X} - \mathbf{g}\mathbf{g}' = (\mathbf{X} - \mathbf{1}\mathbf{g}')'\mathbf{D}(\mathbf{X} - \mathbf{1}\mathbf{g}')$$



g est un facteur principal

Pourquoi? g est vecteur propre de VM associé à la valeur propre 0 car, comme g est χ^2 -orthogonal à W ,

$$VMg = (X - 1g')'D(X - 1g')Mg = 0$$

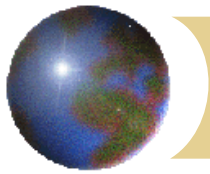
On a donc

$$X'DXMg = VMg + gg'Mg = 0 + g\|g\|_{\chi^2} = g.$$

Les autres valeurs et vecteurs propres de VM et $X'DXM$ sont identiques car, pour tout vecteur $u \perp g$

$$X'DXM u = VM u + gg'M u = VM u + g\langle g, u \rangle_{\chi^2} = VM u.$$

Centrage il est inutile de centrer les tableaux de profil ; on effectue une ACP non centrée et on élimine la valeur propre 1 associée à l'axe principal g et au facteur principal $Mg = 1$.



Calcul de l'ACP

On fait d'abord le calcul pour les profils-lignes.

Facteurs principaux ils sont vecteurs propres de

$$\mathbf{MX}'\mathbf{DX} = (n\mathbf{D}_2^{-1})(\mathbf{D}_1^{-1}\mathbf{N})'\frac{\mathbf{D}_1}{n}(\mathbf{D}_1^{-1}\mathbf{N}) = \mathbf{D}_2^{-1}\mathbf{N}'\mathbf{D}_1^{-1}\mathbf{N}.$$

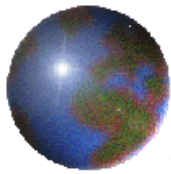
On a donc pour chaque axes principal k

$$\mathbf{D}_2^{-1}\mathbf{N}'\mathbf{D}_1^{-1}\mathbf{N}\mathbf{u}_k = \lambda_k\mathbf{u}_k$$

Composantes principales la composante principale associée au facteur $\mathbf{u}_k$ est $\mathbf{a}_k = \mathbf{X}\mathbf{u}_k = \mathbf{D}_1^{-1}\mathbf{N}\mathbf{u}_k$; ell est vecteur propre de la matrice $\mathbf{D}_1^{-1}\mathbf{N}\mathbf{D}_2^{-1}\mathbf{N}'$ car

$$\mathbf{D}_1^{-1}\mathbf{N}\mathbf{D}_2^{-1}\mathbf{N}'\mathbf{a}_k = \mathbf{D}_1^{-1}\mathbf{N}\mathbf{D}_2^{-1}\mathbf{N}'\mathbf{D}_1^{-1}\mathbf{N}\mathbf{u}_k = \lambda_k\mathbf{D}_1^{-1}\mathbf{N}\mathbf{u}_k = \lambda_k\mathbf{a}_k$$

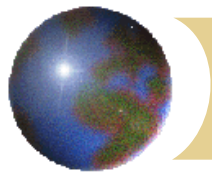
Profils-colonnes on échange les indices 1 et 2 et on transpose $\mathbf{N}$.



Comparaison lignes-colonnes

	<i>ACP profils-lignes</i>	<i>ACP profils-colonnes</i>
<i>Facteurs principaux</i>	Vecteurs propres de $\mathbf{D}_2^{-1} \mathbf{N}' \mathbf{D}_1^{-1} \mathbf{N}$	Vecteurs propres de $\mathbf{D}_1^{-1} \mathbf{N} \mathbf{D}_2^{-1} \mathbf{N}'$
<i>Composantes principales</i>	Vecteurs propres de $\mathbf{D}_1^{-1} \mathbf{N} \mathbf{D}_2^{-1} \mathbf{N}'$ normalisés par $\mathbf{a}'_k \frac{\mathbf{D}_1}{n} \mathbf{a}_k = \lambda_k$	Vecteurs propres de $\mathbf{D}_2^{-1} \mathbf{N}' \mathbf{D}_1^{-1} \mathbf{N}$ normalisés par $\mathbf{b}'_k \frac{\mathbf{D}_2}{n} \mathbf{b}_k = \lambda_k$

Comparaison les deux analyses conduisent aux mêmes valeurs propres et les facteurs principaux de l'une sont les composantes principales de l'autre (à un facteur près).



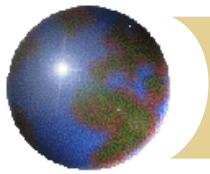
Interprétation des résultats

Coordonnées des points Les coordonnées des points-lignes et points-colonnes s'obtiennent en cherchant les vecteurs propres des produits des deux tableaux de profils. Ce sont les grandeurs principales à obtenir.

Projection des nuages il est possible de projeter les deux nuages de points sur la même représentation. On justifiera plus tard le sens de cette représentation et son interprétation.

Cercle des corrélations il n'a aucun intérêt ici, puisque les véritables variables sont qualitatives.

(non) effet de taille comme les composantes variables sont centrées ($\sum_{i=1}^{m_1} n_{i.} a_{ki} = \sum_{j=1}^{m_2} n_{.j} b_{kj} = 0$), on sait que les coordonnées des a_k et b_k ne peuvent être toutes de même signe ; il n'y a donc jamais d'effet de « taille ».



Contributions à l'inertie

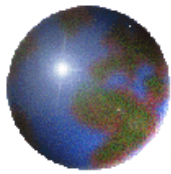
Contribution des profils-lignes On sait que $\lambda_k = \sum_{i=1}^{m_1} \frac{n_{i.}}{n} (a_{ki})^2$, où a_{ki} est la coordonnée du profil-ligne i sur la k ième composante principale de l'ACP sur les profils-lignes. On définit donc la contribution de i à l'axe principal k comme

$$\frac{n_{i.}}{n} \cdot \frac{(a_{ki})^2}{\lambda_k}.$$

On considérera les catégories ayant les influences les plus importantes (typiquement $> n_{i.}/n$) comme constitutives des axes ; on regardera aussi le signe de la coordonnée.

Contribution des profils-colonnes pour les mêmes raisons, la contribution du profil-colonne k est

$$\frac{n_{.j}}{n} \cdot \frac{(b_{kj})^2}{\lambda_k}.$$



Formules de transition

But on cherche une relation entre les vecteurs $\mathbf{a}_k$ et $\mathbf{b}_k$ pour éviter de faire deux diagonalisation de matrices. Par exemple, si $m_1 < m_2$, on diagonalisera la matrice $\mathbf{D}_1^{-1}\mathbf{N}\mathbf{D}_2^{-1}\mathbf{N}'$.

Formules un calcul simple donne les formules suivantes

$$\mathbf{b}_k = \frac{1}{\sqrt{\lambda_k}} \mathbf{D}_2^{-1} \mathbf{N}' \mathbf{a}_k, \text{ soit } b_{kj} = \frac{1}{\sqrt{\lambda_k}} \sum_{i=1}^{m_1} \frac{n_{ij}}{n_{.j}} a_{ki},$$
$$\mathbf{a}_k = \frac{1}{\sqrt{\lambda_k}} \mathbf{D}_1^{-1} \mathbf{N} \mathbf{b}_k, \text{ soit } a_{ki} = \frac{1}{\sqrt{\lambda_k}} \sum_{j=1}^{m_2} \frac{n_{ij}}{n_{i.}} b_{kj}.$$

Méthode comme $\mathbf{a}_k$ est (à une normalisation près) le facteur principal associé à $\mathbf{b}_k$, on sait que $\mathbf{b}_k = \alpha \mathbf{D}_2^{-1} \mathbf{N}' \mathbf{a}_k$. Pour déterminer α , il suffit d'écrire que $\mathbf{b}_k' \frac{\mathbf{D}_2}{n} \mathbf{b}_k = \lambda_k$.