Thème 1

A.C.P

Département Sciences du Numérique
Janvier 2020

On présente ici une **méthode factorielle** d'analyse de données. Il s'agit d'explorer puis décrire un tableau de n individus caractérisés par p **variables quantitatives** actives. Une méthode factorielle va chercher à substituer les p variables (jugées trop nombreuses, typiquement pour une visualisation) par des **facteurs synthétiques**, c'est-à-dire de nouvelles variables, qui expliquent en grande partie la structure des données en présence. Par variables actives, on entend celles qui vont compter dans la construction des facteurs. D'autres variables supplémentaires (par exemple qualitatives) peuvent enrichir l'interprétation.

Un tableau de données est donné en exemple ci-dessous.

	Janv	Févr	Mars	Avri	Mai	Juin	juil	Août	Sept	Octo	Nove	Déce	Lati	Long
Bordeaux	5.6	6.6	10.3	12.8	15.8	19.3	20.9	21	18.6	13.8	9.1	6.2	44.5	-0.34
Brest	6.1	5.8	7.8	9.2	11.6	14.4	15.6	16	14.7	12	9	7	48.24	-4.29
Clermont	2.6	3.7	7.5	10.3	13.8	17.3	19.4	19.1	16.2	11.2	6.6	3.6	45.47	3.05
Grenoble	1.5	3.2	7.7	10.6	14.5	17.8	20.1	19.5	16.7	11.4	6.5	2.3	45.1	5.43
Lille	2.4	2.9	6	8.9	12.4	15.3	17.1	17.1	14.7	10.4	6.1	3.5	50.38	3.04
Lyon	2.1	3.3	7.7	10.9	14.9	18.5	20.7	20.1	16.9	11.4	6.7	3.1	45.45	4.51
Marseille	5.5	6.6	10	13	16.8	20.8	23.3	22.8	19.9	15	10.2	6.9	43.18	5.24
Montpellier	5.6	6.7	9.9	12.8	16.2	20.1	22.7	22.3	19.3	14.6	10	6.5	43.36	3.53
Nantes	5	5.3	8.4	10.8	13.9	17.2	18.8	18.6	16.4	12.2	8.2	5.5	47.13	-1.33
Nice	7.5	8.5	10.8	13.3	16.7	20.1	22.7	22.5	20.3	16	11.5	8.2	43.42	7.15
Paris	3.4	4.1	7.6	10.7	14.3	17.5	19.1	18.7	16	11.4	7.1	4.3	48.52	2.2
Rennes	4.8	5.3	7.9	10.1	13.1	16.2	17.9	17.8	15.7	11.6	7.8	5.4	48.05	-1.41
Strasbourg	0.4	1.5	5.6	9.8	14	17.2	19	18.3	15.1	9.5	4.9	1.3	48.35	7.45
Toulouse	4.7	5.6	9.2	11.6	14.9	18.7	20.9	20.9	18.3	13.3	8.6	5.5	43.36	1.26
Vichy	2.4	3.4	7.1	9.9	13.6	17.1	19.3	18.8	16	11	6.6	3.4	46.08	3.26

Figure 1.1: Températures moyennes par villes avec 2 variables supplémentaires

1.1 Analyse en Composantes Principales (ACP)

1.1.1 Principe

Pour l'ACP, on commence par placer les données d'entrée disponibles dans une matrice $\mathbf{X}$. On suppose que l'on dispose de n vecteurs d'entrées $\mathbf{x}_i \in \mathcal{X} = \mathbb{R}^p$, $i \in 1 \dots n$. Pour suivre la terminologie usuelle en analyse de données, on dira que $\mathbf{x}_i$ est vu comme un individu d'indice i c'est-à-dire une collection de p variables (ou attributs numériques). L'objectif est de représenter chaque individu avec un nombre réduit q d'attributs ou de variables. Si $q \ll p$, on cherche une représentation «compacte» des vecteurs donnés en entrée d'un prédicteur.

Grâce à cette réduction du nombre de variables d'entrée, on pourra simplifier le prédicteur mais aussi revenir à une dimension permettant la *visualisation* des données si $q = 2$ ou 3. Cela permet d'analyser graphiquement la structure des données, d'observer d'éventuels regroupements.

On souhaite donc approcher des individus $\mathbf{x}_i \in \mathcal{X}$ de grande taille dans un espace de dimension réduite en les déformant le moins possible. Si chaque individu est rangé en ligne dans $\mathbf{X}$, on a au départ $\mathbf{X} \in \mathcal{M}_{\mathbb{R}}(n,p)$ (noté abusivement $\mathbb{R}^{n\times p}$ plus loin)

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_i^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix} = \begin{pmatrix} & & \vdots & & \\ & & \vdots & & \\ \cdots & \cdots & x_{ij} & \cdots & \cdots \\ & & \vdots & & \\ & & \vdots & & \end{pmatrix}$$

On a :

- ligne $\mathbf{x}_i^T \in \mathbb{R}^p \rightarrow$ valeurs des p variables de l'individu i
- colonne $\mathbf{v}_j \in \mathbb{R}^n \rightarrow$ valeurs de la variable j prises pour les n individus

On dira qu'un individu est un vecteur $\mathbf{x}_i$ de l'e.v. $\mathbb{R}^p$ appelé *espace des individus*. Dans $\mathbb{R}^p$, on étudie les distances entre individus. De manière complémentaire, on dira qu'une variable est un vecteur $\mathbf{v}_j$ de l'e.v. $\mathbb{R}^n$ appelé *espace des variables*. Dans $\mathbb{R}^n$, on étudie les angles entre variables.

1.1.2 Caractéristique d'ordre 1 : tendance centrale des variables

On note $\bar{\mathbf{x}} \in \mathbb{R}^p$, le vecteur des **moyennes arithmétiques des variables**. Ce vecteur $\bar{\mathbf{x}}$ est aussi appelé **individu moyen** (centre de gravité du nuage de points).

Il s'écrit matriciellement

$$\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i = \frac{1}{n}\underbrace{\begin{bmatrix} \mathbf{x}_1 & | \cdots | & \mathbf{x}_n \end{bmatrix}}_{p\times n} \underbrace{\begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}}_{n\times 1} = \frac{1}{n}\mathbf{X}^\top \mathbf{1}_n$$

Centrage des données : application d'une translation dans $\mathbb{R}^p$ telle que l'individu moyen $\bar{\mathbf{x}}$ soit à l'origine.

- Individu centré : $\mathbf{x}_i^c = \mathbf{x}_i - \bar{\mathbf{x}}$
- Tableau de données centré : $\mathbf{X}^c = \mathbf{X} - \mathbf{1}_n \bar{\mathbf{x}}^\top$

1.1.3 Caractéristiques d'ordre 2 : dispersion et dépendance des variables

Matrice de covariance « empirique » $\mathbf{\Sigma} \in \mathbb{R}^{p \times p}$, définie matriciellement[1] par

$$\mathbf{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top = \frac{1}{n}(\mathbf{X}^c)^\top \mathbf{X}^c = \frac{1}{n}\mathbf{X}^\top\mathbf{X} - \bar{\mathbf{x}}\bar{\mathbf{x}}^\top$$

La matrice de covariance mesure :

- la dispersion des p variables autour de leurs moyennes arithmétiques :
  $\rightarrow$ la diagonale

$$\begin{bmatrix} \ddots & & \\ & \sigma_{jj} = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2 & \\ & & \ddots \end{bmatrix}$$

  contient les p « **variances empiriques** » des variables notées σ_j^2 ;
- les dépendances linéaires entre deux variables $\mathbf{v}_i$ et $\mathbf{v}_k$, $i \neq k$
  $\rightarrow$ les éléments hors-diagonale, notés σ_{ik}, contiennent les « **covariances empiriques** » entre les variables $\mathbf{v}_i$ et $\mathbf{v}_k$.

L'**inertie du nuage de points**, c.-à-d. la dispersion de ceux-ci autour du centre de gravité $\bar{\mathbf{x}}$, mesurée par la moyenne des carrés des distances des individus à $\bar{\mathbf{x}}$, est donnée par

$$\mathcal{I}(\mathbf{X}) = trace(\mathbf{\Sigma}) = \frac{1}{n}\sum_{i=1}^{n}\|\mathbf{x}_i - \bar{\mathbf{x}}\|^2$$

Elle est aussi appelée **variance totale**, sachant que

$$trace(\mathbf{\Sigma}) = \sum_{j=1}^{p}\sigma_j^2$$

[1] On montrera, à titre d'exercice que $(\mathbf{X}^c)^\top\mathbf{X}^c = \mathbf{X}^\top\mathbf{X} - n\bar{\mathbf{x}}\bar{\mathbf{x}}^\top$.

$$\begin{aligned}(\mathbf{X}^c)^\top\mathbf{X}^c &= (\mathbf{X}^\top - \bar{\mathbf{x}}\mathbf{1}_n^\top)(\mathbf{X} - \mathbf{1}_n\bar{\mathbf{x}}^\top)\\ &= \mathbf{X}^\top\mathbf{X} - \underbrace{(\mathbf{X}^\top\mathbf{1}_n)}_{=n\bar{\mathbf{x}}}\bar{\mathbf{x}}^\top - \bar{\mathbf{x}}\underbrace{(\mathbf{1}_n^\top\mathbf{X})}_{=n\bar{\mathbf{x}}^\top} + \bar{\mathbf{x}}\underbrace{(\mathbf{1}_n^\top\mathbf{1}_n)}_{=n}\bar{\mathbf{x}}^\top\\ &= \mathbf{X}^\top\mathbf{X} - n\bar{\mathbf{x}}\bar{\mathbf{x}}^\top - n\bar{\mathbf{x}}\bar{\mathbf{x}}^\top + n\bar{\mathbf{x}}\bar{\mathbf{x}}^\top = \mathbf{X}^\top\mathbf{X} - n\bar{\mathbf{x}}\bar{\mathbf{x}}^\top\end{aligned}$$

1.1.4 Corrélation entre variables

Il faut réduire (normaliser) les données pour analyser les éventuelles corrélations entre variables. On appelle matrice diagonale (p, p) de réduction, la matrice :

$$\mathbf{D}_{1/\sigma} = (\operatorname{diag}\sigma)^{-1} = \begin{pmatrix} \frac{1}{\sigma_1} & & & & \\ & \ddots & & & \\ & & \frac{1}{\sigma_i} & & \\ & & & \ddots & \\ & & & & \frac{1}{\sigma_p} \end{pmatrix}$$

On peut alors réduire le tableau de données centrées :

$$\begin{aligned} \mathbf{X}_{\mathbf{0,1}} &= \begin{pmatrix} & & \frac{x_{1j}-\bar{x}_j}{\sigma_j} & & \\ & & \vdots & & \\ \cdots & \cdots & \frac{x_{ij}-\bar{x}_j}{\sigma_j} & \cdots & \cdots \\ & & \vdots & & \\ & & \frac{x_{nj}-\bar{x}_j}{\sigma_j} & & \end{pmatrix} \\ &= \mathbf{X}^c\mathbf{D}_{1/\sigma} \end{aligned}$$

On appelle finalement matrice de corrélation, la matrice suivante :

$$\begin{aligned} \mathbf{R} &= \begin{pmatrix} & \vdots & \\ \cdots & \frac{1}{n}\sum_{k=1}^{n} \frac{(x_{ki}-\bar{x}_i)}{\sigma_i}\frac{\left(x_{kj}-\bar{x}_j\right)}{\sigma_j} & \cdots \\ & \vdots & \end{pmatrix} \\ &= \frac{1}{n}\mathsf{X}_{\mathbf{0,1}}^{\top}\mathsf{X}_{\mathbf{0,1}} \end{aligned}$$

On peut, grâce à cette matrice, identifier empiriquement une dépendance affine entre les variables $\mathbf{v}_i$ et $\mathbf{v}_j$: si $R_{ij} = 1$ alors $\mathbf{v}_j = a\mathbf{v}_i + b$ (avec $a > 0$, corrélation positive).

Une corrélation positive $R_{ij} \approx 1$ (sachant $|R_{ij}| \leq 1$) signifie donc une forte dépendance entre les deux variables de sorte que si $\mathbf{v}_i$ augmente, il en va statistiquement de même pour $\mathbf{v}_j$.

1.1.5 Analyse en $q = 1$ Composante Principale

L'ACP est une méthode *factorielle linéair*e pour analyser la structure des données. Elle permet d'obtenir une représentation approchée du nuage de points dans un sous-espace de dimension faible $q \ll p$ de $\mathbb{R}^p$, qui conserve le « maximum » d'information et qui soit la plus « proche » possible du nuage initial.

Le problème $\mathscr{P}_{q=1}$

Formulation 1 *Il s'agit de chercher une droite D de $\mathbb{R}^p$, passant par le centre de gravité $\bar{\mathbf{x}}$ et de vecteur directeur $\mathbf{u} \in \mathbb{R}^p$, qui maximise l'inertie du nuage des points projetés sur D, c.-à-d. solution de*

$$\max_{\mathbf{u}\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n} \|\mathbf{y}_i - \bar{\mathbf{y}}\|^2 \tag{1.1}$$

où $\mathbf{y}_i$ dénote la projection de $\mathbf{x}_i$ sur D et $\bar{\mathbf{y}}$ dénote le centre de gravité des points projetés.

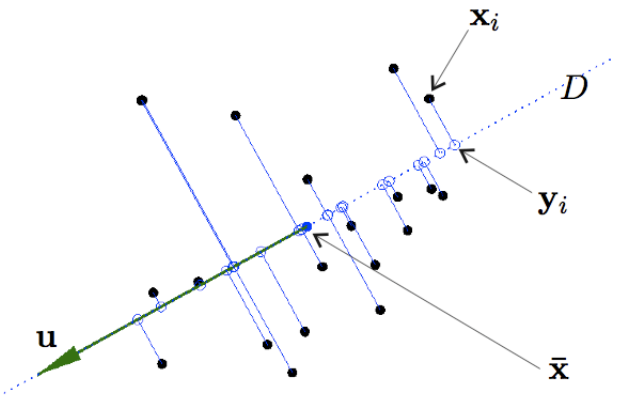

Figure 1.2: La droite recherchée est telle que la moyenne des carrés des distances entre projections sur D, c.-à-d. telle que l'inertie est la plus grande possible : la projection du nuage doit être la plus étalée possible.

> Dit différemment : « l'ACP cherche à remplacer les p variables d'un individu par une nouvelle —appelée (première) composante principale— qui soit de variance maximale et obtenue à partir d'une combinaison linéaires des des variables initiales ».

Un **résultat important** est que maximiser l'inertie revient à minimiser l'erreur d'approximation, c.-à-d. la somme des carrés des distances entre points originaux et points projetés.

$$\arg\max_{\mathbf{u}\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n} \|\mathbf{y}_i - \bar{\mathbf{y}}\|^2 = \arg\min_{\mathbf{u}\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n} \|\mathbf{y}_i - \mathbf{x}_i\|^2$$

Du fait de cette équivalente, l'approximation du nuage de points est la « meilleure » possible.

Projection orthogonale sur une droite

On rappelle que si $\mathbf{u} \in \mathbb{R}^p$ représente un vecteur directeur unitaire ($\|\mathbf{u}\| = 1$) de la droite D, alors la projection orthogonale de $\mathbf{x}_i$ sur D s'écrit

$$\mathbf{y}_i = c_i\mathbf{u} + \bar{\mathbf{x}} \quad \text{où} \quad c_i = \mathbf{u}^\top(\mathbf{x}_i - \bar{\mathbf{x}})$$

La matrice $\mathbf{\Pi} = \mathbf{u}\mathbf{u}^\top$ est un projecteur sur la droite vectorielle parallèle à D ; il vérifie $\mathbf{\Pi} = \mathbf{\Pi}^2$

Dans $\mathbb{R}^p$, l'**approximation du tableau de données** projetées sur D est donc donnée par la matrice $n \times q$ de rang 1

$$\mathbf{Y} = (\mathbf{X} - \mathbf{1_n}\bar{\mathbf{x}}^\top)\mathbf{u}\mathbf{u}^\top + \mathbf{1_n}\bar{\mathbf{x}}^\top$$

dont la ligne i correspond à la projection sur D de l'individu i

$$\mathbf{Y} \overset{\text{déf}}{=} \begin{bmatrix} \vdots \\ \mathbf{y}_i^\top \\ \vdots \end{bmatrix}$$

Définition 1 *On appelle* ***Composante Principale****, la coordonnée* c_i *du point projeté* $\mathbf{y}_i$*, relativement au repère* $\{\bar{\mathbf{x}}; \mathbf{u}\}$*. La droite* D *est appelée* ***Axe Principal*** *et le vecteur directeur* $\mathbf{u}$ *est appelé* ***Vecteur Principal****.*

Résolution de $\mathscr{P}_{q=1}$

On montre que le problème de l'analyse en $q = 1$ Composante Principale est strictement équivalent au problème suivant :

Formulation 2 *Chercher* $\mathbf{u} \in \mathbb{R}^p$ *solution de*

$$\max_{\mathbf{u}\in\mathbb{R}^p} \mathbf{u}^\top\mathbf{\Sigma}\mathbf{u} \quad \textit{sous la contrainte} \quad \|\mathbf{u}\|^2 = 1 \tag{1.2}$$

Pour montrer l'équivalence des deux formulations 1 et 2 précédentes,

- on montre facilement[2] que $\bar{\mathbf{y}} = \bar{\mathbf{x}}$.
- Ainsi le carré de la distance entre $\mathbf{y}_i$ et $\bar{\mathbf{y}}$ s'écrit :

$$\|\mathbf{y}_i - \bar{\mathbf{y}}\|^2 = \left\|\mathbf{u}\mathbf{u}^\top(\mathbf{x}_i - \bar{\mathbf{x}})\right\|^2 = (\mathbf{x}_i - \bar{\mathbf{x}})^\top\mathbf{u}\mathbf{u}^\top\mathbf{u}\mathbf{u}^\top(\mathbf{x}_i - \bar{\mathbf{x}}) = (\mathbf{x}_i - \bar{\mathbf{x}})^\top\mathbf{u}\mathbf{u}^\top(\mathbf{x}_i - \bar{\mathbf{x}})$$

$$= (\mathbf{u}^\top(\mathbf{x}_i - \bar{\mathbf{x}}))^2$$

[2] $$\bar{\mathbf{y}} = \frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{u}\mathbf{u}^\top(\mathbf{x}_i - \bar{\mathbf{x}}) + \bar{\mathbf{x}}\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{u}\mathbf{u}^\top\mathbf{x}_i - \frac{1}{n}\sum_{i=1}^{n}\mathbf{u}\mathbf{u}^\top\bar{\mathbf{x}} + \frac{1}{n}\sum_{i=1}^{n}\bar{\mathbf{x}} = \mathbf{u}\mathbf{u}^\top\bar{\mathbf{x}} - \mathbf{u}\mathbf{u}^\top\bar{\mathbf{x}} + \bar{\mathbf{x}} = \bar{\mathbf{x}}$$

- Il nous suffit de montrer que $\frac{1}{n}\sum_{i=1}^{n}\|\mathbf{y}_i-\bar{\mathbf{y}}\|^2=\mathbf{u}^\top\mathbf{\Sigma}\mathbf{u}$ pour faire le lien entre (1.1) et (1.2):

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\|\mathbf{y}_i-\bar{\mathbf{y}}\|^2 &= \frac{1}{n}\sum_{i=1}^{n}(\mathbf{u}^\top(\mathbf{x}_i-\bar{\mathbf{x}}))^2\\ &= \frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{u}^\top(\mathbf{x}_i-\bar{\mathbf{x}})\right)\left((\mathbf{x}_i-\bar{\mathbf{x}})^\top\mathbf{u}\right)\\ &= \frac{1}{n}\mathbf{u}^\top\underbrace{\left(\sum_{i=1}^{n}(\mathbf{x}_i-\bar{\mathbf{x}})(\mathbf{x}_i-\bar{\mathbf{x}})^\top\right)}_{n\mathbf{\Sigma}}\mathbf{u}=\mathbf{u}^\top\mathbf{\Sigma}\mathbf{u}\end{aligned}$$

■

Le problème d'optimisation sous contrainte $\mathscr{P}_{q=1}$ formulé en équation (1.2) peut être reformulé comme le problème non-contraint

$$\max_{\mathbf{u}\in\mathbb{R}^p}\mathscr{L}(\mathbf{u})$$

en introduisant le Lagrangien

$$\mathscr{L}(\mathbf{u})=\mathbf{u}^\top\mathbf{\Sigma}\mathbf{u}+\lambda(1-\|\mathbf{u}\|^2)$$

Puisque

$$\frac{\partial\mathscr{L}}{\partial\mathbf{u}}=2\mathbf{\Sigma}\mathbf{u}-2\lambda\mathbf{u}.$$

La condition nécessaire d'optimalité $\frac{\partial\mathscr{L}}{\partial\mathbf{u}}=\mathbf{0}$ est exprimée par l'équation aux valeurs propres suivante :

$$\mathbf{\Sigma}\mathbf{u}=\lambda\mathbf{u}. \tag{1.3}$$

La solution exacte de (1.3) est un couple $(\lambda,\mathbf{u})$, correspondant à une valeur propre et un vecteur propre de $\mathbf{\Sigma}$.

Comment choisir ce couple ?

- Comme $\mathbf{\Sigma}$ est symétrique et semi-définie positive, ses valeurs propres sont réelles et positives (ou nulles).
- Puisque $\mathbf{u}^\top\mathbf{\Sigma}\mathbf{u}=\lambda\mathbf{u}^\top\mathbf{u}$, cf. (1.3), maximiser $\mathbf{u}^\top\mathbf{\Sigma}\mathbf{u}$ sous la contrainte $\|\mathbf{u}\|^2=1$, revient à maximiser λ !

Théorème 1 (solution de $\mathscr{P}_{q=1}$) *La solution du problème $\mathscr{P}_{q=1}$ pour $\mathbf{u}_1$ est donnée par le vecteur propre associé à la plus grande valeur propre de $\mathbf{\Sigma}$.*

1.1.6 Cas général : analyse en $q > 1$ composantes principales

Le problème $\mathscr{P}_q$

Formulation 3 *Chercher un sous-espace affine S_q de dimension $q < p$ de $\mathbb{R}^p$, passant par le centre de gravité $\bar{\mathbf{x}}$ et de direction le sous-espace vectoriel $F_q \subset \mathbb{R}^p$, qui maximise l'inertie du nuage des points projetés sur S_q, c.-à-d. solution de*

$$\max_{\substack{F \subset \mathbb{R}^p \\ \dim F = q}} \frac{1}{n} \sum_{i=1}^{n} \|\mathbf{y}_i - \bar{\mathbf{y}}\|^2$$

où $\mathbf{y}_i$ dénote la projection de $\mathbf{x}_i$ sur S_q et $\bar{\mathbf{y}}$ dénote le centre de gravité des points projetés.

On rappelle tout d'abord qu'un sous-espace vectoriel de dimension q peut être engendré par une base formée de q vecteurs indépendants, notés $\{\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_q\}$, qui forment la matrice

$$\mathbf{U}_q = \begin{bmatrix} \mathbf{u}_1 & | \cdots | & \mathbf{u}_q \end{bmatrix}. \tag{1.4}$$

Si la base $\{\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_q\}$ est orthonormée, alors $\mathbf{U}_q$ est orthogonale c.-à-d.

$$\mathbf{U}_q^\top \mathbf{U}_q = \mathbf{I}.$$

Donc, rechercher F_q revient à rechercher une base $\{\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_q\}$ orthonormée qui l'engendre.

Projection orthogonale sur un sous-espace affine

Si $\{\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_q\} \subseteq \mathbb{R}^p$ est une base orthonormée engendrant F_q, alors la projection orthogonale de $\mathbf{x}_i$ sur S_q de direction F_q s'écrit

$$\mathbf{y}_i = \mathbf{U}_q \mathbf{c}_i + \bar{\mathbf{x}} \quad \text{où} \quad \mathbf{c}_i = \mathbf{U}_q^\top (\mathbf{x}_i - \bar{\mathbf{x}}) \tag{1.5}$$

où $\mathbf{U}_q$ est défini en (1.4). La matrice $\mathbf{\Pi} = \mathbf{U}_q \mathbf{U}_q^\top$ est un projecteur sur F_q (direction de S_q) ; elle vérifie $\mathbf{\Pi} = \mathbf{\Pi}^2$.

Définition 2 *On appelle **Composantes Principales** les coordonnées $(c_1, ..., c_q)$ du point projeté $\mathbf{y}_i$, relativement au repère orthornormé $\{\bar{\mathbf{x}}; \mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_q\}$.*

On note $\mathbf{c}_i$ le vecteur des Composantes Principales de l'individu i. L'ensemble des q Composantes Principales sera donné par la matrice notée

$$\mathbf{C}_q = (\mathbf{X} - \mathbf{1}_n \bar{\mathbf{x}}^\top)\mathbf{U}_q = \mathbf{X}^c \mathbf{U}_q$$

sachant que la ligne i correspondra aux Composantes Principales de l'individu i

$$\mathbf{C}_q = \underset{n \times q}{\begin{bmatrix} \vdots & \\ \cdots & c_{ij} \quad \cdots \\ \vdots & \end{bmatrix}} = \begin{bmatrix} \vdots \\ \mathbf{c}_i^\top \\ \vdots \end{bmatrix}$$

Dans $\mathbb{R}^p$, l'**approximation du tableau de données** projetées sur S_q, appelée **reconstruction**, est donnée par la matrice $n \times q$ de rang q

$$\mathbf{Y} = (\mathbf{X} - \mathbf{1_n}\bar{\mathbf{x}}^\top)\mathbf{U}_q\mathbf{U}_q^\top + \mathbf{1}_n\bar{\mathbf{x}}^\top = \mathbf{X^c}\mathbf{U}_q\mathbf{U}_q^\top + \mathbf{1}_n\bar{\mathbf{x}}^\top \tag{1.6}$$

dont la ligne i correspond à la projection sur D de l'individu i

$$\mathbf{Y} \overset{\text{déf}}{=} \begin{bmatrix} \vdots \\ \mathbf{y}_i^\top \\ \vdots \end{bmatrix}$$

Un **résultat important** est que l'approximation du nuage de points est la « meilleure » possible. En effet, maximiser l'inertie revient à minimiser l'erreur d'approximation, c.-à-d. à minimiser la somme des carrés des distances entre points et points projetés (voir le théorème d'Eckart–Young en §1.1.7)

$$\arg \max_{\{\mathbf{u}_1,\mathbf{u}_2,\ldots,\mathbf{u}_q\}} \frac{1}{n}\sum_{i=1}^{n} \|\mathbf{y}_i - \bar{\mathbf{y}}\|^2 = \arg \min_{\{\mathbf{u}_1,\mathbf{u}_2,\ldots,\mathbf{u}_q\}} \frac{1}{n}\sum_{i=1}^{n} \|\mathbf{y}_i - \mathbf{x}_i\|^2$$

Résolution du problème $\mathscr{P}_q$

On supposera ici, sans perte de généralité, que le centre de gravité du nuage de points coïncide avec $\mathbf{0} \in \mathbb{R}^p$. Pour se ramener à ce cas, on effectuera au préalable un centrage des données de telle façon que :

$$\mathbf{x}_i^c = \mathbf{x}_i \quad \text{et} \quad \mathbf{X}^c = \mathbf{X}$$

Théorème 2 *Soit $F_{q-1} \subset \mathbb{R}^p$ un sous-espace vectoriel de dimension $q-1$, solution de $\mathscr{P}_{q-1}$. Alors le sous-espace vectoriel F_q de dimension q, solution de $\mathscr{P}_q$, est la somme directe[a] de F_{q-1} et de la droite vectorielle F_1 de $\mathbb{R}^p$, orthogonale à F_{q-1}, solution de $\mathscr{P}_1$: les solutions sont « emboîtées ».*

[a] La somme directe de deux sev F et G est définie par $F \oplus G = \{f + g \mid f \in F, \quad g \in G\}$

Ce que cela veut dire : pour obtenir F_q, on procède de proche en proche, on cherche d'abord le sous-espace vectoriel de dimension 1 maximisant l'inertie du nuage projeté, puis le sous-espace vectoriel de dimension 1 orthogonal au précédent et maximisant l'inertie, etc.

Théorème 3 *Le sous-espace vectoriel F_q de dimension q, solution de $\mathscr{P}_q$, est engendré par les q vecteurs propres de Σ associés aux q plus grandes valeurs propres.*

On rappelle que les vecteurs propres de $\mathbf{\Sigma}$ sont orthogonaux, sans perte de généralité, on les supposera unitaires. On peut démontrer ces théorèmes du fait que le problème de l'analyse en q Composantes Principales est strictement équivalent au problème $\mathscr{P}_q$ suivant :

résoudre

$$\max_{\mathbf{U}_q \in R^{p\times q}} trace(\mathbf{U}_q^\top \mathbf{\Sigma} \mathbf{U}_q) \quad \text{sous la contrainte} \quad rang\mathbf{U}_q = q$$

Les vecteurs de base recherchés sont alors les colonnes de la solution $\mathbf{U}_q$, cf. (1.4).

1.1.7 Reconstruction du tableau des données au moyen des composantes principales et des facteurs principaux

Bien que l'objectif soit en général de n'utiliser qu'un petit nombre de Composantes Principales, l'ACP en construit initialement p, autant que de variables originales.

Reconstruction de l'individu i

Voir l'équation (1.5).

Reconstruction du tableau des données

Voir l'équation (1.6).

- Si $q = p$: reconstitution/reconstruction exacte ;
- si $q < p$: meilleure approximation de $\mathbf{X}$ par une matrice de rang q au sens des moindres carrés (théorème d'Eckart–Young) :

$$\mathbf{Y} = \arg\min_{\mathbf{M}} \|\mathbf{M} - \mathbf{X}\|_F \quad \text{sous la contrainte} \quad rang(\mathbf{M}) = q$$

  Aucune autre matrice de rang q ne peut rendre l'erreur d'approximation plus faible.

1.1.8 Conclusion et propriétés de l'ACP

Statistiquement parlant, l'ACP a pour objectif de décrire le nuage de points par q nouvelles variables (les Composantes Principales) qui soient une combinaison linéaire des variables initiales et dont la variance totale est maximale.

- **Nombre de CP :**

  « Retenir q Composantes Principales »
  veut dire
  « Remplacer les observations originales par leur projections orthogonales dans le sous-espace à q dimensions défini par les q premières Composantes Principales. »

- **Orthogonalité :**

  Les Composantes Principales sont associées à des directions de l'espace des observations qui sont deux à deux orthogonales. Autrement dit, l'ACP procède à un changement de repère orthogonal, les axes originaux étant remplacées par les Axes Principaux.

- **Décorrélation :**

  D'une part, on a le schéma suivant, permettant de calculer la matrice de covariance $\mathbf{\Sigma}'$ associées aux Composantes Principales

$$\begin{array}{ccc} \{\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_p\} & \overset{\text{changement de base vectorielle}}{\longrightarrow} & \{\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_p\} \\ \mathbf{X}^c & \longrightarrow & \mathbf{C}^c = \mathbf{X}^c\mathbf{U} \\ \mathbf{\Sigma} = (\mathbf{X}^c)^\top \mathbf{X}^c & \longrightarrow & \mathbf{\Sigma}' = (\mathbf{C}^c)^\top \mathbf{C}^c = \mathbf{U}^\top \mathbf{\Sigma} \mathbf{U} \end{array}$$

D'autre part,

$$\begin{aligned} \mathbf{\Sigma}\mathbf{u}_j &= \lambda_j \mathbf{u}_j \\ &\Leftrightarrow \mathbf{\Sigma}\mathbf{U} = \mathbf{U} \begin{bmatrix} \ddots & & \\ & \lambda_j & \\ & & \ddots \end{bmatrix} \Leftrightarrow \mathbf{U}^\top \mathbf{\Sigma} \mathbf{U} = \begin{bmatrix} \ddots & & \\ & \lambda_j & \\ & & \ddots \end{bmatrix} \end{aligned}$$

Les Composantes Principales sont des variables qui s'avèrent être deux à deux décorrélées.

- **Ordre d'importance - contraste :**

  La propriété fondamentale des Composantes Principales est de pouvoir être classées par ordre décroissant d'importance : le meilleur sous-espace à q dimensions dans lequel projeter les données est celui engendré par les q premières Composantes Principales.

  La variance totale des Composantes Principales (c.-à-d. des nouvelles variables) correspond à la somme des valeurs propres

$$\sum_{j=1}^{q} \sigma_j^2 = trace(\mathbf{\Sigma}) = \sum_{j=1}^{q} \lambda_j$$

  Chaque valeur propre mesure la part de variance expliquée par l'Axe Principal correspondant. Le contraste conservé par q Composantes Principales est

$$\frac{\sum_{j=1}^{q} \lambda_j}{\sum_{j=1}^{p} \lambda_j}$$