Variables quantitatives : analyse en composantes principales

Jean-Marc Lasgouttes

`http://ana-donnees.lasgouttes.net/`

Préambule : 3 approches des données

Décrire les données de 3 manières complémentaires
- *statistique* : chaque colonne représente une variable mesurée sur différent individus,
- *matricielle* : le tableau complet de données est une matrice de nombres réels,
- *géométrique* : chaque ligne du tableau représente les coordonnées d'un point dans un espace dont la dimension est le nombre de variables.

Combiner ces trois approches pour définir l'ACP en termes de
- *vision statistique* : moyenne, variance, corrélation ;
- *vision matricielle* : valeurs propres, vecteurs propres ;
- *vision géométrique* : distances, angles, projection.

Conséquences sur le cours
- les trois premières parties sont des préliminaires qui durent la moitié du cours !
- il faut faire attention pour comprendre le rôle des différentes approches

Partie I. Données : vision statistique

Les données quantitatives

Définition On appelle « variable » un vecteur $\mathbf{x}$ de taille n. Chaque coordonnée x_i correspond à un individu. On s'intéresse ici à des valeurs numériques.

Poids Chaque individu peut avoir un poids p_i, tel que $p_1 + \cdots + p_n = 1$, notamment quand les individus n'ont pas la même importance (échantillons redressés, données regroupées,...). On a souvent $p = 1/n$.

Moyenne arithmétique On note

$$\bar{x} = \sum_{i=1}^{n} p_i x_i = p_1 x_1 + p_2 x_2 + \cdots + p_n x_n,$$

ou pour des données non pondérés

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{1}{n} \left[x_1 + x_2 + \cdots + x_n \right].$$

Propriétés la moyenne arithmétique est une mesure de *tendance centrale* qui dépend de toutes les observations et est sensible aux valeurs extrêmes. Elle est très utilisée à cause de ses bonnes propriétés mathématiques.

Variance et écart-type

Définition la *variance* de $\mathbf{x}$ est définie par

$$\operatorname{var}(\mathbf{x}) = \sigma_{\mathbf{x}}^2 = \sum_{i=1}^{n} p_i (x_i - \bar{x})^2 \text{ ou } \operatorname{var}(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

L'*écart-type* $\sigma_{\mathbf{x}}$ est la racine carrée de la variance.

Propriétés La variance satisfait la formule suivante

$$\operatorname{var}(\mathbf{x}) = \sum_{i=1}^{n} p_i x_i^2 - (\bar{x})^2$$

La variance est « la moyenne des carrés moins le carré de la moyenne ». L'écart-type, qui a la même unité que $\mathbf{x}$, est une mesure de *dispersion*.

Attention ! les calculatrices utilisent l'estimateur sans biais de la variance dans lequel le $1/n$ est remplacé par $1/(n-1)$.

Mesure de liaison entre deux variables

Définitions la covariance observée entre deux variables $\mathbf{x}$ et $\mathbf{y}$ est

$$\operatorname{cov}(\mathbf{x}, \mathbf{y}) = \sigma_{\mathbf{xy}} = \sum_{i=1}^{n} p_i (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} p_i x_i y_i - \bar{x}\bar{y}.$$

et le *coefficient de r de Bravais-Pearson* ou coefficient de corrélation est donné par

$$\operatorname{cor}(\mathbf{x}, \mathbf{y}) = r_{\mathbf{xy}} = \frac{\sigma_{\mathbf{xy}}}{\sigma_{\mathbf{x}} \sigma_{\mathbf{y}}} = \frac{\operatorname{cov}(\mathbf{x}, \mathbf{y})}{\sqrt{\operatorname{var}(\mathbf{x})} \sqrt{\operatorname{var}(\mathbf{y})}}.$$

Propriétés
- $\operatorname{cov}(\mathbf{x}, \mathbf{x}) = \operatorname{var}(\mathbf{x})$ et $\operatorname{cor}(\mathbf{x}, \mathbf{x}) = 1$
- $\operatorname{cov}(\mathbf{x}, \mathbf{y}) = \operatorname{cov}(\mathbf{y}, \mathbf{x})$ et donc $\operatorname{cor}(\mathbf{x}, \mathbf{y}) = \operatorname{cor}(\mathbf{y}, \mathbf{x})$.

Propriétés du coefficient de corrélation

Borne On a toujours (inégalité de Cauchy-Schwarz)

$$-1 \leq \operatorname{cor}(\mathbf{x}, \mathbf{y}) \leq 1.$$

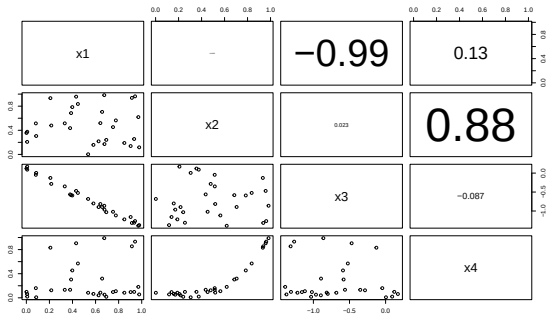
Variables liées $|\operatorname{cor}(\mathbf{x}, \mathbf{y})| = 1$ si et seulement si $\mathbf{x}$ et $\mathbf{y}$ sont linéairement liées :

$$a x_i + b y_i = c, \text{ pour tout } 1 \leq i \leq n.$$

En particulier, $\operatorname{cor}(\mathbf{x}, \mathbf{x}) = 1$.

Variables décorrélées si $\operatorname{cor}(\mathbf{x}, \mathbf{y}) = 0$, on dit que les variables sont *décorrélées*. Cela ne veut pas dire qu'elles sont indépendantes !

Le coefficient de corrélation par l'exemple



Interprétation on a 4 variables numériques avec 30 individus. Les variables 1 et 2 sont « indépendantes » ; les variables 1 et 3 ont une relation linéaire ; les variables 2 et 4 ont une relation non-linéaire.

Que signifie une corrélation linéaire ?

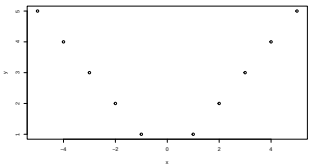
Qu'est ce qui est significatif ? si on a assez de données, on peut considérer qu'une corrélation supérieure à $0,5$ est significative, et une corrélation entre $0,3$ et $0,5$ est faible.

Une corrélation égale à 1 indique que les deux variables sont équivalentes.

Qu'est-ce que cela veut dire ? une corrélation significative indique une liaison entre deux variables, mais pas nécessairement un lien de causalité. Exemple :

> En 2016, 59,2 % des décès ont eu lieu dans des établissements de santé (hôpital ou clinique) et 26% à domicile. L'hôpital est-il dangereux pour la santé ?

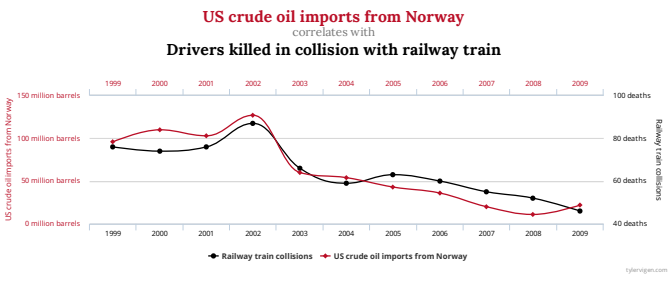
Et une décorrélation ? voici un exemple ou $\text{cor}(\mathbf{x}, \mathbf{y}) = 0$



Fausses corrélations

Quand ? Elles peuvent se trouver quand on a peu de données

Exemple Importations de pétrole brut de la Norvège vers les États-Unis et nombre de conducteurs tués par une collision avec un train : $r = 0,95$ entre 1999 et 2009.



Exemple issu du site *Spurious Correlations*

http://www.tylervigen.com/spurious-correlations.

Partie II. Données : vision matricielle

Notation

Matrice tableau de données, notée par un lettre majuscule grasse (ex : $\mathbf{A}$).

Vecteur matrice à une seule colonne, noté par une lettre minuscule grasse (ex : $\mathbf{x}$).

Cas particuliers matrices zéro ($n \times p$), identité ($n \times n$) et vecteur unité de taille n :

$$\mathbf{0}_{np} = \begin{bmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \\ 0 & & 0 \end{bmatrix}, \quad \mathbf{I}_n = \begin{bmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{bmatrix}, \quad \mathbf{1}_n = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}.$$

Transposition échange des lignes et des colonnes d'une matrice ; on note $\mathbf{A}'$ la transposée de $\mathbf{A}$.

Trace la trace d'une matrice carrée est la somme des termes de sa diagonale

$$\text{Tr}(\alpha\mathbf{A}) = \alpha\text{Tr}(\mathbf{A}), \quad \text{Tr}(\mathbf{A}+\mathbf{B}) = \text{Tr}(\mathbf{A}) + \text{Tr}(\mathbf{B}),$$
$$\text{Tr}(\mathbf{AB}) = \text{Tr}(\mathbf{BA}),$$

$$\text{Tr}(\mathbf{ABC}) = \text{Tr}(\mathbf{CAB}) = \text{Tr}(\mathbf{BCA}) \neq \text{Tr}(\mathbf{CBA})$$

Tableau de données

On note x_i^j la valeur de la *variable* $\mathbf{x}^j$ pour le i^{e} individu. $\mathbf{X} = (\mathbf{x}^1, \ldots, \mathbf{x}^p)$ est une matrice rectangulaire à n lignes et p colonnes.

$$\mathbf{x}^j = \begin{bmatrix} x_1^j \\ x_2^j \\ \\ \vdots \\ \\ x_n^j \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} x_1^1 & x_1^2 & & & \cdots & x_1^p \\ x_2^1 & x_2^2 & & & & \\ & & & \vdots & & \\ \vdots & & \cdots & x_i^j & & \\ & & & & \ddots & \\ x_n^1 & & & & & x_n^p \end{bmatrix}.$$

Un *individu* est représenté par

$$\mathbf{e}_i' = [x_i^1, \ldots, x_i^j, \ldots, x_i^p]$$

La matrice des poids

Définition on associe aux individus un poids p_i tel que

$$p_1 + \cdots + p_n = 1$$

que l'on représente par la matrice diagonale de taille n

$$\mathbf{D_p} = \begin{bmatrix} p_1 & & & 0 \\ & p_2 & & \\ & & \ddots & \\ 0 & & & p_n \end{bmatrix}.$$

Symétrie La matrice $\mathbf{D_p}$ est diagonale et donc symétrique : $\mathbf{D_p'} = \mathbf{D_p}$.

Cas uniforme tous les individus ont le même poids $p_i = 1/n$ et $\mathbf{D_p} = \frac{1}{n}\mathbf{I}_n$.

Point moyen et tableau centré

Point moyen c'est le vecteur $\mathbf{g}$ des moyennes arithmétiques de chaque variable :

$$\mathbf{g}' = (\bar{x}^1, \ldots, \bar{x}^p) = \sum_{i=1}^{n} p_i \mathbf{e}_i'.$$

On peut écrire sous forme matricielle

$$\mathbf{g} = \mathbf{X}'\mathbf{D_p}\mathbf{1}_n.$$

Tableau centré il est obtenu en centrant les variables autour de leur moyenne

$$y_i^j = x_i^j - \bar{x}^j, \quad \text{c'est-à-dire} \quad \mathbf{y}^j = \mathbf{x}^j - \bar{x}^j\mathbf{1}_n$$

ou, en notation matricielle,

$$\mathbf{Y} = \mathbf{X} - \mathbf{1}_n\mathbf{g}' = (\mathbf{I}_n - \mathbf{1}_n\mathbf{1}_n'\mathbf{D_p})\mathbf{X}$$

Matrice de variance-covariance

Définition c'est une matrice *carrée* de dimension p

$$\mathbf{V} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & & & \\ \vdots & & \ddots & \\ \sigma_{p1} & & & \sigma_p^2 \end{bmatrix},$$

où $\sigma_{j\ell}$ est la covariance des variables $\mathbf{x}^j$ et $\mathbf{x}^\ell$ et σ_j^2 est la variance de la variable $\mathbf{x}^j$

Symétrie Comme $\sigma_{j\ell} = \sigma_{\ell j}$, la matrice $\mathbf{V}$ est symétrique : $\mathbf{V}' = \mathbf{V}$.

Formule matricielle

$$\mathbf{V} = \mathbf{X}'\mathbf{D_p}\mathbf{X} - \mathbf{g}\mathbf{g}' = \mathbf{Y}'\mathbf{D_p}\mathbf{Y}.$$

Matrice de corrélation

Définition Si l'on note $r_{j\ell} = \sigma_{j\ell}/\sigma_j\sigma_\ell$, c'est la matrice $p \times p$

$$\mathbf{R} = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & & \\ \vdots & & \ddots & \\ r_{p1} & & & 1 \end{bmatrix},$$

Symétrie Comme $r_{j\ell} = r_{\ell j}$, la matrice $\mathbf{R}$ est symétrique : $\mathbf{R}' = \mathbf{R}$.

Formule matricielle $\mathbf{R} = \mathbf{D}_{1/\boldsymbol{\sigma}}\mathbf{V}\mathbf{D}_{1/\boldsymbol{\sigma}}$, où

$$\mathbf{D}_{1/\boldsymbol{\sigma}} = \begin{bmatrix} \frac{1}{\sigma_1} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{\sigma_p} \end{bmatrix}$$

Les données centrées réduites

Définition c'est la matrice $\mathbf{Z}$ contenant les données

$$z_i^j = \frac{y_i^j}{\sigma_j} = \frac{x_i^j - \bar{x}^j}{\sigma_j}, \quad \text{c'est-à-dire} \quad \mathbf{z}^j = \frac{\mathbf{y}^j}{\sigma_j}$$

qui se calcule matriciellement comme $\mathbf{Z} = \mathbf{Y}\mathbf{D}_{1/\boldsymbol{\sigma}}$

Pourquoi réduites ?

— pour que les distances soient indépendantes des unités de mesure,

— pour ne pas privilégier les variables dispersées.

Covariances Les covariances des $\mathbf{z}^j$ sont des corrélations :

$$\operatorname{cov}(\mathbf{z}^k, \mathbf{z}^\ell) = \sum_{i=0}^{n} p_i z_i^k z_i^\ell = \frac{1}{\sigma_k\sigma_\ell}\sum_{i=0}^{n} p_i y_i^k y_i^\ell = \operatorname{cor}(\mathbf{x}^k, \mathbf{x}^\ell),$$

où on a utilisé le fait que $\bar{z}^j = \bar{y}^j = 0$. La matrice de variance-covariance des variables centrées-réduites est donc la matrice de corrélation $\mathbf{R}$.

Partie III. Données : vision géométrique

L'analyse de composantes principales (ACP)

Contexte chaque individu est considéré comme un point d'un espace vectoriel F de dimension p. Ses coordonnées dans F sont

$$(x_i^1, x_i^2, \ldots, x_i^p).$$

L'ensemble des individus est un *nuage de points* dans F et $\mathbf{g}$ est son *centre de gravité*.

Principe on cherche à réduire le nombre p de variables tout en préservant au maximum la structure du problème.

Pour cela on projette le nuage de points sur un sous-espace de dimension inférieure.

Distance entre individus

Motivation afin de pouvoir considérer la structure du nuage des individus, il faut définir une distance, qui induira une géométrie.

Distance euclidienne classique la distance la plus simple entre deux points de $\mathbb{R}^p$ est définie par

$$d^2(\mathbf{u}, \mathbf{v}) = \sum_{j=1}^{p}(u_j - v_j)^2 = \|\mathbf{u} - \mathbf{v}\|^2$$

Généralisation simple on donne un poids $m_j > 0$ à la variable j

$$d^2(\mathbf{u}, \mathbf{v}) = \sum_{j=1}^{p} m_j(u_j - v_j)^2$$

Cela revient à multiplier la coordonnée j par $\sqrt{m_j}$

Métrique

Définition soit $\mathbf{M} = \text{diag}(m_j)$, où $m_1, \ldots, m_p$ sont des réels strictement positifs. On pose

$$\|\mathbf{u}\|^2_{\mathbf{M}} = \sum_{j=1}^{p} m_j u_j^2 = \mathbf{u}'\mathbf{M}\mathbf{u}, \quad d^2_{\mathbf{M}}(\mathbf{u},\mathbf{v}) = \|\mathbf{u}-\mathbf{v}\|^2_{\mathbf{M}}.$$

Espace métrique il est défini par le produit scalaire

$$\langle \mathbf{u},\mathbf{v}\rangle_{\mathbf{M}} = \sum_{j=1}^{p} m_j u_j v_j = \mathbf{u}'\mathbf{M}\mathbf{v}, \quad \langle \mathbf{u},\mathbf{u}\rangle_{\mathbf{M}} = \|\mathbf{u}\|^2_{\mathbf{M}}.$$

Propriétés Le produit scalaire est commutatif, linéaire et satisfait l'identité

$$\|\mathbf{u}+\mathbf{v}\|^2_{\mathbf{M}} = \|\mathbf{u}\|^2_{\mathbf{M}} + \|\mathbf{v}\|^2_{\mathbf{M}} + 2\langle \mathbf{u},\mathbf{v}\rangle_{\mathbf{M}}$$

Orthogonalité on dit que $\mathbf{u}$ et $\mathbf{v}$ sont $\mathbf{M}$-orthogonaux si $\langle \mathbf{u},\mathbf{v}\rangle_{\mathbf{M}} = 0$.

Cas particuliers

Métrique usuelle Si $m_1, \ldots, m_p = 1$, alors $\mathbf{M} = \mathbf{I}_p$ et on note $\langle \mathbf{u},\mathbf{v}\rangle = \langle \mathbf{u},\mathbf{v}\rangle_{\mathbf{I}}$.

Métrique réduite diviser les variables par σ_j est équivalent à prendre $m_j = 1/\sigma_j^2$. On a $\mathbf{D}_{1/\boldsymbol{\sigma}^2} = \mathbf{D}_{1/\boldsymbol{\sigma}}\mathbf{D}_{1/\boldsymbol{\sigma}}$ et donc

$$\langle \mathbf{D}_{1/\boldsymbol{\sigma}}\mathbf{u}, \mathbf{D}_{1/\boldsymbol{\sigma}}\mathbf{v}\rangle = \mathbf{u}'\mathbf{D}_{1/\boldsymbol{\sigma}}\mathbf{D}_{1/\boldsymbol{\sigma}}\mathbf{v} = \mathbf{u}'\mathbf{D}_{1/\boldsymbol{\sigma}^2}\mathbf{v} = \langle \mathbf{u},\mathbf{v}\rangle_{\mathbf{D}_{1/\boldsymbol{\sigma}^2}}.$$

Travailler avec la métrique $\mathbf{D}_{1/\boldsymbol{\sigma}^2}$, c'est comme utiliser la métrique $\mathbf{I}$ sur des variables réduites.

La plupart du temps en ACP, on fait l'analyse avec la métrique usuelle sur les données centrées-réduites.

Partie IV. L'analyse en composantes principales

Inertie

Définition l'inertie en un point $\mathbf{v}$ du nuage de points est

$$I_{\mathbf{v}} = \sum_{i=1}^{n} p_i \|\mathbf{e}_i - \mathbf{v}\|^2_{\mathbf{M}} = \sum_{i=1}^{n} p_i (\mathbf{e}_i - \mathbf{v})'\mathbf{M}(\mathbf{e}_i - \mathbf{v}).$$

Inertie totale La plus petite inertie possible est $I_{\mathbf{g}}$, donnée par

$$I_{\mathbf{g}} = \sum_{i=1}^{n} p_i \|\mathbf{e}_i - \mathbf{g}\|^2_{\mathbf{M}} = \sum_{i=1}^{n} p_i (\mathbf{e}_i - \mathbf{g})'\mathbf{M}(\mathbf{e}_i - \mathbf{g})$$

qui est la seule intéressante puisque $I_{\mathbf{v}} = I_{\mathbf{g}} + \|\mathbf{v}-\mathbf{g}\|^2_M$.

Autres relations $I_{\mathbf{g}}$ mesure la moyenne des carrés des distances entre les individus

$$2I_{\mathbf{g}} = \sum_{i=1}^{n}\sum_{j=1}^{n} p_i p_j \|\mathbf{e}_i - \mathbf{e}_j\|^2_{\mathbf{M}}.$$

Interprétation L'inertie totale mesure l'étalement du nuage de points

Calcul de l'inertie

Forme matricielle L'inertie totale est aussi donnée par la trace de la matrice $\mathbf{VM}$ (ou $\mathbf{MV}$)

$$I_{\mathbf{g}} = \text{Tr}(\mathbf{VM}) = \text{Tr}(\mathbf{MV})$$

Métrique usuelle $\mathbf{M} = \mathbf{I}_p$ correspond au produit scalaire usuel et

$$I_{\mathbf{g}} = \text{Tr}(\mathbf{V}) = \sum_{j=1}^{p} \sigma_i^2$$

Métrique réduite obtenue quand $\mathbf{M} = \mathbf{D}_{1/\boldsymbol{\sigma}^2} = \mathbf{D}^2_{1/\boldsymbol{\sigma}}$

$$I_{\mathbf{g}} = \text{Tr}(\mathbf{D}_{1/\boldsymbol{\sigma}^2}\mathbf{V}) = \text{Tr}(\mathbf{D}_{1/\boldsymbol{\sigma}}\mathbf{V}\mathbf{D}_{1/\boldsymbol{\sigma}}) = \text{Tr}(\mathbf{R}) = p.$$

Variables centrées réduites On se retrouve encore dans le cas où

$$I_{\mathbf{g}} = \text{Tr}(\mathbf{R}) = p.$$

L'analyse de composantes principales (version 2)

Principe on cherche à projeter $\mathbf{M}-$orthogonalement le nuage de points sur un espace F_k de dimension $k < p$, sous la forme

$$\mathbf{e}_i^* - \mathbf{g} = c_{i1}\mathbf{a}_1 + c_{i2}\mathbf{a}_2 + \cdots + c_{ik}\mathbf{a}_k$$

Les vecteurs $\mathbf{a}_1, \ldots, \mathbf{a}_k$ définissent l'espace F_k et les $c_{i\ell}$ sont les coordonnées de $\mathbf{e}_i^*$.

Critère on veut que la moyenne des carrés des distances entre les points $\mathbf{e}_i$ et leur projetés $\mathbf{e}_i^*$ soit minimale. Comme on a toujours (théorème de Pythagore)

$$\|\mathbf{e}_i - \mathbf{g}\|^2 = \|\mathbf{e}_i - \mathbf{e}_i^*\|^2 + \|\mathbf{e}_i^* - \mathbf{g}\|^2,$$

cela revient à maximiser l'inertie du nuage projeté.

On cherche donc F_k, sous espace de dimension k de F_p, qui maximise l'inertie du nuage projeté sur F_k.

Résultat principal

Propriété Il existe p réels $\lambda_1, \ldots, \lambda_p$ positifs ou nuls et p vecteurs $\mathbf{a}_1, \ldots, \mathbf{a}_p$, tels que

$$\mathbf{VM}\mathbf{a}_k = \lambda_k \mathbf{a}_k.$$

— Les λ_k sont les *valeurs propres* de $\mathbf{VM}$ et sont classées par ordre décroissant :

$$\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \cdots \geq \lambda_p \geq 0.$$

— Les $\mathbf{a}_k$ sont les vecteurs propres de $\mathbf{VM}$ et sont « $\mathbf{M}$-orthonormaux » :

$$\langle \mathbf{a}_k, \mathbf{a}_k\rangle_{\mathbf{M}} = 1, \qquad \langle \mathbf{a}_k, \mathbf{a}_\ell\rangle_{\mathbf{M}} = 0 \text{ si } k \neq \ell.$$

Théorème principal *La projection sur k variables qui maximise l'inertie projetée est obtenue en considérant les k premières valeurs propres $\lambda_1, \ldots, \lambda_k$ et les $\mathbf{a}_1, \ldots, \mathbf{a}_k$ correspondants, appelés* axes principaux.

Le calcul ne dépend pas du nombre de variables retenues.