Imputation des données manquantes

1. Imputation simple : remplacer la valeur manquante par une seule valeur "plausible"

   - imputation par moyenne (quantitative) ou mode (qualitative)
   - imputation par régression = remplacer la valeur manquante par une valeur prédite par un modèle de régression. (simple, en général)
   - imputation à l'aide des analyses multivariées (ACP, ACM)
     idée = utiliser la formule de reconstitution des données à l'aide de l'ACP.

Dans la suite on va s'intéresser à cette dernière méthode qui sera appellée NIPALS (Algorithme NIPALS pour données manquantes).

NIPALS = Nonlinear Iterative PArtial Least Squares.

L'algorithme NIPALS

Observation importante : dans le cas des données sans valeur manquante

NIPALS = formule de reconstitution des données en ACP.

NIPALS avec données complètes :

Soit $X = \begin{pmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1p} \\ x_{21} & \cdots & x_{2j} & \cdots & x_{2p} \\ \vdots & & & & \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{ip} \\ \vdots & & & & \\ x_{n1} & \cdots & x_{nj} & \cdots & x_{np} \end{pmatrix}$

$X_1, X_2, \ldots, X_p$ = variables quantitatives

- il est important de bien savoir
  - la régression linéaire
  - l'analyse en composantes principales

On rappelle ces deux aspects dans les pages suivantes.

Rappel 1 : Régression linéaire simple

On dispose de deux variables X et Y observées sur n individus :

	X	Y
1	x_1	y_1
2	x_2	y_2
⋮	⋮	⋮
i	x_i	y_i
⋮	⋮	⋮
n	x_n	y_n

L'équation de régression linéaire est

$$Y = \beta_0 + \beta_1 X + \varepsilon$$ et sous les hypothèses classiques de la régression on obtient les estimateurs :

$$\begin{cases} \hat{\beta}_0 = \overline{Y} - \beta_1 \overline{X} \\ \hat{\beta}_1 = \dfrac{\frac{1}{n}\sum x_i y_i - \bar{x}\,\bar{y}}{s_x^2} \end{cases}$$

avec $\bar{x} = \frac{1}{n}\sum x_i$, $\bar{y} = \frac{1}{n}\sum y_i$

$$s_x^2 = \frac{1}{n}\sum (x_i - \bar{x})^2$$

Observons que si les variables sont

centrées ($\bar{x}=0$, $\bar{y}=0$) et réduites ($s_x^2=1$, $s_y^2=1$) alors :

$$\hat{\beta}_0 = 0 \quad \text{et} \quad \hat{\beta}_1 = \frac{1}{n}\sum x_i y_i$$

et le modèle s'écrit donc :

$$Y = \left(\frac{1}{n}\sum x_i y_i\right) \cdot X + \varepsilon, \text{ ou, en forme matricielle}$$

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \left(\frac{1}{n}\sum x_i y_i\right) \cdot \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

$$\begin{pmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{pmatrix} = \hat{b}_1 \cdot \begin{pmatrix} x_1 \\ \vdots \\ x_i \\ \vdots \\ x_n \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_i \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Notation : $\sum x_i y_i = \langle x, y \rangle$ ← produit scalaire

Remarque importante : régression linéaire multiple avec variables indépendantes (ou non-corrélées).

On a :

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{pmatrix} \quad \text{et} \quad X = \begin{bmatrix} x_1 & x_2 & & x_p \\ \downarrow & \downarrow & & \downarrow \\ x_{11} & x_{1j} & \cdots & x_{1p} \\ x_{21} & x_{2j} & \cdots & x_{2p} \\ \vdots & & & \\ x_{i1} & x_{ij} & \cdots & x_{ip} \\ \vdots & & & \\ x_{n1} & x_{nj} & \cdots & x_{np} \end{bmatrix}$$

On suppose que :

- $X_1, X_2, \dots X_p$ sont indépendantes et
- de plus centrées et réduites.

Alors, dans ce cas, la matrice de var-cov de X

$$V = \mathrm{Cov}(X) = R(X) = \frac{1}{n} X^T X = I_p = \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix}_{p \times p}$$

Dans le modèle linéaire

$$y = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{pmatrix} = (X^T X)^{-1} X^T y$$

ou encore :

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_j \\ \vdots \\ \hat{\beta}_p \end{pmatrix} = \underbrace{\begin{pmatrix} \frac{1}{n} & & & 0 \\ & \frac{1}{n} & & \\ & & \ddots & \\ 0 & & & \frac{1}{n} \end{pmatrix}}_{I_n : (n \times n)} \cdot \underbrace{\begin{pmatrix} x_{11} & x_{21} & \cdots & x_{n1} \\ x_{12} & x_{22} & \cdots & x_{n2} \\ \vdots & & & \\ 1 & & & \\ x_{p1} & x_{p2} & \cdots & x_{pn} \end{pmatrix}}_{X^T \ (p \times n)} \cdot \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

On obtient que

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_j \\ \vdots \\ \hat{\beta}_p \end{pmatrix} = \begin{pmatrix} \frac{1}{n} \sum x_{i1} \cdot y_i = \frac{1}{n} \langle x_1, Y \rangle \\ \frac{1}{n} \sum x_{i2} \cdot y_i = \frac{1}{n} \langle x_2, Y \rangle \\ \vdots \\ \frac{1}{n} \sum x_{ip} y_i = \frac{1}{n} \langle x_p, Y \rangle \end{pmatrix}$$

La leçon à tirer est que les coefficients des variables X_j, les $\hat{\beta}_j$, sont en fait les mêmes comme si on avait fait juste la régression linéaire simple entre Y et X_j.

Pour résumer en "deux" mots : la régression linéaire multiple avec p variables indépendantes se réduit à p régressions simples $(Y \sim X_j \ ; j = 1 \cdots p)$

Fin Rappel 1.

Rappel 2 : Analyse en composantes principales (ACP)

Construire des composantes qui resume "au mieux" l'information dans les données.

composante (linéaire) :

$$c = u_1 X_1 + u_2 X_2 + \dots + u_p X_p \quad \text{avec}$$

$u = \begin{pmatrix} u_1 \\ \vdots \\ u_p \end{pmatrix}$ des poids définissant la composante c.

Les poids sont normalisés

$$\sum_{i=1}^{p} u_i^2 = 1 \quad \text{car le critère}$$

pour déterminer c est celui de variance (information maximale).

Donc, ACP consiste à chercher c t.q.

$Var(c)$ soit maximale

parmi toutes les combinaisons linéaires possibles de $X_1, \dots, X_p$.

Solution : $Vu = \lambda u$, u associé à la valeur propre la plus grande de X.

V ↓ matrice de covariance de X.

La matrice de variance-covariance V (qui coincide avec la matrice de corrélation R quand les données sont centrées et réduites) a les propriétés suivantes :

- elle est symétrique et positive définie
- elle admet p valeurs propres positives ou nulles

$$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$$

- à chaque valeur propre on associe un vecteur propre u_i :

$$\lambda_1 \geq \lambda_2 \dots \geq \lambda_p$$
$$\downarrow \qquad \downarrow \qquad \downarrow$$
$$u_1 \in \mathbb{R}^p \quad u_2 \in \mathbb{R}^p \quad u_p \in \mathbb{R}^p$$

tels que :

$$\|u_i\|^2 = \sum_{j=1}^{p} u_j^2 = 1 \quad , \quad \forall\, i = 1 \dots p \quad j = 1 \dots p.$$

(a) $$u_i = \begin{pmatrix} u_{i1} \\ u_{i2} \\ \vdots \\ u_{ip} \end{pmatrix} = \text{le } i\text{ème vecteur propre de } V.$$

et

(b) u_i et u_j sont orthogonaux entre eux pour $i \neq j$

$$\langle u_i, u_j \rangle = \sum_{k=1}^{p} u_{ik} \cdot u_{jk} = 0$$

Le point b) nous indique que

⚠ $\{u_1, u_2 \dots, u_p\}$ forment une base dans $\mathbb{R}^p$.

Du coup, tout élément x de $\mathbb{R}^p$ peut s'écrire dans la base $\{u_1, u_2 \dots u_p\}$.

$$x = \alpha_1 \cdot u_1 + \alpha_2 u_2 + \dots + \alpha_p u_p$$

$$\| \quad \begin{pmatrix} x_1 \\ \vdots \\ x_p \end{pmatrix}$$

avec α_i, $i = 1 \dots p$, les coefficients de développement.

$$\alpha_i = \langle u_i, x \rangle = \sum_{k=1}^{p} u_{ik} \cdot x_k$$

- De manière vectorielle :

$$\begin{pmatrix} x_1 \\ \vdots \\ x_p \end{pmatrix} = \alpha_1 \cdot \begin{pmatrix} u_{11} \\ \vdots \\ u_{1p} \end{pmatrix} + \dots \quad \alpha_p \begin{pmatrix} u_{p1} \\ \vdots \\ u_{pp} \end{pmatrix}$$

u_1 $\qquad$ u_p

\# écriture en colonne.

ou

$$(x_1 \dots x_p) = \alpha_1 \cdot (u_{11} \dots u_{1p}) + \dots + \alpha_p (u_{p1}, \dots, u_{pp})$$

\# écriture en ligne

On définit donc les composantes principales de X, comme étant les variables définis par les vecteurs $\{u_1, u_2, \dots, u_p\}$.

On a donc p composantes principales :

$$u_1 \rightarrow C_1 = u_{11} X_1 + u_{12} X_2 + \dots + u_{1p} X_p$$

$$u_2 \rightarrow C_2 = u_{21} X_1 + u_{22} X_2 + \dots + u_{2p} X_p$$

$$\vdots$$

$$u_p \rightarrow C_p = u_{p1} X_1 + u_{p2} X_2 + \dots + u_{pp} X_p$$

ou matriciellement :

$$C_1 = \begin{pmatrix} c_{1,1} \\ c_{2,1} \\ \vdots \\ c_{n,1} \end{pmatrix} = u_{11} \cdot \begin{pmatrix} x_{11} \\ x_{21} \\ \vdots \\ x_{n1} \end{pmatrix} + u_{12} \cdot \begin{pmatrix} x_{12} \\ x_{22} \\ \vdots \\ x_{n2} \end{pmatrix} + \dots + u_{1p} \begin{pmatrix} x_{1p} \\ x_{2p} \\ \vdots \\ x_{np} \end{pmatrix}$$

$\uparrow X_1$ $\uparrow X_2$ $\uparrow X_p$.

Vous remplacez C_1 par C_i et vous obtenez l'expression de toute composante C_i.

Notez que $C_1 = \begin{pmatrix} c_{11} \\ c_{21} \\ \vdots \\ c_{n1} \end{pmatrix} = \begin{pmatrix} \langle u_1, (x_{11}\ x_{12}\ \dots\ x_{1p}) \rangle \\ \vdots \\ \langle u_1, x_{n1}\ x_{n2}\ \dots\ x_{np} \rangle \end{pmatrix} = \begin{pmatrix} \langle u_1, ind_1 \rangle \\ \vdots \\ \langle u_1, ind_n \rangle \end{pmatrix}$.

Ces composantes principales sont donc des nouvelles variables aléatoires construites à partir des variables $X_1, \dots X_p$ et qui ont les propriétés suivantes :

	X_1	X_2	----	X_p	C_1	C_2	----	C_p
1	X_{11}	X_{12}	----	X_{1p}	C_{11}	C_{12}	....	C_{1p}
2	X_{21}	X_{22}	----	X_{2p}	C_{21}	C_{22}		C_{2p}
.								
.								
n	X_{n1}	X_{n2}		X_{np}	C_{n1}	C_{n2}		C_{np}
	Données d'origine				composants principales			

Propriétés :

- [données centrées et réduites] : $\overline{C_i} = 0$ et $V(C_i) = \lambda_i$

(moyenne nulle et variance = val propre).

⚠ Attention : discussion sur les val. propres nulles $\lambda_i = 0$.

- $\mathrm{cor}(C_i, C_j) = 0 \quad \forall\, i, j \quad i \neq j$

variables non corrélées

Comment les composantes principales résument-elles l'information contenue dans X ?

Information = variance

Alors, l'information totale contenue dans le tableau de données X est donnée par l'inertie totale, qui après petit calcul, donne

$$I_{totale} = V(x_1) + \dots + V(x_p) \quad \left(= p \text{ si } x_i \text{ réduites}\right)$$

⚠

A leur tour, l'information contenue dans la composante c_1 est $Var(c_1) = \lambda_1$.

On dit que la composante c_1 explique $\frac{\lambda_1}{I_{tot}}$ de l'information.

Pareil pour une composante quelconque, c_i, à elle seule, elle explique $\frac{\lambda_i}{I_{totale}}$ de l'information.

Maintenant, si on met ensemble plusieurs composantes, on a le pouvoir explicatif suivant :

- C_2 toute seule : $\dfrac{\lambda_1}{I_{tot}} = \uparrow \dfrac{\lambda_1}{p}$ si ACP normée

Donc C_1 toute seule resume $\dfrac{\lambda_1}{I_{tot}}$ de l'information de X.

suj	X_1 --- X_p
1	x_{11} -- x_{1p}
2	
⋮	
n	x_{n1} x_{np}

$\simeq$ approx

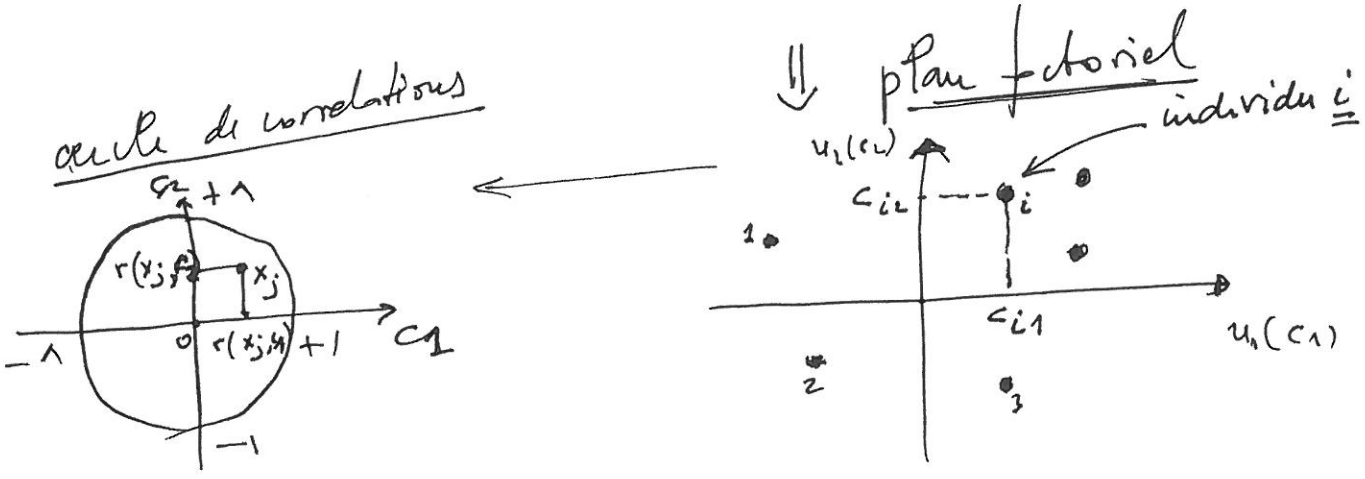
suj	C_1
1	$c_{1,1}$
2	$c_{2,1}$
⋮	
n	$c_{n,1}$

- C_1 et C_2 : $\dfrac{\lambda_1+\lambda_2}{I_{total}} = \uparrow \dfrac{\lambda_1+\lambda_2}{p}$ si ACP normalisée

	X_1 --- X_p
1	x_{11} x_{1p}
2	
i	x_{i1} x_{ip}
⋮	
n	x_{n1} x_{np}

$\simeq$ approx

	C_1	C_2
1	c_{11}	c_{12}
2	c_{21}	c_{22}
⋮		
i	c_{i1}	c_{i2}
⋮		
n	c_{n1}	c_{n2}



Et ainsi de suite.

On garde en général un nombre de composantes qui assurent un bon taux d'information expliquée (90%, par exemple, mais il n'y a pas de règle précise).

- Supposons que $q \leq p$ les valeurs propres de V sont non-nulles : $\lambda_i > 0$, $\forall\, i = 1 \dots q$.

et que $\lambda_{q+j} = 0 \quad \forall\, j \in [\![1, p-q]\!]$.

$$\lambda_1 \geq \lambda_2 \cdots \geq \lambda_q > 0 = \lambda_{q+1} = \cdots \lambda_p.$$

Ce qui est un peu moins connu de l'ACP est la formule de reconstitution des données.

Le fait que $\{u_1, u_2, \dots u_p\}$ forment une base de $\mathbb{R}^p$ cela veut dire que l'individu i

$$ind_i = (x_{i1}, x_{i2}, \dots x_{ip})$$ peut

être exprimé dans cette base, et donc

$$\widehat{ind}_i = \langle ind_i, u_1 \rangle \cdot u_1 + \langle ind_1, u_2 \rangle \cdot u_2 + \cdots \langle ind_1, u_p \rangle u_p$$

$$\widehat{ind}_i = \underbrace{c_{i1}}_{\in \mathbb{R}} \cdot \underbrace{u_1}_{\in \mathbb{R}^p} + c_{i2} \cdot u_2 + \cdots + c_{ip} u_p.$$

Ou encore :

$$(x_{i1}, \dots, x_{ip}) = c_{i1} \cdot (u_{11} \dots u_{1p}) + c_{i2}(u_{21}, \dots, u_{2p}) + \dots + c_{ip}(u_{p1}, \dots, u_{pp})$$

Remarque : évidemment, si $\lambda_j = 0$ alors $c_{ij} = 0$ car $V(c_j) = 0$ et $c_j \equiv 0$ ⚠ pour tous les individus..

[Donc, chaque ligne de la matrice X peut s'écrire comme une combinaison linéaire des vecteurs propres de $\bar{V}$.

(On savoit que chaque composante c_i est une combinaison des variables x_j, $j = 1 \dots p$.)

En mettant la décomposition de $(x_{i1}, \dots x_{ip})$ pour tous les i sous forme matricielle, on obtient :

$$\begin{pmatrix} x_{11} \dots x_{1j} \dots x_{1p} \\ \vdots \\ x_{i1} \dots x_{ij} \dots x_{ip} \\ \vdots \\ x_{n1} \dots x_{nj} \dots x_{np} \end{pmatrix} = \begin{pmatrix} c_{11} \\ c_{21} \\ \vdots \\ c_{i1} \\ \vdots \\ c_{n1} \end{pmatrix} (u_{11}, \dots, u_{1p}) + \begin{pmatrix} c_{12} \\ c_{22} \\ \vdots \\ c_{i2} \\ \vdots \\ c_{in} \end{pmatrix} (u_{21}, u_{22} \dots u_{2p}) + \dots$$

$$X = c_1 \cdot u_1^T + c_2 \cdot u_2^T + \dots c_q \cdot u_q^T$$

⇒ Formule de reconstitution de l'ACP.

On obtient donc la formule de reconstitution des données à l'aide de l'ACP :

Elle nous dit que les données d'origine peuvent être reconstituées à partir des composantes principales $(c_i, i=1\dots q)$ et des facteurs principaux $(u_i, i=1\dots p)$.

$$\begin{pmatrix} x_{11} & \dots & x_{1j} & \dots & x_{1p} \\ \vdots & & & & \\ x_{i1} & \dots & x_{ij} & \dots & x_{ip} \\ \vdots & & & & \\ x_{n1} & \dots & x_{nj} & \dots & x_{np} \end{pmatrix} = \begin{pmatrix} c_{11} \\ \vdots \\ c_{i1} \\ \vdots \\ c_{n1} \end{pmatrix} \cdot (u_{11}, \dots u_{1p}) + \dots \begin{pmatrix} c_{1q} \\ c_{2q} \\ \vdots \\ c_{iq} \\ \vdots \\ c_{in} \end{pmatrix} \cdot (u_{q1}, \dots, u_{qp})$$

⚠ relation exacte

Si l'on pense que les dernières composantes principales (peu explicatives) sont en fait liées à un "bruit" présent dans les données d'origine, alors on peut retenir dans cette décomposition uniquement $r < q$ composantes pour "débruiter" les données. On aurait alors :

$$\underbrace{X}_{\text{données observées}} = \underbrace{c_1 \cdot u_1^T + \dots + c_r \cdot u_r^T}_{\text{données débruitées}} + \underbrace{c_{r+1} \cdot u_{r+1}^T + \dots c_q \cdot u_q^T}_{\text{bruit.}}$$

Fin rappel 2.

Maintenant on est en mesure de présenter l'algorithme NIPALS.

NIPALS sur un tableau de données sans valeurs manquantes.

NIPALS $\Longleftrightarrow$ ACP.

On s'intéresse dans un premier temps au calcul de la 1ère composante principale, c_1.

c_1 est donnée par ses valeurs prises sur chaque individu :

$$c_1 = \begin{pmatrix} c_{11} \\ \vdots \\ c_{i1} \\ \vdots \\ c_{n1} \end{pmatrix}$$

On rappelle la formule de reconstitution (page 26, top)

$$(x_{i1}, \dots, x_{ip}) = c_{i1}\underbrace{(u_{11}, \dots u_{1p})}_{u_1} + c_{i2}\underbrace{(u_{21}, \dots u_{2p})}_{u_2} + \dots c_{ip}\underbrace{(u_{p1} \dots + u_{np})}_{u_p}$$

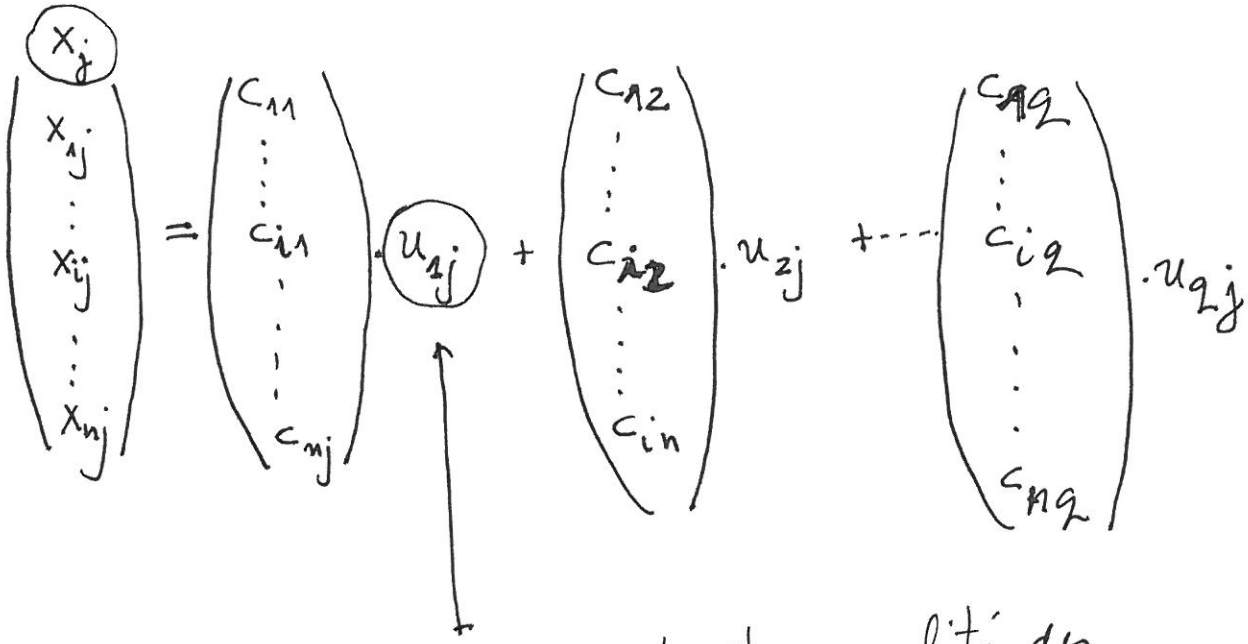
Remarque importante :

On constate ici que c_{i1}, $i = 1, \dots, n$ est le coefficient de la régression linéaire simple du vecteur $(x_{i1}, \dots, x_{ip})$ sur le vecteur $\underbrace{(u_{11}, \dots u_{1p})}_{u_1}$

Donc, si on connaît le vecteur $u_1 = (u_{11}, \dots u_{1p})$ on connaîtrait aussi le vecteur $c_1 = (c_{11}, \dots c_{1n})$ en réalisant n régressions simples.

Comment trouver le vecteur $u_1 = (u_{11} \dots u_{1p})$

Toujours grâce à la formule de reconstitution, on constate que la variable X_j (comme colonne) s'écrit :



De nouveau, grâce à l'orthogonalité des composantes c_i, on constate que

<u>Remarque</u>

u_{1j} est le coefficient de la régression linéaire simple de la variable X_j sur la variable c_1.

Donc pour ~~obtenir~~ trouver $u_1 = (u_{11} \dots u_{1p})$ on doit faire p regression linéaires simples entre $(X_j$ et $c_1)$, $j = 1 \dots p$.

On obtient donc l'algorithme <u>suivant :</u>

1. Initialisation :

$c_1 = X_1$ (la première composante est égale à la variable X_1)

↳ on pourrait prendre une autre si l'on veut.

2. Calcul de u_1 : effectuer donc p régressions linéaires simples.

for (j in 1 : p)

{ u_{1j} = coeff de la régression de X_j sur c_1

$\Updownarrow$

(*) $u_{1j} = \langle X_j, c_1 \rangle = \sum_{k=1}^{n} x_{kj} \cdot c_{k1}$

$\Updownarrow$

u_{1j} = coef (lm ($X_j \sim c_1$)) [2]

}

on normalise u_1 :

$$u_{1j} = u_{1j} / \sqrt{\sum u_{1j}^2}$$

3. Calcul de c_1 : on effectue n régressions linéaires entre les individus et u_1

for (i in 1 : n)

(**) { $c_{1i} = \sum_{k=1}^{p} x_{ik} \cdot u_{1k}$ # régression de l'individu i sur u_1

$\Updownarrow$

}. c_{1i} = coef (lm ($X[i,\cdot] \sim c_1$)) [2]

(4) Itérer les pas 2 et 3 jusqu'à la convergence.

↓

Soit on se donne un nombre fixé d'itérations

↳ soit on regarde d'un pas à l'autre l'évolution de

$$\| u_1^{(s)} - u_1^{(s+1)} \|$$

et on s'arrête lorsque cette quantité est inférieure à un seuil fixé ($\varepsilon = 10^{-5}$ par exemple).

Fin algorithme NIPALS pour le calcul de la composante 1

Comment on calcule la composante c_2 après avoir calculé la composante c_1 ?

On sait que

$$X = c_1 \cdot u_1^T + c_2 \cdot u_2^T + \dots + c_q \cdot u_q^T$$

$$X - c_1 \cdot u_1^T = c_2 \cdot u_2^T + \dots + c_q \cdot u_q^T$$

$$X_1 = c_2 \cdot u_2^T + \dots + c_q \cdot u_q^T$$

et on applique NIPALS pour la 1ère composante à la matrice X_1. ⚠

Après avoir obtenu

$(c_1, u_1), (c_2, u_2), \ldots, (c_h, u_h)$ pour obtenir c_{h+1} et u_{h+1} on applique l'algorithme NIPALS pour chercher la 1ère composante au tableau

$$X_h = X - c_1 \cdot u_1^T - c_2 \cdot u_2^T \cdots - c_h \cdot u_h^T .$$

FIN NIPALS données complètes

Et en présence de données manquantes ?

La force de NIPALS est qu'il est basé uniquement sur des régressions linéaires simples. Et ces régressions sont faisables en présence de données manquantes (approximation de coefficients sur les données (complètes) complètes !).

C'est-à-dire que si X et Y sont deux variables avec des données manquantes

$$X = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}, \quad Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$ alors, en travaillant sur les couples (x_i, y_i) où les données sont présentes,

$\hat{\beta}_1$ peut être estimé par

$$\hat{\beta}_1 = \sum_{\substack{i=1 \\ \{i \mid (x_i, y_i) \text{ non manquants}\}}}^{n} x_i \cdot y_i$$ à une constante près $\left(\frac{1}{\text{nb couples non manquants}}\right)$

Du coup, dans l'algorithme NIPALS (page 30) les pas (*) et (**) peuvent être modifiés par

(*) $$u_{1j} = \sum_{k=1}^{n} x_{kj} \cdot c_{k1} \Longrightarrow u_{1j} = \sum_{\substack{k=1 \\ \{k : c_{k1} \text{ existe}\}}}^{n} x_{kj} \cdot c_{k1}$$

(**) $$c_{1i} = \sum_{k=1}^{p} x_{ik} \cdot u_{1k} \Rightarrow c_{1i} = \sum_{\substack{k=1 \\ \{k : x_{ik} \text{ existe}\}}}^{p} x_{ik} \cdot u_{1k}.$$

A la fin de l'algorithme NIPALS appliqué pour le calcul des $\underline{r}$ composantes principales et r facteurs principaux on aura :

$$\hat{C}_1 = \begin{pmatrix} \hat{c}_{11} \\ \vdots \\ \hat{c}_{n1} \end{pmatrix}, \quad \hat{u}_1 = (\hat{u}_{11}, \dots \hat{u}_{1p})$$

$$\hat{C}_2 = \begin{pmatrix} \hat{c}_{12} \\ \vdots \\ \hat{c}_{n2} \end{pmatrix}, \quad \hat{u}_2 = (\hat{u}_{21}, \dots \hat{u}_{2p})$$

$$\hat{C}_r = \begin{pmatrix} \hat{c}_{1r} \\ \vdots \\ \hat{c}_{nr} \end{pmatrix}, \quad \hat{u}_r = (u_{r1}, \dots u_{rp})$$

avec données complètes ! (sans "trous").

Avec la formule de reconstitution des données on obtient

$$X \simeq \hat{C}_1 \cdot \hat{u}_1^T + \hat{C}_2 \cdot \hat{u}_2^T + \dots + \hat{C}_r \cdot \hat{u}_r^T$$

X : avec d.m.

$\hat{C}_1 \cdot \hat{u}_1^T + \hat{C}_2 \cdot \hat{u}_2^T + \dots + \hat{C}_r \cdot \hat{u}_r^T$: pas des d.m.

On obtient donc une estimation des données manquantes :

Si x_{ij} est l'observation de la variable j sur l'individu i alors son estimation à l'aide des r composantes principales est donnée par (voir page 29)

$$\hat{x}_{ij} = \hat{c}_{i1} \cdot \hat{u}_{1j} + \hat{c}_{i2} \cdot \hat{u}_{2j} + \dots + \hat{c}_{ir} \cdot \hat{u}_{rj}$$

- si x_{ij} est manquante, alors il s'agit d'une imputation.
- si x_{ij} est présente alors $\hat{x}_{ij}$ est une approximation de x_{ij} si $r < q$. (valeur exacte si $r = q$.).

Remarque importante :

Nous avons considéré au départ que les données du tableau X sont centrées et réduites. Avant d'appliquer NIPALS il faut donc centrer et réduire les variables (avec les valeurs manquantes !).

$$X \xrightarrow[\text{+ réduire}]{\text{centrer}} Z \xrightarrow{\text{NIPALS}} \hat{Z}$$

$\{\bar{x}_1, \dots \bar{x}_p\}$
$\{s_1^2, \dots s_p^2\}$

$$\hat{x}_{ij} = \bar{x}_j + s_j \cdot \hat{z}_{ij}$$

Fiche TP (R)

1 simulation :

1.1 générer un tableau X de $\underbrace{n = 100}_{\text{individus}} \times \underbrace{p = 4}_{\text{variables}}$

provenant d'une distribution normale multivariée

$N(\mu, \Sigma)$ avec

$\mu = (1, 2, 4, 3)$ et

$$\Sigma = \begin{pmatrix} 0.7 & 0 & 1.3 & 0.5 \\ 0 & 0.2 & -0.3 & -0.1 \\ 1.3 & -0.3 & 3.1 & 1.3 \\ 0.5 & -0.1 & 1.3 & 0.6 \end{pmatrix}.$$

Nb : Utiliser le package "mvtnorm"

1.2 Realiser l'ACP normée de X

- utiliser le package "FactoMineR"
- regarder les composantes principales et les facteurs propres.

1.3 programmer NIPALS. Comparer les resultats fournis par NIPALS avec ceux donnés par FactoMineR.

1.4. générer aléatoirement des valeurs manquantes dans le tableau X. On va supposer que $\mathbb{P}(X_{ij} \text{ soit manquante}) = 0.05$ (Données MCAR donc !).

(1.5) On considère maintenant

$\widetilde{X}$ le tableau avec données manquantes obtenu en 1.4.

Appliquer NIPALS pour estimer les données manquantes. Comparer avec les vraies valeurs

(2). Application réelle.

Imputer par NIPALS les données manquantes présentes dans la base "airquality" (base de donnée existante en R)

```
> help(airquality)
```

— Enjoy! —

Fin imputation simple. Passons à l'imputation multiple.