Article ACP

Benjamin GUIGON

Janvier 2020

1 Introduction

Lorsqu'on étudie simultanément un nombre important de variables quantitatives (ne serait-ce que 4 !), comment en faire un graphique global ? La difficulté vient de ce que les individus étudiés ne sont plus représentés dans un plan, espace de dimension 2, mais dans un espace de dimension plus importante (par exemple 4). L'objectif de l'Analyse en Composantes Principales (ACP) est de revenir à un espace de dimension réduite (par exemple 2) en déformant le moins possible la réalité (cf. l'introduction élémentaire à l'ACP). Il s'agit donc d'obtenir le résumé le plus pertinent possible des données initiales.

2 Partie 1

Soit p variables réelles X^J $(J = 1, ...p)$, observés sur n individus $(i = 1,...,n)$ affectés de poids ω_i.

$\forall i \in [1, n] : \omega_i > 0$ et $\sum \omega_i = 1$
$\forall i \in [1, n] : x_i^j = X^j(i)$, mesure sur le i-ème individus

Ces mesures sont regroupés dans une matrice X d'ordre $(n \times p)$, $\mathbb{R}$ avec la base ε.

	X^1	$\cdots$	X^j	$\cdots$	X^p
1	x_1^1	$\cdots$	x_1^j	$\cdots$	x_1^p
$\vdots$	$\vdots$		$\vdots$		$\vdots$
i	x_i^1	$\cdots$	x_i^j	$\cdots$	x_i^p
$\vdots$	$\vdots$		$\vdots$		$\vdots$
n	x_n^1	$\cdots$	x_n^j	$\cdots$	x_n^p

FIGURE 1 − Matrice initiale X d'ordre nxp

À chaque individu i est associé le vecteur x_i contenant la i-ème ligne de X mise en colonne. C'est un élément d'un espace vectoriel noté E de dimension p ; nous choisissons $\mathbb{R}^p$ muni de la base canonique E et d'une métrique de matrice M.

Les colonnes sont centrées (la moyenne de chaque colonnes lui ait retranchée)

$$D = \text{diag}(\omega_1, ..., \omega_n).$$

2.1 Métrique des poids

- Moyenne empirique de X_j :

$$\bar{x^j} = \langle Xe^j, \mathbb{1}_n \rangle_D = e^{j'} X' D \mathbb{1}_n.$$

Cette expression utilise le principe de produit scalaire par rapport à la mesure D représenté par les symboles "$\langle ... \rangle_D$". e^j représente le j-ième vecteur de la base canonique de l'espace dans lequel on se place, donc la multiplication de X par ce vecteur va permettre de récupérer le j-ième vecteur de la matrice X. La multiplication par l'indicatrice de n $\mathbb{1}_n$ va permettre de récuperer de deviser le j-ième par n. On peut simplifier l'équation et la passer sous forme matricielle. On obtient la 2ème partie de l'égalité.

- Barycentre des individus :

$$\bar{x} = X' D \mathbb{1}_n.$$

Cette formule, sous forme matricielle, permet de calculer le vecteur des barycentres de chaque individus.

- Matrice des données centrées :

$$\bar{X} = X - \mathbb{1}_n x'.$$

Une fois calculé le vecteur des barycentres, on le retranche à X pour obtenir la matrice X centrée.

- Covariance de X^j et X^k :

$$x^{j'} D x^k = \langle x^j, x^k \rangle_D.$$

Si on reprend la formule de la covariance d'une variable, on obtient $\frac{1}{N} \sum (x_i - \bar{x})^2$. La covariance de de $(x^j)'$ et x^k n'est donc rien d'autre que le produit scalaire par la mesure des D des deux.

- Ecart-type de X^j :

$$\sigma_j = (x^j D x^j)^{1/2} = \|x^j\|_D$$

L'équart-type est la racine de la variance, soit la racine de la formule juste au dessus.

- Matrice des covariances

$$S = \sum_{i=1}^{n} \omega_i (x_i - \bar{x})(x_i - \bar{x})' = \bar{X}' D \bar{X}$$

Pour retrouver la matrice de covariance, donc des covariances de toutes les variables entre elles, on utilise la formule du calcul de la covariance entre 2 variables mais en utilisant la matrice X pour avoir tous les couples de variables.

- Corrélation de X^j et X^k :

$$\frac{\langle x^j, x^k \rangle_D}{\|x^j\|_D \|x^k\|_D} = \cos \theta_D(x^j, x^k)$$

Par définition du produit scalaire, nous retrouvons que le produit scalaire de 2 variables ne donne le cosinus de l'angle entre ces 2 variables.

A noter que : la longueur d'un vecteur représente un écart-type et le cosinus d'un angle entre deux vecteurs représente une corrélation.

2.2 Objectifs

Les objectifs poursuivis par une ACP sont :

• La représentation graphique "optimale" des individus (lignes), minimisant les déformations du nuage des points, dans un sous-espace Eq de dimension q $(q < p)$.

• La représentation graphique des variables dans un sous-espace Fq en explicitant au "mieux" les liaisons initiales entre ces variables.

• La réduction de la dimension (compression), ou approximation de X par un tableau de rang q $(q < p)$.

L'ACP admet des définitions équivalentes selon que l'on s'attache à la représentation des individus, à celle des variables ou encore à leur représentation simultanée.

2.3 Modèle

Les notations sont celles du paragraphe précédent :

• X désigne le tableau des données issues de l'observation de p variables quantitatives Xj sur n individus i de poids wi,

• E est l'espace des individus muni de la base canonique et de la métrique de matrice M,

• F est l'espace des variables muni de la base canonique et de la métrique des poids D = diag(w1, . . . , wn).

De façon générale, un modèle s'écrit :

$$\textbf{Observation} = \textbf{Modele} + \textbf{Bruit}$$

assorti de différents types d'hypothèses et de contraintes sur le modèle et sur le bruit.

En ACP, la matrice des données est supposée être issue de l'observation de n vecteurs aléatoires indépendants $[x_1, ..., x_n]$, de même matrice de covariance $\sigma^2\Gamma$, mais d'espérances différentes z_i, toutes contenues dans un sousespace affine de dimension q $(q < p)$ de E. Dans ce modèle, $E(x_i) = z_i$ est un paramètre spécifique attaché à chaque individu i et appelé *effet fixe*, le modèle étant dit *fonctionnel*. Ceci s'écrit en résumé :

- $[x_i; i = 1, ..., n]$, n vecteurs aléatoires indépendants de E,
- $x_i = z_i + \varepsilon_i, i = 1, ..., n$ avec

$$\begin{cases} E(\varepsilon_i) = 0,\ var(\varepsilon_i) = \sigma^2\Gamma \\ \sigma > 0 \text{ inc. } \Gamma \text{ rég. et connue,} \end{cases} \tag{1}$$

- $\exists A_q$, sous sous-espace affine de $dim q$ de E tel que $\forall i$, $z_i \in \mathrm{A}_q$ $(q < p)$.

Du fait que E est un espace linéaire, nous pouvons retrouver la 2ème condition, si on prend $E(x_i)$, on utilise la linéarité pour retrouver la valeur de $E(\varepsilon_i)$ et $var(\varepsilon_i)$

Soit $\bar{z} = \sum_{i=1}^{n} \omega_i z_i$. Les hypothèses du modèle entraînent que $\bar{z}$ appartient à A_q. Soit donc E_q le sous-espace vectoriel de E de dimension q tel que :

$$A_q = \bar{z} + E_q$$

Nous retrouvons ce resultat grace à la stabilité de E par combinaison linéaire.

Les paramètres à estimer sont alors E_q et z_i, $i = 1, ..., n$, éventuellement σ ; z_i est la part systématique, ou effet, supposée de rang q ; éliminer le bruit revient donc à réduire la dimension.

2.4 Theorie

On considère p variable statistiques centrées $X^1, ...X^p$. Une combinaison linéaire de coefficients $\mathbf{f}_j$ de ces variables,

$$\mathbf{c} = \sum_{j=1}^{p} f_j \mathbf{x}^j = \bar{\mathbf{X}}\mathbf{f},$$

[On retrouve bien ici une combinaison linéaire des variables X ainsi que les coefficients f_j. L'autre partie de l'égalité est la forme matricielle.]

définit une nouvelle variable centrée C qui, à tout individu i, associe la "mesure"

$$C(i) = (x_i - \bar{x})' f.$$

En retranchant $\bar{x}$, on centre la variable $C(i)$, conformement à la définition de $\bar{x}$.

$PROPOSITION2$. — Soient p variables quantitatives centrées $X^1, ...X^p$ observées sur n individus de poids ω_i ; l'ACP de $(\bar{X}, M, D)$ est aussi la recherche des q combinaisons linéaires normées des X^j , non corrélées et dont la somme des variances soit maximale.

Nous allons chercher à minimiser la variance totale pour déterminer les axes principaux de l'ACP.

• Les vecteurs $f^k = M\mathbf{v}^k$ sont les *facteurs principaux*. Ils permettent de définir les combinaisons linéaires des X$j$optimales au sens ci-dessus.

• Les vecteurs $c^k = \bar{\mathbf{X}} f^k$ sont les *composantes principales*.

• Les variables C^k associées sont centrées, non corrélées et de variance λ_k ; ce sont les *variables principales*;

$$cov(C^k, C^l) = (\bar{\mathbf{X}} f^k)' D \bar{\mathbf{X}} \mathbf{f}^l = \mathbf{f^{k\prime} S f^l}$$

$$= v^{k\prime} \mathbf{MSM} v^l = \lambda_l v^{k\prime} \mathbf{M} v^l = \lambda_l {\delta_k}^l$$

Pour retrouver cette formule il faut utiliser les termes vus dans les points plus hauts. Une fois remplacé dans l'équation, il faut aller voir dans l'annexe pour retrouver que $SMv^l = \lambda_l v^l$ d'après les propriétés des valeurs propres et des vecteurs propres associés. Et enfin on retrouve que $v^{k\prime} \mathbf{M} v^l$ est équivalent à l'identité. Ce qui fait que l'on retrouve $\lambda_l {\delta_k}^l$ avec ${\delta_k}^l$ égale à 0 si $l \neq k$ et 1 si $l = k$.

On obtient donc :

$$\mathbf{C = \bar{X} F = \bar{X} M V = U \Lambda^{1/2}}$$

Ou C est la matrice des composantes principales.

Les axes définis par les vecteurs D-orthonormés u^k sont appelés axes factoriels.

2.5 Graphique

Maintenant que nous avons trouvé les composantes principales ainsi que les combinaisons linéaires qui composent les nouveaux axes. Nous allons projeter les individus sur ces nouveaux axes pour avoir une représentation graphique qui nous permettent d'exploiter les données.

2.5.1 Projection

Chaque individu i représenté par x_i est approché par sa projection Morthogonale $\widehat{z_i}^q$ sur le sous-espace $\widehat{E_q}$ engendré par les q premiers vecteurs principaux $(v^1, ..., v^q)$. En notant e_i un vecteur de la base canonique de l'espace des variables, la coordonnée de l'individu i sur v^k est donnée par :

$$\langle x_i - \bar{x}, v^k \rangle_M = (x_i - \bar{x})' \mathbf{M} v^k = {e_i}' \bar{X} \mathbf{M} v^k = c_i^k$$

En faisant le produit scalaire de la matrice X centrée, par le vecteur v^k, qui est le k-ième vecteur princiale, on obtient un scalaire, qui est la coordonnée de l'individu i sur le k-ième axe principal. Donc pour illustrer ces propos. Regardons le graphique ci-dessous.

Nous retrouvons une ACP de 32 villes de france par rapport à la température moyenne par mois pendant 12 ans. Si l'on fait un parallèle avec les calculs précédents, si $k = 1$, on obtient les coordonnées de toutes les villes sur l'axe 1 qui est l'axe principale de notre ACP.

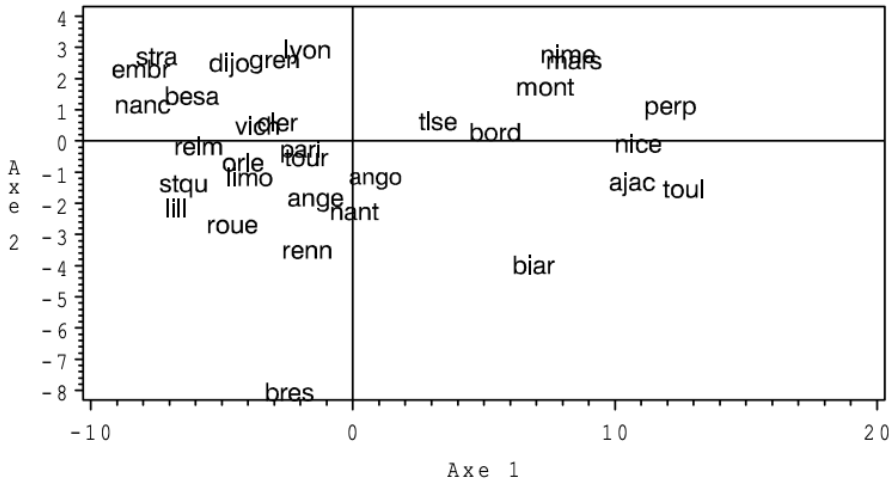


FIGURE 2 - Graphique de l'*ACP* selon l'Axe 1 et l'Axe 2

2.5.2 Qualités

La "qualité globale" des représentations est mesurée par *la part de dispersion expliquée* :

$$r_q = \frac{tr\mathbf{SM}\widehat{\mathbf{P}}_\mathbf{q}}{tr\mathbf{SM}} = \frac{\sum_{k=1}^{q} \lambda_k}{\sum_{k=1}^{p} \lambda_k}$$

En calculant la trace de la matrice $\mathbf{SM}\widehat{\mathbf{P}}_\mathbf{q}$ on obtient en effet la somme des valeurs propres, car la la matrice $\mathbf{SM}\widehat{\mathbf{P}}_\mathbf{q}$ est une matrice diagonale dans laquelle les coefficients diagonaux sont les valeurs propres. Pour trouver $\widehat{\mathbf{P}}_\mathbf{q}$ il faut aller cherche dans l'annexe. On retrouve que c'est égale à :

$$\widehat{\mathbf{P_q}} = \mathbf{V}_q \mathbf{V}_q' \mathbf{M}$$

On sait que la trace d'une matrice est invariante suivant le repaire de la matrice, on sait que λ_k sont les valeurs propres de SM par définition de SM. Donc en multipliant SM par $\widehat{\mathbf{P_q}}$, les valeurs propres sont invariantes par changement de base. Donc la trace de la matrice est bien égale à la somme des valeurs propres λ_k.

Remarque. — La dispersion d'un nuage de points unidimensionnel par rapport à sa moyenne se mesure par la variance. Dans le cas multidimensionnel, la dispersion du nuage $\mathcal{N}$ par rapport à son barycentre $\bar{x}$ se mesure par *l'inertie*, généralisation de la variance :

$$I_q(\mathcal{N}) = \sum_{i=1}^{n} \omega_i \|x_i - \bar{x}\|_M^2 = \|\bar{\mathbf{X}}\|_{M,D}^2 = tr(\bar{\mathbf{X}}'\mathbf{D}\bar{\mathbf{X}}\mathbf{M}) = tr(\mathbf{SM})$$

On retrouve la formule des moindres carrés sans la minimisation. En effet on cherche à trouver la dispersion du nuage de point.k On retrouve ensuite $\bar{x}$ de par sa définition. Enfin pour passer de $\bar{x}$ il faut utiliser la définition de la norme de *Frobenius* qui nous dit que $\|A\|_F := (tr(A * A)^{1/2}$. Donc on retrouve :

$$\|\bar{\mathbf{X}}\|_{M,D}^2 = tr(\bar{\mathbf{X}}'\mathbf{D}\bar{\mathbf{X}}\mathbf{M}).$$

La qualité de la représentation de chaque xi est donnée par le cosinus carré de l'angle qu'il forme avec sa projection :

$$[\cos\theta(x_i - \bar{x}, \widehat{z_i}^q)]^2 = \frac{\|\widehat{\mathbf{P_q}}(x_i - \bar{x})\|_M^2}{\|x_i - \bar{x}\|_M^2} = \frac{\sum_{k=1}^{q}(c_i^k)^2}{\sum_{k=1}^{p}(c_i^k)^2}$$

2.5.3 Contributions

Les contributions de chaque individu à l'inertie de leur nuage

$$\gamma_i = \frac{\omega_i \|x_i - \bar{x}\|_M^2}{tr\mathbf{SM}} = \frac{\omega_i \sum_{k=1}^{p}(c_i^k)^2}{\sum_{k=1}^{p}\lambda_k},$$

ainsi qu'à la variance d'une variable principale

$$\gamma_i^k = \frac{\omega_i (c_i^k)^2}{\lambda_k}$$

permettent de déceler les observations les plus influentes et, éventuellement, aberrantes. Ces points apparaissent visiblement lors du tracé des diagrammesboîtes parallèles des composantes principales qui évitent ainsi une lecture fastidieuse de ce tableau des contributions.

La formule de γ_i découle directement des définitions plus hautes. Par définition de c_i.

2.6 Biplot

Maintenant que nous avons les resultats ainsi que les indicateurs tels que la *qualité* et la *contribution*, nous pouvons passer à l'analyse graphique.

À partir de la décomposition en valeurs singulières de ($\bar{X}$,M, D), on remarque que chaque valeur

$$x_i^j - \bar{x}^j = \sum_{k=1}^{p} \sqrt{\lambda_k} u_i^k v_k^j = [U\Lambda^{1/2}V']_i^j$$

[On reprend ici la décomposition en valeur singulière faite dans l'annèxe.]

s'exprime comme produit scalaire usuel des vecteurs

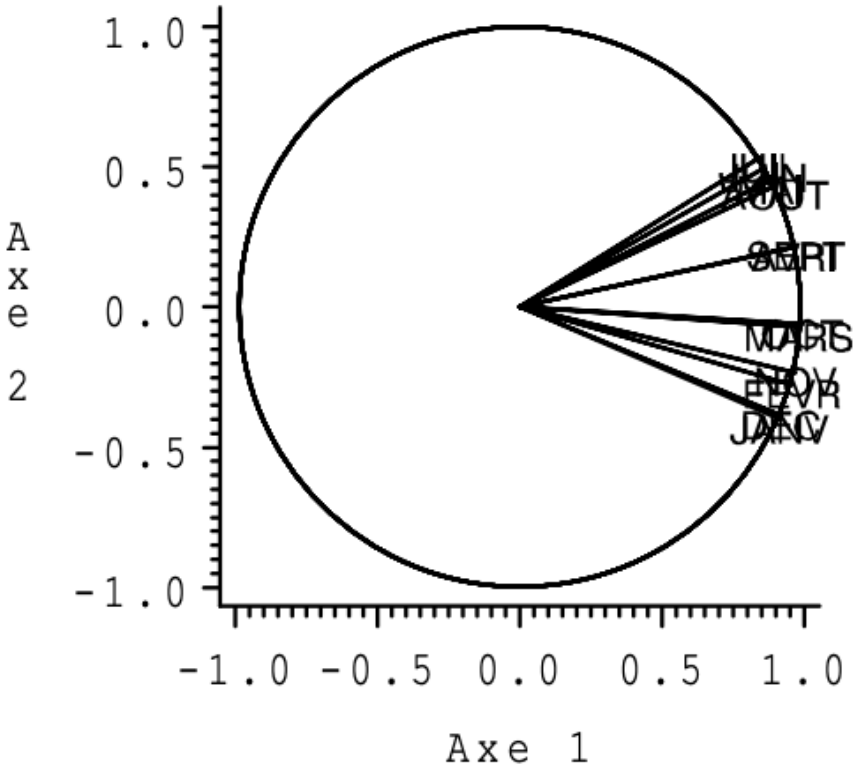
$$c_i = [U\Lambda^{1/2}]_i \text{ et } v^j \text{ ou encore } u_i \text{ et } [V\Lambda^{1/2}]_j$$

On remarque en effet que $[U\Lambda^{1/2}V']_i^j$ est l'équivalent du produit scalaire de $[U\Lambda^{1/2}]_i$ par v^j et pareil pour u_i.

Nous pouvons dresser 2 types de graphique pour l'ACP, un graphique *isométrique ligne* et *isométrique colonne*.

1) La preprésentation *isométrique ligne* :
La représentation isométrique ligne utilise les matrices C et V ; elle permet d'interpréter les distances entre individus ainsi que les produits scalaires entre un individu et une variable qui sont, dans le premier plan principal, des approximations des valeurs observées $X^j(\omega_i)$;

2) La preprésentation *isométrique colonne* :
La représentation isométrique colonne utilise les matrices U et $V\Lambda^{1/2}$; elle permet d'interpréter les angles entre vecteurs variables (corrélations) et les produits scalaires comme précédemment.


1.0
0.5
0.0
-0.5
-1.0
Axe 2
-1.0 -0.5 0.0 0.5 1.0
Axe 1

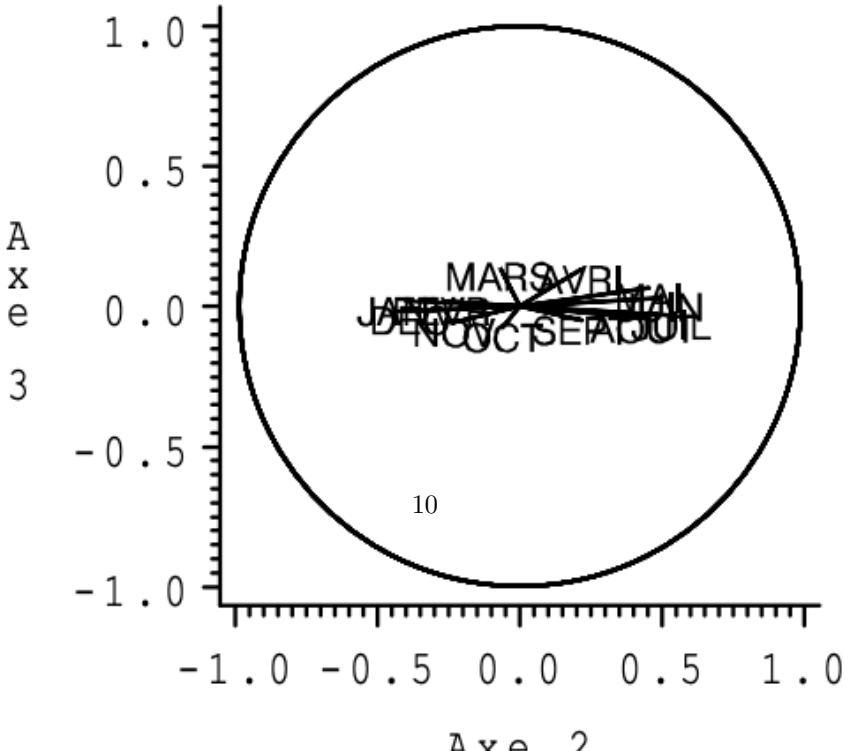

1.0
0.5
0.0
-0.5
-1.0
Axe 3
-1.0 -0.5 0.0 0.5 1.0
Axe 2

FIGURE 3 - Représentation *isometriqueligne* et *isometriquecolonne*

2.7 Annexe

2.7.1 Estimation

PROPOSITION 1. — *L'estimation des paramètres de* (1) *est fournie par l'ACP de* (X, M, D) *c'est-à-dire par la décomposition en valeurs singulières de* $(\bar{X}, M, D)$:

$$\widehat{Z_q} = \sum_{k=1}^{q} \lambda_k^{1/2} \mathbf{u}^k \mathbf{v}^k = \mathbf{U}_q \Lambda^{1/2} \mathbf{V'}_q.$$

Nous ne detaillerons pas la preuve, mais elle découle du théorème de la décomposition en valeurs singulières et du théorième d'approximation matricielles.

• Les $\mathbf{u}^k$ sont les vecteurs propres D-orthonormés de la matrice $\bar{X}M\bar{X}'D$ associés aux valeurs propres λ_k rangées par ordre décroissant.
• Les $\mathbf{v}^k$, appelés vecteurs principaux, sont les vecteurs propres M-orthonormés de la matrice $\bar{X}'D\bar{X}M = SM$ associés aux mêmes valeurs propres ; ils engendrent des sous espaces vectoriels (s.e.v) de dimension 1 appelés axes principaux.
• λ_k est la k-ième valeurs propres
• Λ est un lambda majuscule qui représente la matrice des valeurs propres.
La 2ème partie de l'équation n'est autre que la forme matricielle de la formule.

Les estimations sont donc données par :

$$\widehat{\bar{z}} = \bar{\mathbf{x}},$$

$$\widehat{Z_q} = \sum_{k=1}^{q} \lambda_k^{1/2} \mathbf{u}^k \mathbf{v}^k = \mathbf{U}_q \Lambda^{1/2} \mathbf{V'}_q = \bar{\mathbf{X}} \widehat{P'_q} \ ,$$

ou $\widehat{P_q} = \mathbf{V}_q \mathbf{V}'_q M$ est la matrice de projection M-orthogonale sur $\widehat{E_q}$,

$$\widehat{E_q} = vect(\mathbf{v}^1, ..., \mathbf{v}^q) \ ,$$

$\widehat{E_2}$ est appelé plan principal,

$$\widehat{\mathbf{z}_i} = \widehat{P}_q \mathbf{x}_i + \bar{\bar{x}}$$