Analyse de données

L. BELLANGER

Master 1 Ingénierie Statistique
Dpt de Mathématiques - Université de Nantes

Plan

I. Outils de représentation d'un échantillon

Plan du ch. I

1. Structure d'un tableau de données

Questions préalables

Avant toute analyse de données recueillies après *expérience ou enquête*, il faut se poser un certain nombre de questions préalables:

- ***Que sont les données ?*** Sont-elles en nombre important, quelle est leur précision ?
- **D'où proviennent-elles ?** Les valeurs sont-elles raisonnables ?
- ***Comment et quand ont-elles été mesurées ?*** Des biais sont-ils possibles ? L'observateur sait-il arrondir les nombres ?
- ***Comment ont-elles été acquises ?*** A-t-on mesuré toutes les unités de la population, ou bien a-t-on fait un échantillonnage ?
- ***Existe-t-il une structure logique entre les observations ?*** Les observations ont-elles en commun certains facteurs d'environnement ?

Après avoir répondu à ces questions seulement, il est possible de pousser l'analyse plus loin !

1. Structure d'un tableau de données

La forme du tableau de données

- **Tableau individus×caractères**
  - Représentation des données sous la forme d'un *tableau rectangulaire* à :
    - I lignes et J colonnes ;
    - Formellement, un tableau de données $\mathbf{X}$ est de la forme :

$$\mathbf{X} = [x_{ij}] = \begin{bmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_I^T \end{bmatrix} = [\mathbf{x}^1 \quad \dots \quad \mathbf{x}^J] \; i = 1,\dots,I \text{ (ou } n\text{)}; \; j = 1,\dots,J \text{ (ou } p\text{)}$$

  - Les deux indices i et j repèrent la valeur ligne i, colonne j.
  - **Mais la représentation et le rôle d'une ligne (individu) ou d'une colonne (variable) sont ≠ :**
    - Le vecteur ligne $\boldsymbol{x}_i \in \mathbb{R}^J$ **mis en colonne** représente la $i^{\text{ème}}$ **observation élémentaire** sur laquelle ont été mesurées les J caractéristiques ou variables.
    - Le vecteur colonne $\boldsymbol{x}^j \in \mathbb{R}^I$ représente la $j^{\text{ème}}$ *variable ou indicateur*[1].

[1] Dans la terminologie du modèle linéaire, on emploie le mot ***facteur***.

1. Structure d'un tableau de données

La forme du tableau de données

- **Tableau individus×caractères**

Exemple (Bouroche & Saporta (1992))

		Caractères					
		Age x^1	Revenu imposable x^2	…	Salaire brut x^j	…	Ancien-neté x^p
Individus	1	x_1^1	x_1^2	…	x_1^j	…	x_1^p
	2	x_2^1	x_2^2		x_2^j	…	x_2^p
		…	…				
	i	x_i^1	x_i^2	…	x_i^j	…	x_i^p
		…					…
	n	x_n^1	x_n^2	…	x_n^j	…	x_n^p

1. Structure d'un tableau de données

La forme du tableau de données

- **Tableau individus×caractères**

Exemple suite (Bouroche & Saporta (1992))

Sur les mêmes individus, on aurait aussi pu observer les caractères « **sexe** » et « **situation matrimoniale** ». Pour être traitées numériquement, ces caractères doivent être représentés sous la forme d'un **tableau de variables indicatrices** prenant les valeurs 0 ou 1.

On dit alors que les données sont représentées sous **forme disjonctive complète**

		Caractères					
		Sexe		Situation matrioniale			
		F	M	Marié	Pacsé	Céliba-taire	Veuf, divorcé
Individus	1	1	0	1	0	0	0
	2	0	1	0	1	0	0
		…	…				
	i	1	0	0	0	1	0
		…					…
	n	0	1	0	0	0	1

1. Structure d'un tableau de données

La forme du tableau de données $\mathbf{X}$

• Tableau individus×caractères

- $\mathbf{X}$ est souvent constitué par **deux sous-tableaux** : $\mathbf{X} = [\mathbf{X}^1 \quad |\mathbf{X}^2]$
  - $\mathbf{X}^1$ (resp. $\mathbf{X}^2$) est un tableau à I lignes et $J = p$ (resp, q) colonnes.
  - L'individu i (ligne i de $\mathbf{X}$), est formée de p variables et q indicateurs ; les p variables définissent le ***vecteur observation*** lui-même « repéré » par un ***vecteur de contrôle*** .
  - L'ensemble des méthodes statistiques se rattachent :
    - soit à des ***méthodes unidimensionnelles*** *pour* l'analyse séparée de chaque colonne de $\mathbf{X}$.
    - soit à des ***méthodes multidimensionnelles*** pour l'analyse simultanée de l'ensemble des colonnes de $\mathbf{X}^1$.
  - L'existence d'un vecteur de contrôle induit une ***structure dans le tableau de données*** :
    - glt utilisée dans l'analyse des données ou
    - uniqt rôle d'aide à l'interprétation.

1. Structure d'un tableau de données

La forme du tableau de données $\mathbf{X}$

• Tableau individus×caractères

- $\mathbf{X}$ peut être **complété** par d'autres :
  - **observations** sur les mêmes J colonnes ; on parlera:
    - ***d'observations supplémentaires,*** si ces individus ne sont pas pris en compte dans l'analyse initiale
    - par opposition aux premières qui sont des ***observations actives***.
  - **colonnes** sur les mêmes individus ; on parlera de :
    - ***variables supplémentaires*** (les secondes) et
    - de ***variables actives*** (les premières).

1. Structure d'un tableau de données

La forme du tableau de données $\mathbf{X}$

• Tableau individus×caractères

- Schématiquement le tableau $\mathbf{X}$ **peut donc être complété** de la manière suivante :

$$\begin{bmatrix} \mathbf{X} & \mathbf{X}^{C_{sup}} \\ \mathbf{X}_{L_{sup}} & / \end{bmatrix}$$

le symbole « / » peut être remplacé par **des variables supplémentaires associées à des individus supplémentaires**.

1. Structure d'un tableau de données

La forme du tableau de données $\mathbf{X}$

• Tableau de contingence

- Un **tableau de contingence** est le croisement de 2 variables qualitatives A et B :

le coefficient n_{ij} du tableau = l'**effectif** (nb d'individus) présentant à la fois la modalité i de A et la modalité j de B.

- Dans un tel tableau, les individus ont été regroupés et ne peuvent plus être distingués.
- **Autre représentation** : à chacun des caractères nominaux, on associe un tableau de variables indicatrices (une variable par modalité).

1. Structure d'un tableau de données

La forme du tableau de données X

- **Tableau de contingence**
  - **Autre représentation : exemple**

$$X_1 = \begin{bmatrix} 1\,0\,0 \\ 0\,1\,0 \\ 0\,1\,0 \\ 1\,0\,0 \\ 0\,0\,1 \\ 0\,0\,1 \\ 0\,1\,0 \\ 0\,0\,1 \end{bmatrix} \quad X_2 = \begin{bmatrix} 1\,0 \\ 1\,0 \\ 0\,1 \\ 1\,0 \\ 0\,1 \\ 0\,1 \\ 1\,0 \\ 1\,0 \end{bmatrix}$$

, on obtient le tableau de contingence :

$$(X_1)^T X_2 = \begin{bmatrix} 2\,0 \\ 2\,1 \\ 1\,2 \end{bmatrix}$$

1. Structure d'un tableau de données

La forme du tableau de données X

- **Tableau de proximité**
  - On dispose de mesures de ressemblance ou de dissemblance entre tous les objets pris 2 à 2.
    - **ex** : tableau des distances entre les principales villes de France
  - Tableau glt ***symétrique*** contenant des ***nbs*** ≥ 0 analogues à des distances (ou à des inverses de distances) ; mais n'en possédant pas tjs toutes les propriétés axiomatiques (inégalité triangulaire)
    - **Rappel** : d est une ***distance*** si
      - (a) $d(a,b) = 0 \Leftrightarrow a = b$;
      - (b) $d(a,b) = d(b,a)$ (symétrie) ;
      - (c) $d(a,b) \leq d(a,c) + d(b,c)$ (*inégalité triangulaire*)
  - Si (c) n'est pas vérifiée, on parlera plutôt de ***dissimilarité***.

1. Structure d'un tableau de données

Notion de type de variables

- **Variables quantitatives**
  - ***continues*** (ou d'échelle) : dont les valeurs forment un sous-ensemble infini de $\mathbb{R}$
    (**ex** : poids, taille, revenu)
  - ***discrètes*** : dont les valeurs forment un sous-ensemble fini ou infini de $\mathbb{N}$
    (**ex** : nombre d'enfants)
- **Variables catégorielles (ou qualitatives)**
  - dont l'ensemble des valeurs est fini — ces valeurs sont numériques ou alphanumériques, mais quand elles sont numériques, ce ne sont que des codes et non des quantités (**ex** : n° de département)

1. Structure d'un tableau de données

Passage d'un codage simple à un codage disjonctif complet

Qd la valeur d'une variable ne correspond pas à une structure numérique, on peut s'y ramener par un simple *codage*.

Exemple : Couleur présentant trois caractéristiques {vert, jaune, marron}

On peut soit :

- établir la correspondance :
  {vert, jaune, marron} <=> {1, 2, 3}
- utiliser pour chaque réalisation une variable à deux valeurs slt, 0 ou 1. Chaque modalité de la variable définit une **variable dichotomique** ; la correspondance est alors :

Tableau 1 - **Exemple de passage d'un codage simple à un codage disjonctif complet.**

	Codage initial	vert	jaune	marron
Vert	1	1	0	0
Jaune	2	0	1	0
marron	3	0	0	1

Pour un traitement statistique, on utilise ce dernier codage appelé ***codage disjonctif complet***. Ainsi un objet vert sera analysé par l'ensemble des trois valeurs {1,0,0}, un jaune par {0,1,0} et un marron par {0,0,1}.

1. Structure d'un tableau de données

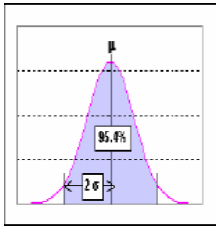
En résumé :

*Tableau 2 - **Les différents types de variables ou d'indicateurs, nature et exemples.***

Variable	Type	Nature	Exemples
quantitative	continu		poids,taille, revenu
	entier	discrète	nombre d'enfants
qualitative	polytomie ordonnée	ordinale	(mauvais,moyen, bon),
	polytomie non ordonnée	nominale	profession,couleur, variété végétale
	dichotomie	binaire	(vrai,faux), (oui, non), (présence, absence)

2. Résumés unidimensionnels

Pourquoi ?



- *Explorer la distribution* de chaque variable prise une à une pour :
  - vérifier la fiabilité des variables ;
  - détecter :
    - les valeurs incohérentes ou **manquantes**
      => **imputation** ou **suppression**
    - les valeurs extrêmes ou aberrantes (**outliers**) à éliminer ?
- Variables continues
  - tester la Normalité des variables (surtout si petits effectifs) et les transformer pour augmenter la Normalité.
- *Variables discrètes*
  - regrouper certaines modalités trop nombreuses ou avec des effectifs trop petits (poids trop grand).

2. Résumés unidimensionnels

Graphiques

Histogramme, **boîte à moustaches** (ou box-plot), ***graphique de densité*** *(après estimation de celle-ci)* et ***graphique de Normalité*** *(qq-plot ou pp-plot).*

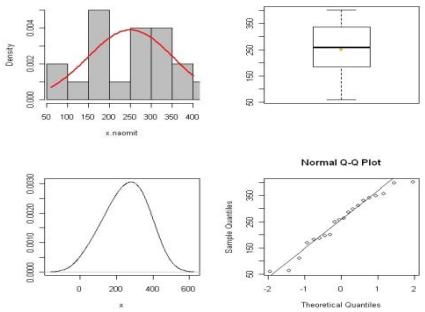


Figure 1. Eaux1 : Graphiques variable HCO_3

2. Résumés unidimensionnels

Graphiques

Histogramme

- Parlant uniquement si n est suffisamment élevé (>100),
- souvent utile de le remplacer ou ajouter le graphique de densité, version lissée de l'histogramme.

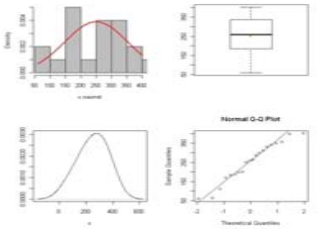


Figure 1. Eaux1 : Graphiques variable HCO_3

Sous R : `hist()`

2. Résumés unidimensionnels

Graphiques

Boîte à moustaches (ou boxplot) - *Sous R :* **`boxplot()`**

soit $Q_{0.25}$, $Q_{0.50}$ (la **médiane**) et $Q_{0.75}$ les valeurs tq 25%, 50% et les 75% des observations leur soient inférieures (i.e. les **quartiles**).

On trace une boîte comme celle de la *Fig* 2 :

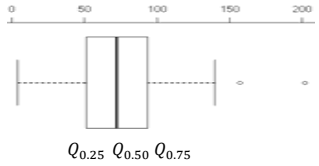


Figure 2. Eaux1 : Boxplot de Ca

- Les extrémités indiquées par « | » sont à :

  $\max(min, Q_{0.25} - 1.5(Q_{0.75} - Q_{0.25}))$ et $\min(\max, Q_{0.75} + 1.5(Q_{0.75} - Q_{0.25}))$
- les valeurs extérieures (« ° ») : suspectes ou extrêmes

↪ Très facile à interpréter ! On voit très rapidement si la distribution est symétrique.

2. Résumés unidimensionnels

Paramètres numériques classiques

- ***Position*** **: *moyenne*** $\bar{x}$**, *médiane*** Me**, *mode*** Mo

- La **moyenne**, notée $\bar{x}$, est définie par :

$$\bar{x} = \sum_{i=1}^{n} x_i / n$$

Si on a n_i valeurs x_i identiques, et si le nombre total est $n = \sum n_i$, on associe à chaque observation une **pondération** $p_i = n_i/n$, et donc $\sum p_i = 1$; alors on a :

$$\bar{x} = \sum_{i=1}^{n} p_i x_i$$

Sous R : **`mean()`** et **`summary(base)`**

On appelle **variable centrée** la variable x_c de coordonnées : $x_i - \bar{x}$;
↪ la moyenne de cette variable est donc nulle.

2. Résumés unidimensionnels

Paramètres numériques classiques

- ***Position*** **: *moyenne*** $\bar{x}$**, *médiane*** Me**, *mode*** Mo

- La **médiane**, notée Me, est la valeur qui divise les n obs. en 2 parties égales (50 % lui sont inférieures et 50 % supérieures)
- paramètre plus stable que $\bar{x}$ et moins sensible qu'elle aux valeurs suspectes. On dit que c'est un paramètre **robuste**.

Sous R : **`median()`**et **`summary(base)`**

2. Résumés unidimensionnels

Paramètres numériques classiques

- ***Position*** **: *moyenne*** $\bar{x}$**, *médiane*** Me**, *mode*** Mo

- Le **mode**, notée Mo, est la valeur ou la classe de fréquence maximale. Pour une variable quantitative, c'est la valeur la plus probable.
  - contrairement aux autres paramètres de position, le mode peut ne pas être unique : on parlera alors de **distribution multimodale**.
  - Pour une distribution Normale : $\bar{x} = Me = Mo$.

2. Résumés unidimensionnels

Paramètres numériques classiques

- *Dispersion* : ***ces paramètres*** définissent la variabilité de la distribution.

- Le plus utilisé, la **variance**, définie par :

$$s^2 = var(\boldsymbol{x}) = \sum_{i=1}^{n} (x_i - \bar{x})^2 \Big/ (n-1)$$

  - ➢ division par $(n-1)$ justifiée pour des raisons statistiques ; quelquefois la division est faite par n (cas de l'analyse des données …après) ;
  - ➢ la variance est alors la moyenne des carrés des écarts à la moyenne.

  On note s l'**écart-type** qui est $\sqrt{var(\boldsymbol{x})}$. Il s'exprime dans la même unité que la variable étudiée.

On appelle **variable centrée réduite** notée x_{CR} la variable de coordonnées : $(x_i - \bar{x})/s$; la moyenne de x_{CR} est donc 0 et son écart-type vaut 1.

Sous R : **`var(base)`** et **`sd(base)`**
avec ….division par $(n-1)$

2. Résumés unidimensionnels

Paramètres numériques classiques

- *Dispersion* :

- **Étendue** (souvent peu significative à cause des extrêmes) :

$$\max_i x_i - \min_i x_i$$

  - ➢ paramètre très sensible aux **valeurs extrêmes**.

- **Écart interquartile** : $Q_{0.75} - Q_{0.25}$
  - ➢ plus stable que l'étendue.

- **Coefficient de variation** : $CV(\%) = 100s/\bar{x}$
  - ➢ X dispersée si $CV > 25\%$
  - ➢ grandeur sans unité ⟹ utile pour comparer la dispersion de deux échantillons d'une même variable.

2. Résumés unidimensionnels

Paramètres numériques classiques

- *Centrer et réduire une variable*

  - Pour **centrer** une variable, on ôte à chaque individu la valeur moyenne $\bar{x}$.
  - Pour **réduire** une variable, on divise chaque individu par l'écart-type s_x .
  - La moyenne devient égale à 0. L'écart-type devient égal à 1.
  - On peut ramener ainsi les moyennes et écart-types des variables quantitatives d'un tableau de données à des valeurs identiques.

$$x_{CR} = \frac{x - \bar{x}}{s_x}$$

Sous R : **`scale(x, center = TRUE, scale = TRUE)`**

2. Résumés unidimensionnels

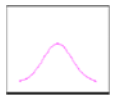
Paramètres numériques classiques

- *Paramètres de forme*

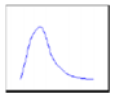
- **Coefficient d'asymétrie** (« skewness »)

$$b_1 = \frac{m_3}{{m_2}^{3/2}} = \frac{1}{n}\left[\sum_{i=1}^{n}(x_i - \bar{x})^3\right] \Big/ \left(\frac{1}{n}\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]\right)^{3/2}$$

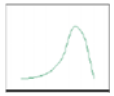
  - ➢ $= 0$ si la série de données est **symétrique**



  - ➢ > 0 si elle est **allongée vers la droite**
    - fréquent dans les données économiques



  - ➢ < 0 si elle est **allongée vers la gauche**



Sous R : **`skewness(e1071 ; agricolae ; moments)`**

2. Résumés unidimensionnels

Paramètres numériques classiques

- *Paramètres de forme (suite)*

▪ **Coefficient d'aplatissement** (« kurtosis »)

$$b_2 = \frac{m_4}{{m_2}^2} = \frac{1}{n}\left[\sum_{i=1}^{n}(x_i - \bar{x})^4\right] \Big/ \left(\frac{1}{n}\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]\right)^2$$

➢ $= 3$ si aplatie comme Gauss 

➢ > 3 si plus concentrée que Gauss 

➢ < 3 si plus aplatie que Gauss 

On normalise souvent le kurtosis en soustrayant **3** ; SAS et SPSS le font.

Sous R aussi : `kurtosis(e1071 ; agricolae ; moments)` …

2. Résumés unidimensionnels

Les variables qualitatives

Pour chaque modalité :

- Effectif : n_i nombre d'individus présentant la modalité i
- Fréquence :

$$f_i = \frac{n_i}{\sum_{i=1}^{I} n_i} = \frac{n_i}{n_+}$$

Pour une variable qualitative ordinale

- Fréquence cumulée : $F_i = \sum_{j=1}^{i} f_i$

Niveau d'études

		Effectifs	Pourcentage	Pourcentage valide	Pourcentage cumulé
Valide	<=Bac	12	12,0	12,0	12,0
	IUT	35	35,0	35,0	47,0
	Bac+4	26	26,0	26,0	73,0
	Ecoles	16	16,0	16,0	89,0
	Doctorat	11	11,0	11,0	100,0
	Total	100	100,0	100,0	

2. Résumés unidimensionnels

Les variables qualitatives

Représentations graphiques

Les représentations classiques pour une variable qualitative sont :

- Les **diagrammes en bâtons**
- Les **graphiques circulaires** (camemberts)

Pour chacun de ces graphiques on peut choisir d'afficher les effectifs ou les pourcentages.

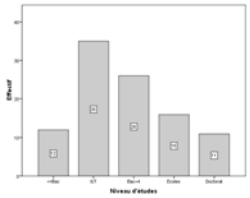
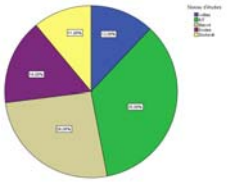


Diagramme en bâtons



Graphique circulaire

Sous R : `barplot( ), pie( ), mosaicplot( )`

3. Résumés multidimensionnels

La forme du tableau de données

un tableau de données **X** est une **représentation rectangulaire** à n lignes et p colonnes de la forme :

$$\mathbf{X} = \begin{bmatrix} x_1^1 & \cdots & x_1^j & \cdots & x_1^p \\ \vdots & \ddots & & \iddots & \vdots \\ x_i^1 & & x_i^j & & x_i^p \\ \vdots & \iddots & & \ddots & \vdots \\ x_n^1 & \cdots & x_n^j & \cdots & x_n^p \end{bmatrix}$$

← *individu i : vecteur ligne dans $\mathbb{R}^p$*

variable j : vecteur colonne dans $\mathbb{R}^n$

2 nuages de pts:

- nuage des n individus dans $\mathbb{R}^p$
- nuage des p variables dans $\mathbb{R}^n$

3. Résumés multidimensionnels

Espaces de représentation

Interprétation **géométrique** des lignes et les colonnes du tableau $\mathbf{X} \in \mathcal{M}_{n \times p}$ par des points dans 2 espaces différents:

l'espace des individus et **l'espace des variables** :

- *les espaces de représentation* :
  - **celui des n individus,** de dim p, noté $\Re^p \subset \mathbb{R}^p$
    - Les n lignes mis en colonne sont considérées comme n pts de **l'espace des individus** à p dimensions.

    2 points sont très proches si les p coord. de ces 2 pts sont très proches (mêmes valeurs pour les différentes variables).
  - **celui des p variables,** de dim n, noté $\Re^n \subset \mathbb{R}^n$
    - Les p colonnes sont considérées comme p pts de **l'espace des variables** à n dimensions.

    2 variables sont proches si leurs n coordonnées sont très voisines (i.e. ces variables mesurent la même chose ou sont liées par une relation particulière).

3. Résumés multidimensionnels

Un jeu de données est constitué par un **triplet** $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$ défini par les 3 éléments suivants:

1. $\mathbf{X} = \left[x_i^j\right]$ matrice des données brutes n mesures de p variables, quantitatives ou non
2. $\mathbf{Q}$, $p \times p$, métrique Euclidienne sur l'espace $\mathbb{R}^p$ des lignes $\boldsymbol{x}_i$ de $\mathbf{X}$ **(**transformées en colonne)
3. $\mathbf{D}$, $n \times n$, métrique Euclidienne sur l'espace $\mathbb{R}^n$ des colonnes $\boldsymbol{x}^j$ de $\mathbf{X}$, **tjs diagonale**. $\mathbf{D} = diag(p_1, \ldots, p_n)$

Les espaces Euclidiens $(\mathbb{R}^n, \mathbf{D})$ et $(\mathbb{R}^p, \mathbf{Q})$ sont resp. les **espaces des variables** et **des individus**.

Notation

$r = rg(\mathbf{X}) \leq min(n, p)$

3. Résumés multidimensionnels

Rappels d'algèbre linéaire (cf. WikiStat)

1. Notations
2. Matrices
3. Espaces euclidiens
4. Eléments propres
5. Optimisation

3. Résumés multidimensionnels

Rappels d'algèbre linéaire (cf. WikiStat)

- Notations et rappels d'algèbre linéaire de niveau L.
- Introduction des principaux théorèmes d'approximation matricielle par décomposition en valeurs singulières, à la base des méthodes statistique factorielles.

1. Notations

E et F, 2 espaces vectoriels réels munis respectivement des bases canoniques $\mathcal{E} = \{\boldsymbol{e}_j; j = 1,,,,,p\}$ et $\mathcal{F} = \{\boldsymbol{f}_i; i = 1,,,,,n\}$.

3. Résumés multidimensionnels

Rappels d'algèbre linéaire

2. Matrices

2.1 Notations et définitions

La matrice d'ordre $(n \times p)$ associée à une application linéaire de E dans F est décrite par un tableau :

$$\mathbf{A} = \begin{bmatrix} a_1^1 & \cdots & a_1^j & \ldots & a_1^p \\ \vdots & & & & \\ a_i^1 & \ldots & a_i^j & \ldots & a_i^p \\ \vdots & & & & \\ a_n^1 & \ldots & a_n^j & \ldots & a_n^p \end{bmatrix} \in \mathcal{M}_{n \times p}$$

On note par la suite :

- $a_i^j = [\mathbf{A}]_i^j$ le terme général de la matrice $\mathbf{A}$;
- $\boldsymbol{a}_i = \begin{bmatrix} a_i^1 & \ldots & a_i^p \end{bmatrix}^{\mathrm{T}}$ un vecteur-ligne *mis en colonne* ;
- $\boldsymbol{a}^j = \begin{bmatrix} a_1^j & \ldots & a_n^j \end{bmatrix}^{\mathrm{T}}$ un vecteur-colonne.

3. Résumés multidimensionnels

Rappels d'algèbre linéaire

2. Matrices

2.2 Définitions

Une matrice est dite :

- vecteur-ligne (colonne) si $n = 1$ $(p = 1)$;
- vecteur-unité d'ordre p si elle vaut $\mathbf{1}_p = [1 \quad \ldots \quad 1]^{\mathrm{T}}$;
- scalaire si $n = 1$ et $p = 1$;
- carrée si $n = p$.

Une matrice **carrée d'ordre** p est dite :

- identité notée $\mathbf{I}_p$ si $a_i^j = \delta_i^j = \begin{cases} 0 \; si \; i \neq j \\ 1 \; si \; i = j \end{cases}$;
- diagonale si $a_i^j = 0$ lorsque $i \neq j$; <u>sous R :</u> **`diag()`**
- symétrique si $a_i^j = a_j^i \; \forall (i,j)$;
- triangulaire supérieure (resp. inférieure) si $a_i^j = 0$ lque $i > j$ (resp. $i < j$).

3. Résumés multidimensionnels

Rappels d'algèbre linéaire

2. Matrices

2.2 Définitions

Une matrice est dite partitionnée en blocs si ses éléments sont eux-mêmes des matrices.

Par exemple :

$$\mathbf{A}_{n \times p} = \begin{bmatrix} \mathbf{A}^1_{1\,(r \times s)} & \mathbf{A}^2_{1\,(r \times (p-s))} \\ \mathbf{A}^1_{2\,((n-r) \times s)} & \mathbf{A}^2_{2\,(n-r) \times (p-s)} \end{bmatrix}$$

<u>Sous R :</u>
Créer une matrice **`matrix( )`**

3. Résumés multidimensionnels

Rappels d'algèbre linéaire

2. Matrices

2.2 Opérations sur les matrices

- ***Somme*** : $[\mathbf{A} + \mathbf{B}]_i^j = a_i^j + b_i^j$, pour $\mathbf{A}$ et $\mathbf{B}$ de même ordre $n \times p$;
- ***Multiplication par un scalaire*** : $[\alpha \mathbf{A}]_i^j = \alpha a_i^j, \forall \alpha \in \mathbb{R}$;
- **Transposition** : notation $\mathbf{A}^{\mathrm{T}}$ ou $\mathbf{A}'$
  - $[\mathbf{A}^{\mathrm{T}}]_i^j = a_j^i, \mathbf{A}^{\mathrm{T}} \in \mathcal{M}_{p \times n}$
  - $[\mathbf{A}^{\mathrm{T}}]^{\mathrm{T}} = \mathbf{A}$;
  - $[\mathbf{A} + \mathbf{B}]^{\mathrm{T}} = \mathbf{A}^{\mathrm{T}} + \mathbf{B}^{\mathrm{T}}$;
  - $(\mathbf{AB})^{\mathrm{T}} = \mathbf{B}^{\mathrm{T}} \mathbf{A}^{\mathrm{T}}$

  et
  - $\begin{bmatrix} \mathbf{A}^1_{1\,(r \times s)} & \mathbf{A}^2_{1\,(r \times (p-s))} \\ \mathbf{A}^1_{2\,((n-r) \times s)} & \mathbf{A}^2_{2\,(n-r) \times (p-s)} \end{bmatrix}^{\mathrm{T}} = \begin{bmatrix} \left[\mathbf{A}^1_{1\,(r \times s)}\right]^{\mathrm{T}} & \left[\mathbf{A}^1_{2\,((n-r) \times s)}\right]^{\mathrm{T}} \\ \left[\mathbf{A}^2_{1\,(r \times (p-s))}\right]^{\mathrm{T}} & \left[\mathbf{A}^2_{2\,(n-r) \times (p-s)}\right]^{\mathrm{T}} \end{bmatrix}$
  - <u>Sous R :</u> fonction **`t( )`**

3. Résumés multidimensionnels

Rappels d'algèbre linéaire

2. Matrices

2.2 Opérations sur les matrices

- **Produit scalaire élémentaire :** $\boldsymbol{a}^{\mathrm{T}}\boldsymbol{b} = \sum_{i=1}^{n} a_i b_i$ où $\boldsymbol{a}$ et $\boldsymbol{b}$ sont des vecteurs colonnes de $\mathbb{R}^n$;
- **Produit :** $[\mathbf{AB}]_i^j = \boldsymbol{a}_i{}^{\mathrm{T}}\boldsymbol{b}^j$ avec $\mathbf{A} \in \mathcal{M}_{n\times p}$, $\mathbf{B} \in \mathcal{M}_{p\times q}$ et $\mathbf{AB} \in \mathcal{M}_{n\times q}$

Et pour des matrices blocs :

$$\begin{bmatrix} \mathbf{A}_1^1 & \mathbf{A}_1^2 \\ \mathbf{A}_2^1 & \mathbf{A}_2^2 \end{bmatrix}\begin{bmatrix} \mathbf{B}_1^1 & \mathbf{B}_1^2 \\ \mathbf{B}_2^1 & \mathbf{B}_2^2 \end{bmatrix} = \begin{bmatrix} \mathbf{A}_1^1\mathbf{B}_1^1 + \mathbf{A}_1^2\mathbf{B}_2^1 & \mathbf{A}_1^1\mathbf{B}_1^2 + \mathbf{A}_1^2\mathbf{B}_2^2 \\ \mathbf{A}_2^1\mathbf{B}_1^1 + \mathbf{A}_2^2\mathbf{B}_2^1 & \mathbf{A}_2^1\mathbf{B}_1^2 + \mathbf{A}_2^2\mathbf{B}_2^2 \end{bmatrix}$$

Sous réserve de compatibilité des dimensions !

Sous R : `A %*% B`

$\Longrightarrow$ Faire exercice 3 feuille de Td/TP.

3. Résumés multidimensionnels

Rappels d'algèbre linéaire

2. Matrices

2.3 Propriétés des matrices carrées

- ***Trace***

Soit $\mathbf{A} \in \mathcal{M}_{p\times p}$ on définit la trace de $\mathbf{A}$ par : $tr\mathbf{A} = \sum_{j=1}^{p} a_j^j$

Propriétés de la trace :

- $tr\alpha = \alpha$;
- $tr\alpha\mathbf{A} = \alpha tr\mathbf{A}$; $tr\mathbf{A}^{\mathrm{T}} = tr\mathbf{A}$;
- $tr(\mathbf{A}+\mathbf{B}) = tr\mathbf{A} + tr\mathbf{B}$;
- $tr\mathbf{AB} = tr\mathbf{BA}$, reste vrai si $\mathbf{A} \in \mathcal{M}_{n\times p}$ et $\mathbf{B} \in \mathcal{M}_{p\times n}$;
- $tr\mathbf{C}\mathbf{C}^{\mathrm{T}} = tr\mathbf{C}^{\mathrm{T}}\mathbf{C} = \sum_{i=1}^{n}\sum_{j=1}^{p}\left(c_i^j\right)^2$ avec $\mathbf{C} \in \mathcal{M}_{n\times p}$.
- Sous R : `sum(diag( ))`

3. Résumés multidimensionnels

Rappels d'algèbre linéaire

2. Matrices

2.3 Propriétés des matrices carrées

- **Déterminant :**

« Initialement introduit en algèbre, pour résoudre un système d'équations linéaires comportant autant d'équations que d'inconnues. Il se révèle un outil très puissant dans de nombreux domaines. Il intervient ainsi dans l'étude des endomorphismes, la recherche de leurs valeurs propres, les propriétés d'indépendance linéaire de certaines familles de vecteurs, … » http://fr.wikipedia.org/

On note $det(\mathbf{A})$ ou $|\mathbf{A}|$, le déterminant de la matrice carrée $\mathbf{A} \in \mathcal{M}_{p\times p}$

Propriétés du déterminant :

- $det(\mathbf{A}) = \prod_{j=1}^{p} a_j^j$ si $\mathbf{A}$ est triangulaire ou diagonale ;
- $det(\alpha\mathbf{A}) = \alpha^p det(\mathbf{A})$; $det(\mathbf{AB}) = det(\mathbf{A})det(\mathbf{B})$;
- $det\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{0} & \mathbf{C} \end{bmatrix} = det(\mathbf{A})det(\mathbf{C})$;
- $det\begin{bmatrix} \mathbf{A}_1^1 & \mathbf{A}_1^2 \\ \mathbf{A}_2^1 & \mathbf{A}_2^2 \end{bmatrix} = det(\mathbf{A}_1^1)det\left(\mathbf{A}_2^2 - \mathbf{A}_2^1(\mathbf{A}_1^1)^{-1}\mathbf{A}_1^2\right) = det(\mathbf{A}_2^2)det\left(\mathbf{A}_1^1 - \mathbf{A}_1^2(\mathbf{A}_2^2)^{-1}\mathbf{A}_2^1\right)$ si $(\mathbf{A}_1^1)^{-1}$ et $(\mathbf{A}_2^2)^{-1}$ existent
- Sous R : `det( )`

3. Résumés multidimensionnels

Rappels d'algèbre linéaire

2. Matrices

2.3 Propriétés des matrices carrées

- **Inverse**

On note $\mathbf{A}^{-1}$, la matrice unique lorsqu'elle existe, inverse de $\mathbf{A}$, tq :

$$\mathbf{A}^{-1}\mathbf{A} = \mathbf{A}\mathbf{A}^{-1} = \mathbf{I} ;$$

elle existe si et seulement si $det(\mathbf{A}) \neq 0$.

Propriétés de l'inverse :

- $(\mathbf{A}^{-1})^{\mathrm{T}} = (\mathbf{A}^{\mathrm{T}})^{-1}$;
- $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$;
- $det(\mathbf{A}^{-1}) = {}^{1}/_{det(\mathbf{A})}$
- Sous R : `solve( )`

3. Résumés multidimensionnels

Rappels d'algèbre linéaire

2. Matrices

2.3 Propriétés des matrices carrées

Définitions :

Une matrice carrée est dite:

- ***symétrique*** si $\mathbf{A}^{\mathrm{T}} = \mathbf{A}$;
- ***singulière*** si $det(\mathbf{A}) = 0$;
- ***régulière*** si $det(\mathbf{A}) \neq 0$;
- ***idempotente*** si $\mathbf{AA} = \mathbf{A}^2 = \mathbf{A}$;
- ***définie-positive*** si, $\forall \boldsymbol{x} \in \mathbb{R}^p, \boldsymbol{x}^{\mathrm{T}}\mathbf{A}\boldsymbol{x} \geq 0$ et $\boldsymbol{x}^{\mathrm{T}}\mathbf{A}\boldsymbol{x} = 0 \Longrightarrow \boldsymbol{x} = \mathbf{0}$;
- ***positive***, *ou* ***semi-définie-positive*** si, $\forall \boldsymbol{x} \in \mathbb{R}^p, \boldsymbol{x}^{\mathrm{T}}\mathbf{A}\boldsymbol{x} \geq 0$;
- ***orthogonale*** si $\mathbf{AA}^{\mathrm{T}} = \mathbf{A}^{\mathrm{T}}\mathbf{A} = \mathbf{I}$ $\left(ie\ \mathbf{A}^{\mathrm{T}} = \mathbf{A}^{-1}\right)$.

3. Résumés multidimensionnels

Rappels d'algèbre linéaire

3. Espaces euclidiens

Soit E un espace vectoriel réel de dimension p isomorphe à $\mathbb{R}^p$

Rappels succincts :
un **isomorphisme** entre deux ensembles structurés est une application bijective qui préserve la structure et dont la réciproque préserve aussi la structure.
un **espace vectoriel** est un ensemble muni d'une structure permettant d'effectuer des combinaisons linéaires.

3.1 Sous-espaces (exercice 8 feuille de Td/TP)

- Un sous-ensemble E_q de E est un ***sous-espace vectoriel*** (s.e.v.) de E s'il est non vide et stable par combinaisons linéaires : $\forall (\boldsymbol{x}, \boldsymbol{y}) \in E_q^2, \forall \alpha \in \mathbb{R}, \alpha(\boldsymbol{x} + \boldsymbol{y}) \in E_q$.
- Le q-uple $\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_q\}$ de E constitue un ***sytème linéairement indépendant*** si et slt si : $\sum_{i=1}^{q} \alpha_i \boldsymbol{x}_i = 0 \Longrightarrow \alpha_1 = \cdots = \alpha_q = 0$.
  Dans le cas contraire ils sont dits *linéairement dépendants*.
- Un système linéairement indépendant $\varepsilon_q = \{\boldsymbol{e}_1, \dots, \boldsymbol{e}_q\}$ qui engendre le s.e.v. $E_q = vec\{\boldsymbol{e}_1, \dots, \boldsymbol{e}_q\}$ en constitue une **base** i.e. tout vecteur de E_q s'exprime de manière unique comme combinaison linéaire des éléments de ε_q ; on a alors : $dim(E_q) = card(\varepsilon_q) = q$.

3. Résumés multidimensionnels

Rappels d'algèbre linéaire

3. Espaces euclidiens

3.2 Rang d'une matrice $\mathbf{A} \in \mathcal{M}_{n \times p}$

Soit $\mathbf{A}$ la matrice d'une application linéaire de $E = \mathbb{R}^p$ dans $F = \mathbb{R}^n$.

Le **rang d'une matrice** : rang de l'application linéaire qu'elle représente ou encore rang de la famille de ses vecteurs colonnes (ie la dimension du sous-espace vectoriel engendré par cette famille).
Le **rang d'une application linéaire** f de E dans F : dimension de son image, s.e.v. de F.

Définition
- $Im(\mathbf{A}) = vect\{\boldsymbol{a}^1, \dots, \boldsymbol{a}^p\}$ est le s.e.v. de F ***image*** de $\mathbf{A}$;
- $Ker(\mathbf{A}) = \{\boldsymbol{x} \in E\ ; \mathbf{A}\boldsymbol{x} = \mathbf{0}\}$ est le s.e.v. de E ***noyau*** de $\mathbf{A}$;

Théorème du rang : Il relie dim(E) à celle du noyau de f et au rang de f.

$$\dim(E) = dim\left(Im(\mathbf{A})\right) + \dim\left(Ker(\mathbf{A})\right)$$

et $\quad \mathrm{rg}(\mathbf{A}) = dim\left(Im(\mathbf{A})\right)$ est le rang de $\mathbf{A}$

3. Résumés multidimensionnels

Rappels d'algèbre linéaire

3. Espaces euclidiens

3.2 Rang d'une matrice $\mathbf{A} \in \mathcal{M}_{n \times p}$

Le rang d'une matrice $\mathbf{A}$ est :

- le nombre maximal de vecteurs lignes (ou colonnes) linéairement indépendants ;
- la dimension du s.e.v. engendré par les vecteurs lignes (ou colonnes) de $\mathbf{A}$.

Propriétés :

- $\mathrm{rg}(\mathbf{A}) = dim\left(Im(\mathbf{A})\right)$; $0 \leq \mathrm{rg}(\mathbf{A}) \leq min(n, p)$;
- $\mathrm{rg}(\mathbf{A}) = \mathrm{rg}\left(\mathbf{A}^{\mathrm{T}}\right)$;
- $\mathrm{rg}(\mathbf{A} + \mathbf{B}) \leq \mathrm{rg}(\mathbf{A}) + \mathrm{rg}(\mathbf{B})$;
- $\mathrm{rg}(\mathbf{AB}) \leq \min\left(\mathrm{rg}(\mathbf{A}), \mathrm{rg}(\mathbf{B})\right)$;
- $\mathrm{rg}(\mathbf{A}) = \mathrm{rg}\left(\mathbf{AA}^{\mathrm{T}}\right) = \mathrm{rg}\left(\mathbf{A}^{\mathrm{T}}\mathbf{A}\right)$;
- Si $\mathbf{B} \in \mathcal{M}_{p \times q}$ de rang $q < p$ et $\mathbf{A} \in \mathcal{M}_{p \times p}$ de rang p alors la matrice $\mathbf{B}^{\mathrm{T}}\mathbf{AB}$ est de rang q.

3. Résumés multidimensionnels

Rappels d'algèbre linéaire

3. Espaces euclidiens

3.2 Rang d'une matrice $\mathbf{A} \in \mathcal{M}_{n\times p}$

Remarques :

- La Décomposition en Valeurs Singulières (DVS) de $\mathbf{A}$ est un outil efficace pour déterminer $\mathrm{rg}(\mathbf{A})$.
- $\mathbf{A}$ est carrée d'ordre p inversible $\Leftrightarrow det(\mathbf{A}) \neq 0$
  $\Leftrightarrow$ $\mathbf{A}$ de plein rang (i.e. $\mathrm{rg}(\mathbf{A}) = \mathrm{p}$).
- Sous R : `qr( )`

3. Résumés multidimensionnels

Rappels d'algèbre linéaire

3. Espaces euclidiens

3.3 Métrique euclidienne (exercice 9 feuille de Td/TP)

Soit $\mathbf{M} \in \mathcal{M}_{p\times p}$, matrice carrée, symétrique, définie-positive ; $\mathbf{M}$ définit sur l'espace E :

- un ***produit scalaire*** : $\langle \boldsymbol{x}, \boldsymbol{y} \rangle_{\mathbf{M}} = \boldsymbol{x}^{\mathrm{T}}\mathbf{M}\boldsymbol{y}$;
- une ***norme*** : $\|\boldsymbol{x}\|_{\mathbf{M}} = \langle \boldsymbol{x}, \boldsymbol{x} \rangle_{\mathbf{M}}^{1/2}$;
- une ***distance*** : $d_{\mathbf{M}}(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|_{\mathbf{M}}$;
- des ***angles*** : $\cos\theta_{\mathbf{M}}(\boldsymbol{x}, \boldsymbol{y}) = \frac{\langle \boldsymbol{x}, \boldsymbol{y} \rangle_{\mathbf{M}}}{\|\boldsymbol{x}\|_{\mathbf{M}} \|\boldsymbol{y}\|_{\mathbf{M}}}$.

La matrice $\mathbf{M}$ étant donnée, on dit que :

- une matrice $\mathbf{A}$ est $\mathbf{M}$-*symétrique* si $(\mathbf{MA})^{\mathrm{T}} = \mathbf{MA}$;
- deux vecteurs $\boldsymbol{x}$ et $\boldsymbol{y}$ sont $\mathbf{M}$-*orthogonaux* si $\langle \boldsymbol{x}, \boldsymbol{y} \rangle_{\mathbf{M}} = 0$;
- un vecteur $\boldsymbol{x}$ est $\mathbf{M}$-*normé* si $\|\boldsymbol{x}\|_{\mathbf{M}} = 1$;
- une base $\varepsilon_q = \{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_q\}$ est $\mathbf{M}$-*orthonormée* si $\forall (i,j), \langle \boldsymbol{e}_i, \boldsymbol{e}_j \rangle_{\mathbf{M}} = \delta_i^j$

3. Résumés multidimensionnels

Rappels d'algèbre linéaire

3. Espaces euclidiens

3.3 Métrique euclidienne

Remarque : Obtention de $\boldsymbol{u}_1$ $\mathbf{M}$-*normé*

Soient $\boldsymbol{u}$ vecteur de $\mathbb{R}^p$ et $\mathbf{M} \in \mathcal{M}_{p\times p}$, matrice carrée, symétrique, définie-positive :

1. On commence par calculer $\|\boldsymbol{u}\|_{\mathbf{M}} = \sqrt{\boldsymbol{u}^{\mathrm{T}}\mathbf{M}\boldsymbol{u}}$, $\mathbf{M}$-norme de $\boldsymbol{u}$;
2. On définit ensuite :

$$\boldsymbol{u}_1 = \frac{\boldsymbol{u}}{\|\boldsymbol{u}\|_{\mathbf{M}}} \ \mathbf{M}\text{-}normé \ \text{(ie } \|\boldsymbol{u}_1\|_{\mathbf{M}} = 1)$$

3. Résumés multidimensionnels

Rappels d'algèbre linéaire

3. Espaces euclidiens

3.4 Factorisation de Cholesky

Soit $\mathbf{M} \in \mathcal{M}_{p\times p}$, matrice carrée, symétrique, définie-positive,

$\exists\, \mathbf{T}$ matrice triangulaire supérieure avec $t_i^i > 0$ tq

$\mathbf{M} = \mathbf{T}^{\mathrm{T}}\mathbf{T}$; cette décomposition est unique !

Sous R : `chol( )`

Intérêt :

$\mathbf{M}$ carrée, symétrique, définie-positive est souvent utilisée en statistique comme métrique sur $\mathbb{R}^p$. On définit ainsi le produit scalaire : $\langle \boldsymbol{x}, \boldsymbol{y} \rangle_{\mathbf{M}} = \boldsymbol{x}^{\mathrm{T}}\mathbf{M}\boldsymbol{y}$; $\forall\, \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^p$

La décomposition de Cholesky de $\mathbf{M}$ donne : $\langle \boldsymbol{x}, \boldsymbol{y} \rangle_{\mathbf{M}} = \langle \mathbf{T}\boldsymbol{x}, \mathbf{T}\boldsymbol{y} \rangle_{\mathbf{I}_p}$

Changer de métrique

$\Leftrightarrow$

effectuer une transformation linéaire des données.

=> Voir plus loin en ACP ….

3. Résumés multidimensionnels

Rappels d'algèbre linéaire

3. Espaces euclidiens

3.5 Projection

Soit W un s.e.v. de E et $\mathcal{B} = \{\boldsymbol{b}^1, \dots, \boldsymbol{b}^q\}$ une base de W ;
$\mathbf{P} \in \mathcal{M}_{p\times p}$ est une matrice de projection **M**- orthogonale sur W

$$\Leftrightarrow$$

$$\forall \boldsymbol{y} \in E, \mathbf{P}\boldsymbol{y} \in W \text{ et } \langle \mathbf{P}\boldsymbol{y}, \boldsymbol{y} - \mathbf{P}\boldsymbol{y}\rangle_{\mathbf{M}} = 0$$

Propriété :
Toute matrice $\mathbf{P}$ *idempotente* ($\mathbf{P}^2 = \mathbf{P}$) et **M**-symétrique ($(\mathbf{MP})^{\mathrm{T}} = \mathbf{MP}$) est une matrice de projection **M**-orthogonale et réciproquement.

3. Résumés multidimensionnels

Rappels d'algèbre linéaire

3. Espaces euclidiens

3.5 Projection

Autres propriété :

- Les valeurs propres de $\mathbf{P}$ sont 0 ou 1 :
  - $\boldsymbol{u} \in W$; $\mathbf{P}\boldsymbol{u} = \boldsymbol{u}$; $\lambda = 1$, de multiplicité $\dim(W)$;
  - $\boldsymbol{v} \perp W$; (on note $\boldsymbol{v} \in W^{\perp}$) ; $\mathbf{P}\boldsymbol{v} = 0$, $\lambda = 0$, de multiplicité $\dim(W^{\perp})$.
- $tr\mathbf{P} = \dim(W)$;
- $\mathbf{P} = \mathbf{B}\left(\mathbf{B}^{\mathrm{T}}\mathbf{M}\mathbf{B}\right)^{-1}\mathbf{B}^{\mathrm{T}}\mathbf{M}$ où $\mathbf{B} = [\boldsymbol{b}^1, \dots, \boldsymbol{b}^q]$
- Dans le cas particulier où
  - les $\boldsymbol{b}^j$ sont **M**-orthonormés alors $\mathbf{P} = \mathbf{B}\mathbf{B}^{\mathrm{T}}\mathbf{M}$
  - $q = 1$ alors $\mathbf{P} = \frac{\boldsymbol{b}\boldsymbol{b}^{\mathrm{T}}}{\boldsymbol{b}^{\mathrm{T}}\mathbf{M}\boldsymbol{b}}\mathbf{M}$
- Si $\mathbf{P}_1, \dots, \mathbf{P}_q$ sont des matrices de projection **M**-orthogonales alors la somme $\mathbf{P}_1 + \cdots + \mathbf{P}_q$ est une matrice de projection **M**-orthogonale $\Leftrightarrow \mathbf{P}_k\mathbf{P}_j = \delta_j^k\mathbf{P}_j$.
- La matrice $\mathbf{I}_p - \mathbf{P}$ est la matrice de projection **M**-orthogonale sur $W^{\perp}$.

3. Résumés multidimensionnels

Rappels d'algèbre linéaire

4. Eléments propres : décomposition d'une matrice

Pour §4.1 à §4.3, soit $\mathbf{A} \in \mathcal{M}_{p\times p}$ une matrice **carrée**

4.1 Définitions (exercices 4, 5, 10, 11 feuille de Td/TP)

- Un vecteur $\boldsymbol{v} \neq \mathbf{0} \in \mathbb{R}^p$ est appelé vecteur propre de $\mathbf{A}$ s'il existe $\lambda \in \mathbb{C}$ tq :

$$\mathbf{A}\boldsymbol{v} = \lambda\boldsymbol{v}$$

  le scalaire λ est appelé valeur propre de $\mathbf{A}$ associé à $\boldsymbol{v}$
- Si λ est une valeur propre de $\mathbf{A}$, le noyau $ker(\mathbf{A} - \lambda\mathbf{I})$ est un s.e.v. de E, appelé sous-espace propre. Sa dimension est majorée par l'ordre de multiplicité de λ.
- Le spectre (ens. des valeurs propres) de $\mathbf{A}$ correspond aux racines, avec leur multiplicité, du *polynôme caractéristique* :

$$\mathrm{P}_{\mathbf{A}}(\lambda) = det(\mathbf{A} - \lambda\mathbf{I}) = 0$$

ensuite pour trouver $\boldsymbol{v}$ associé à λ résoudre : $(\mathbf{A} - \lambda\mathbf{I})\boldsymbol{v} = 0$.

Remarque : les vecteurs propres $\boldsymbol{v} \in ker(\mathbf{A} - \lambda\mathbf{I})$ n'étant définis qu'à une homothétie près, le plus souvent, on choisit de les normer à 1 en utilisant la norme euclidienne.

3. Résumés multidimensionnels

Rappels d'algèbre linéaire

4. Eléments propres

4.2 Propriétés

Soit $\{\lambda_1, \dots, \lambda_p\}$, le spectre de $\mathbf{A}$ tq $\lambda_1 \geq \cdots \geq \lambda_p$ alors

1. si $\lambda_k \neq \lambda_j$, $\boldsymbol{v}^k \perp_{\mathbf{M}} \boldsymbol{v}^j$;
2. $det(\mathbf{A}) = \prod_{i=1}^{p}\lambda_i$ et $tr(\mathbf{A}) = \sum_{i=1}^{p}\lambda_i$

Par conséquent :

- $\mathbf{A}$ est régulière $\Leftrightarrow \forall k, \lambda_k \neq 0$ (aucune v.p. nulle);
- $\mathbf{A}$ est positive (resp. définie-positive) $\Leftrightarrow \lambda_p \geq (resp.>)0$;
- si $(\lambda, \boldsymbol{v})$ est un élément propre de $\mathbf{A}$ alors

$$\mathbf{A}^k\boldsymbol{v} = \lambda^k\boldsymbol{v}, \forall \in \mathbb{N}^* \text{ et si } \mathbf{A} \text{ est régulière } \mathbf{A}^{-1}\boldsymbol{v} = \lambda^{-1}\boldsymbol{v}$$

3. Résumés multidimensionnels

Rappels d'algèbre linéaire

4. Eléments propres

4.2 Propriétés

Décomposition spectrale :
Soit $\{\lambda_1, \dots, \lambda_p\}$, le spectre de $\mathbf{A} \in \mathcal{M}_{p\times p}$ tq $\lambda_1 \geq \cdots \geq \lambda_p$ alors

$$\mathbf{\Lambda} = \mathbf{V}^{-1}\mathbf{A}\mathbf{V} \Leftrightarrow \mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{-1}$$

où
$\mathbf{\Lambda} = diag(\lambda_1, \dots, \lambda_p)$ et
$\mathbf{V} = [\boldsymbol{v}^1 \quad \cdots \quad \boldsymbol{v}^p]$ (matrice de passage de l'ancienne base à la base diagonale) matrice contenant les vecteurs propres de $\mathbf{A}$ associés aux valeurs propres rangées par ordre décroissant dans $\mathbf{\Lambda}$.

Quelques matrices diagonalisables particulières :
symétriques, $\mathbf{M}$-symétriques, définies-positives, de projection $\mathbf{M}$-orthog.

3. Résumés multidimensionnels

Rappels d'algèbre linéaire

4. Eléments propres

4.3 Théorèmes

Théorème de Cayley-Hamilton (pour ex. 11) :
Si le polynôme caractéristique $\mathrm{P}_{\mathbf{A}}(\lambda) = \lambda^p + P_{p-1}\lambda^{p-1} + \cdots + P_1\lambda + P_0$ alors
$\mathrm{P}_{\mathbf{A}}(\mathbf{A}) = \mathbf{A}^p + P_{p-1}\mathbf{A}^{p-1} + \cdots + P_1\mathbf{A} + P_0\mathbf{I} = \mathbf{0}$

Théorème 1.— Soit deux matrices $\mathbf{A} \in \mathcal{M}_{n\times p}$ et $\mathbf{B} \in \mathcal{M}_{p\times n}$
Les valeurs propres non nulles de $\mathbf{AB}$ et $\mathbf{BA}$ sont identiques avec le même degré de multiplicité.
Si $\boldsymbol{u}$ est vecteur propre de $\mathbf{BA}$ associé à la valeur propre différente de zéro, alors $\boldsymbol{v} = \mathbf{A}\boldsymbol{u}$ est vecteur propre de la matrice $\mathbf{AB}$ associé à la même valeur propre.

Décomposition spectrale pour des types particuliers de matrices.

3. Résumés multidimensionnels

Rappels d'algèbre linéaire

4. Eléments propres

4.3 Théorèmes

Théorème 2. — Une matrice $\mathbf{A}$ ***symétrique réelle*** admet p valeurs propres **réelles**. Ses vecteurs propres peuvent être choisis pour constituer une **base orthonormée** de E ; $\mathbf{A}$ se décompose en :

$$\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{\mathrm{T}} = \sum_{k=1}^{p} \lambda_k \boldsymbol{v}^k \left(\boldsymbol{v}^k\right)^{\mathrm{T}}$$

où
$\mathbf{\Lambda} = diag(\lambda_1, \dots, \lambda_p)$ et
$\mathbf{V} = [\boldsymbol{v}^1 \quad \cdots \quad \boldsymbol{v}^p]$ matrice orthogonale contenant les vecteurs propres de $\mathbf{A}$ associés aux valeurs propres réelles rangées par ordre décroissant dans $\mathbf{\Lambda}$.

(pour ex. 10)

3. Résumés multidimensionnels

Rappels d'algèbre linéaire

4. Eléments propres

4.3 Théorèmes

Théorème 3.— Une matrice $\mathbf{A}$ ***$\mathbf{M}$-symétrique réelle*** admet p valeurs propres réelles. Ses vecteurs propres peuvent être choisis pour constituer une base $\mathbf{M}$-orthonormée de E ; $\mathbf{A}$ se décompose en :

$$\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{\mathrm{T}}\mathbf{M} = \textstyle\sum_{k=1}^{p} \lambda_k \boldsymbol{v}^k \left(\boldsymbol{v}^k\right)^{\mathrm{T}} \mathbf{M}$$

où
$\mathbf{\Lambda} = diag(\lambda_1, \dots, \lambda_p)$ et
$\mathbf{V} = [\boldsymbol{v}^1 \quad \cdots \quad \boldsymbol{v}^p]$ matrice $\mathbf{M}$-orthogonale $(\mathbf{V}^{\mathrm{T}}\mathbf{M}\mathbf{V} = \mathbf{I}_p$ et $\mathbf{V}\mathbf{V}^{\mathrm{T}} = \mathbf{M}^{-1})$ contenant les vecteurs propres de $\mathbf{A}$ associés aux valeurs propres réelles rangées par ordre décroissant dans $\mathbf{\Lambda}$.

Par définition, si $\mathbf{A}$ est aussi def. **positive**, on note la racine carrée de $\mathbf{A}$:

$$\mathbf{A}^{1/2} = \mathbf{V}\mathbf{\Lambda}^{1/2}\mathbf{V}^{\mathrm{T}}\mathbf{M} = \textstyle\sum_{k=1}^{p} \sqrt{\lambda_k}\, \boldsymbol{v}^k \left(\boldsymbol{v}^k\right)^{\mathrm{T}} \mathbf{M}$$

Remarques :

- Les décompositions ne sont pas uniques :
  - pour une valeur propre simple (de multiplicité 1) le vecteur propre normé est défini à un signe près, tandis que
  - pour une valeur propre multiple, une infinité de bases $\mathbf{M}$-orthonormées peuvent être extraites du sous-espace propre unique associé.
- $rg(\mathbf{A}) = rg(\mathbf{\Lambda})$ où $rg(\mathbf{\Lambda})$ n'est autre que le nombre (répétées avec leurs multiplicités) de valeurs propres non nulles.
- Sous R : `eigen()`

4.4 Décomposition en valeurs singulières (DVS)

↪ Construction de la décomposition d'une matrice ***rectangulaire*** $\mathbf{X} \in \mathcal{M}_{n\times p}$

Pour une matrice rectangulaire, la notion de valeur propre n'a pas de sens ; néanmoins :

- les matrices carrées $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ et $\mathbf{X}\mathbf{X}^{\mathrm{T}}$ sont sym., semi-définie positives ;
- $rg(\mathbf{X}) = rg(\mathbf{X}^{\mathrm{T}}\mathbf{X}) = rg(\mathbf{X}\mathbf{X}^{\mathrm{T}}) = r$, les valeurs propres $\neq 0$ (positives) de $\mathbf{X}\mathbf{X}^{\mathrm{T}}$ et $\mathbf{X}^{\mathrm{T}}\mathbf{X}$.

Définition : On appelle **valeurs singulières** de $\mathbf{X}$, les racines carrées des valeurs propres $\neq 0$ de $\mathbf{X}\mathbf{X}^{\mathrm{T}}$ et $\mathbf{X}^{\mathrm{T}}\mathbf{X}$.

4.4 Décomposition en valeurs singulières (DVS)

Version « maigre » : Toute matrice $\mathbf{X}_{n\times p}$ peut être décomposée de la façon suivante :

$$\mathbf{X}_{n\times p} = \mathbf{U}_{n\times r}\mathbf{\Theta}_{r\times r}\left(\mathbf{V}_{p\times r}\right)^{\mathrm{T}}$$

Où

- $\mathbf{\Theta}_{r\times r} = \mathbf{\Lambda}_r{}^{1/2}$ matrice **diagonale** contenant les r **valeurs singulières** de la matrice $\mathbf{X}$; *i.e.* les racines carrées des v.p. **positives et non nulles** λ_i de $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ ou de $\mathbf{X}\mathbf{X}^{\mathrm{T}}$; $\lambda_1 \geq \lambda_2 \geq \cdots \lambda_r > 0$
- $r = rg(\mathbf{X}) \leq min(n,p)$. Si $r = p$, la matrice $\mathbf{X}$ est dite de ***plein rang***.
- $\mathbf{U}_{n\times r} = [\boldsymbol{u}^1 \quad \ldots \quad \boldsymbol{u}^r]$ unitaire $n \times r$ et tq $\boldsymbol{u}^i$ est vecteur propre de $\mathbf{X}\mathbf{X}^{\mathrm{T}}$ associé à la valeur propre non nulle λ_i

  $\Rightarrow$ *Vecteurs principaux de l'espace des colonnes*
- $\mathbf{V}_{p\times r} = [\boldsymbol{v}^1 \quad \ldots \quad \boldsymbol{v}^r]$ unitaire $p \times r$ et tq $\boldsymbol{v}^i$ est vecteur propre de $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ associé à la valeur propre non nulle λ_i

  $\Rightarrow$ *Vecteurs principaux de l'espace des lignes*

$\mathbf{U}$ et $\mathbf{V}$ sont des matrices colonnes-orthonormales (i.e. matrices contenant des vecteurs colonnes unitaires et orthogonaux 2 à 2).

4.4 Décomposition en valeurs singulières (DVS)

Remarques DVS :

- $[\boldsymbol{u}^1 \quad \ldots \quad \boldsymbol{u}^r]$ vecteurs principaux de l'espace des colonnes ; $\|\boldsymbol{u}^i\| = 1$ et si $i \neq j$ alors $\langle \boldsymbol{u}^i; \boldsymbol{u}^j \rangle = 0$.
- $[\boldsymbol{v}^1 \quad \ldots \quad \boldsymbol{v}^r]$ vecteurs principaux de l'espace des lignes ; $\|\boldsymbol{v}^i\| = 1$ et si $i \neq j$ alors $\langle \boldsymbol{v}^i; \boldsymbol{v}^j \rangle = 0$.
- $\sqrt{\lambda_1} \geq \sqrt{\lambda_2} \geq \cdots \geq \sqrt{\lambda_r} > 0$ valeurs singulières.
- Cette décomposition **n'est pas unique**.
- Sous R : `svd( )`

Rappels d'algèbre linéaire

4.4 Décomposition en valeurs singulières (DVS)

En pratique :

- Le nb de valeurs singulières fournit le rg de la matrice $\mathbf{X}$. Ces valeurs singulières sont ordonnées, ce qui induit un ordre sur les colonnes de $\mathbf{U}$ et $\mathbf{V}$.
- Calcul de $\mathbf{U}$ et $\mathbf{V}$: Calcul des vecteurs propres de $\mathbf{X}^T\mathbf{X}$ ou de $\mathbf{X}\mathbf{X}^T$ (prendre la matrice de + petite dim). Les vecteurs propres de l'autre se déduisent par les *formules de transition* :

$$\begin{cases} \mathbf{U} = \mathbf{X}\mathbf{V}\mathbf{\Lambda}_r^{-1/2} \\ \mathbf{V} = \mathbf{X}^T\mathbf{U}\mathbf{\Lambda}_r^{-1/2} \end{cases}$$

Ex 12 feuille de Td/TP

Rappels d'algèbre linéaire

4.4 Décomposition en valeurs singulières (DVS)

Sous R : **svd( )**

Applications : Rang d'une matrice

$$\begin{bmatrix} 2 & 1 & 3 & 4 \\ -1 & 6 & -3 & 0 \\ 1 & 20 & -3 & 8 \end{bmatrix} = \begin{bmatrix} -0.08682 & 0.84068 \\ -0.26247 & -0.53689 \\ -0.96103 & 0.07069 \end{bmatrix} \begin{bmatrix} 22.650 & 0 \\ 0 & 6.081 \end{bmatrix} \begin{bmatrix} -0.03851 & -0.92195 & 0.15055 & -0.35477 \\ 0.37642 & -0.15902 & 0.64476 & 0.64600 \end{bmatrix}$$

$r = rg(\mathbf{X}) = 2 \Longrightarrow \mathbf{X}$ *n'est pas de plein rang*

Rappels d'algèbre linéaire

4.4 Décomposition en valeurs singulières (DVS)

Applications : Décomposition d'une matrice carrée

Considérons une matrice de variance-covariance $\mathbf{S} = \mathbf{X}^T\mathbf{D}\mathbf{X}$

$$S_{2*2} = \mathbf{U}_{2*2}\mathbf{\Theta}_{2*2}\mathbf{V}^T_{2*2}$$

$$\begin{bmatrix} 9.2 & 1.6 \\ 1.6 & 5.8 \end{bmatrix} = \begin{bmatrix} -0.8944 & -0.4472 \\ -0.4472 & 0.8944 \end{bmatrix} \begin{bmatrix} 9 & 0 \\ 0 & 5 \end{bmatrix} \begin{bmatrix} -0.8944 & -0.4472 \\ -0.4472 & 0.8944 \end{bmatrix}$$

- Les valeurs singulières de $\mathbf{S}$ obtenues sur la diagonale de la matrice $\mathbf{\Theta}$ sont égales aux ***valeurs propres*** de $\mathbf{S}$ qui sont réelles : ceci est vrai pour toute matrice carrée symétrique.
- La matrice $\mathbf{U}$ (resp. $\mathbf{V}$) contient des vecteurs colonnes identiques aux vecteurs propres obtenus par diagonalisation de $\mathbf{S}$, leurs signes peuvent varier en fonction du programme utilisé.

Rappels d'algèbre linéaire

4.4 Décomposition en valeurs singulières (DVS)

Applications : Inverse généralisée d'une matrice

Inversion de $\mathbf{X}^T\mathbf{X}$ quand $\mathbf{X}$ n'est pas de plein rg (pb en régression quand pb de colinéarité).

On peut montrer que :

$$\left(\mathbf{X}^T\mathbf{X}\right)^- = \mathbf{V}\mathbf{\Theta}^-\mathbf{U}^T$$

où

$\mathbf{\Theta}^-$ est l'inverse généralisée de $\mathbf{\Theta}$ obtenue en prenant dans cette dernière matrice l'inverse des éléments diagonaux non nuls.

3. Résumés multidimensionnels

Rappels d'algèbre linéaire

4.4 Décomposition en valeurs singulières (DVS)

Généralisation :

En Analyse des Données (AD) : définition de métriques sur chacun des espaces associés aux lignes et aux colonnes d'une matrice réelle.

On définit le triplet $(\mathcal{M}_{n\times p}, \mathbf{Q}, \mathbf{D})$ par la donnée de :

- $\mathcal{M}_{n\times p}$ espace vectoriel des matrices réelles $n \times p$, de dim np
- $\mathbf{Q} \in \mathcal{M}_{p\times p}$ métrique sur l'espace des lignes $\mathcal{R}^p \subset \mathbb{R}^p$;
- $\mathbf{D} \in \mathcal{M}_{n\times n}$ métrique sur l'espace des colonnes $\mathcal{R}^n \subset \mathbb{R}^n$.

DVS de $\mathbf{X} \in \mathcal{M}_{n\times p}$ relativement à 2 matrices symétriques et positives $\mathbf{Q}$ et $\mathbf{D}$

3. Résumés multidimensionnels

Rappels d'algèbre linéaire

4.4 Décomposition en valeurs singulières (DVS)

Généralisation : Version « maigre »

Théorème 4.— Une matrice $\mathbf{X} \in \mathcal{M}_{n\times p}$ de rang r peut s'écrire :

$$\mathbf{X}_{n\times p} = \mathbf{U}_{n\times r}\mathbf{\Theta}_{r\times r}\left(\mathbf{V}_{p\times r}\right)^{\mathrm{T}} = \sum_{k=1}^{r}\sqrt{\lambda_k}\boldsymbol{u}^k\left(\boldsymbol{v}^k\right)^{\mathrm{T}}$$

- $\mathbf{U}_{n\times r}$ contient les vecteurs propres $\mathbf{D}$-orthonormés ($\mathbf{U}^{\mathrm{T}}\mathbf{D}\mathbf{U} = \mathbf{I}_r$) de la matrice $\mathbf{D}$-symétrique positive $\mathbf{X}\mathbf{Q}\mathbf{X}^{\mathrm{T}}\mathbf{D}$ associés aux r v.p. $\lambda_k \neq 0$ rangées par ordre décroissant dans la matrice diagonale $\mathbf{\Theta}_{r\times r}{}^2 = \mathbf{\Lambda}_r$;
- $\mathbf{V}_{p\times r}$ contient les vecteurs propres $\mathbf{Q}$-orthonormés ($\mathbf{V}^{\mathrm{T}}\mathbf{Q}\mathbf{V} = \mathbf{I}_r$) de la matrice $\mathbf{Q}$-symétrique positive $\mathbf{X}^{\mathrm{T}}\mathbf{D}\mathbf{X}\mathbf{Q}$ associés aux mêmes valeurs propres.

De plus,

$$\mathbf{U} = \mathbf{X}\mathbf{Q}\mathbf{V}\mathbf{\Lambda}_r^{-1/2} \text{ et } \mathbf{V} = \mathbf{X}^{\mathrm{T}}\mathbf{D}\mathbf{U}\mathbf{\Lambda}_r^{-1/2}$$

3. Résumés multidimensionnels

Rappels d'algèbre linéaire

5. Optimisation

5.1 Norme d'une matrice

Soit le triplet $(\mathcal{M}_{n\times p}, \mathbf{Q}, \mathbf{D})$ par la donnée de :

- $\mathcal{M}_{n\times p}$ espace vectoriel des matrices réelles $n \times p$, de dim np
- $\mathbf{Q} \in \mathcal{M}_{p\times p}$ métrique sur l'espace des lignes $\mathcal{R}^p \subset \mathbb{R}^p$;
- $\mathbf{D} \in \mathcal{M}_{n\times n}$ métrique sur l'espace des colonnes $\mathcal{R}^n \subset \mathbb{R}^n$.

Définition : On munit $\mathcal{M}_{n\times p}$ du *produit scalaire de Hilbert-Schmidt* :

$$\langle \mathbf{X}, \mathbf{Y}\rangle_{\mathbf{Q},\mathbf{D}} = tr\left(\mathbf{X}\mathbf{Q}\mathbf{Y}^{\mathrm{T}}\mathbf{D}\right) ; \forall\, \mathbf{X}, \mathbf{Y} \in \mathcal{M}_{n\times p}$$

Cas particulier : $\mathbf{Q} = \mathbf{I}_p$ et $\mathbf{D} = \mathbf{I}_n$, on a :

$$\langle \mathbf{X}, \mathbf{Y}\rangle_{\mathbf{I}_p,\mathbf{I}_n} = tr\left(\mathbf{X}\mathbf{Y}^{\mathrm{T}}\right) = \sum_{i=1}^{n}\sum_{j=1}^{p} x_i^j y_i^j$$

3. Résumés multidimensionnels

Rappels d'algèbre linéaire

5. Optimisation

5.1 Norme d'une matrice

- La norme associée au *produit scalaire de Hilbert-Schmidt* est encore appelée *norme trace* ou *norme de Fröbenius* :

$$\|\mathbf{X}\|^2_{\mathbf{Q},\mathbf{D}} = tr\left(\mathbf{X}\mathbf{Q}\mathbf{X}^{\mathrm{T}}\mathbf{D}\right) ; \forall\, \mathbf{X} \in \mathcal{M}_{n\times p}$$

Cas particulier : $\mathbf{Q} = \mathbf{I}_p$ et $\mathbf{D} = \mathbf{I}_n$, on a :

$$\|\mathbf{X}\|^2_{\mathbf{I}_p,\mathbf{I}_n} = tr\left(\mathbf{X}\mathbf{X}^{\mathrm{T}}\right) = \sum_{i=1}^{n}\sum_{j=1}^{p}\left(x_i^j\right)^2 = SSQ(\mathbf{X})$$

SSQ : « sum of squares ».

- Si $\mathbf{D} = \mathbf{diag}(p_1, \ldots, p_n)$, la distance associée à cette norme s'écrit :

$$d^2(\mathbf{X}, \mathbf{Y}) = \|\mathbf{X} - \mathbf{Y}\|^2_{\mathbf{Q},\mathbf{D}} = \sum_{i=1}^{n} p_i \|\boldsymbol{x}_i - \boldsymbol{y}_i\|^2_{\mathbf{Q}}$$

appelé *critère des moindres carrés ordinaires*.

3. Résumés multidimensionnels

Rappels d'algèbre linéaire

5. Optimisation

5.2 Approximation d'une matrice

Soient $\mathbf{X} \in \mathcal{M}_{n\times p}$ de rang r, $\mathbf{Q} \in \mathcal{M}_{p\times p}$ et $\mathbf{D} \in \mathcal{M}_{n\times n}$

Objectif: Trouver la matrice $\mathbf{Z}_q$ de rang $q < r$, qui soit la + proche possible de $\mathbf{X}$.

Théorème 5. La solution du problème :

$$\min_{\mathbf{Z}}\{\|\mathbf{X}-\mathbf{Z}\|^2_{\mathbf{Q},\mathbf{D}} ; \mathbf{Z} \in \mathcal{M}_{n\times p}, rg(\mathbf{Z}) = q < r\}$$

est donnée par la somme des q 1ers termes de la DVS (Théo 4) de $\mathbf{X}$:

$$\mathbf{Z}_q = \sum_{k=1}^{q} \sqrt{\lambda_k}\boldsymbol{u}^k(\boldsymbol{v}^k)^{\mathrm{T}} = \mathbf{U}_{n\times q}\boldsymbol{\Theta}_{q\times q}(\mathbf{V}_{p\times q})^{\mathrm{T}}$$

- Le minimum atteint est : $\|\mathbf{X}-\mathbf{Z}_q\|^2_{\mathbf{Q},\mathbf{D}} = \sum_{k=q+1}^{r}\lambda_k$
- Les matrices $\mathbf{U}_{n\times q}$ et $\mathbf{V}_{p\times q}$ contiennent les q 1ers vecteurs et valeurs propres donnés par la DVS de $\mathbf{X}$.

$\mathbf{Z}_q$ est appelée approximation de rang q de $\mathbf{X}$.

3. Résumés multidimensionnels

Espaces de représentation

Interprétation **géométrique** des lignes et les colonnes du tableau $\mathbf{X} \in \mathcal{M}_{n\times p}$ par des points dans 2 espaces différents:

l'espace des individus et **l'espace des variables** :

- *les espaces de représentation* :
  - **celui des n individus,** de dim p, noté $\Re^p \subset \mathbb{R}^p$
    - Les n lignes mis en colonne sont considérées comme n pts de **l'espace des individus** à p dimensions.

    2 points sont très proches si les p coord. de ces 2 pts sont très proches (mêmes valeurs pour les différentes variables).
  - **celui des p variables,** de dim n, noté $\Re^n \subset \mathbb{R}^n$
    - Les p colonnes sont considérées comme p pts de **l'espace des variables** à n dimensions.

    2 variables sont proches si leurs n coordonnées sont très voisines (i.e. ces variables mesurent la même chose ou sont liées par une relation particulière).

3. Résumés multidimensionnels

Un jeu de données est constitué par un triplet $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$ défini par les 3 éléments suivants:

1. $\mathbf{X} = \left[x_i^j\right]$ matrice des données brutes n mesures de p variables, quantitatives ou non
2. $\mathbf{Q}$, $p \times p$, métrique Euclidienne sur l'espace $\mathbb{R}^p$ des n lignes $\boldsymbol{x}_i$ de $\mathbf{X}$ (transformées en colonne)
3. $\mathbf{D}$, $n \times n$, métrique Euclidienne sur l'espace $\mathbb{R}^n$ des p colonnes $\boldsymbol{x}^j$ de $\mathbf{X}$, **tjs diagonale**. $\mathbf{D} = diag(p_1, \ldots, p_n)$

Les espaces Euclidiens $(\mathbb{R}^n, \mathbf{D})$ et $(\mathbb{R}^p, \mathbf{Q})$ sont resp. les espaces des variables et des individus.

Notation

$r = rg(\mathbf{X}) \leq min(n, p)$

3. Résumés multidimensionnels

Espaces de représentation

- Le **point moyen du nuage des n points** de $\mathbb{R}^p$, aussi appelé ***centre de gravité ou centroïde***, a comme coordonnées celles du ***vecteur moyenne*** , **calculées par :**

$$\bar{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i = \frac{1}{n}\mathbf{X}^{\mathrm{T}}\mathbf{1}_{n\times 1} \in \mathbb{R}^p$$

où $\mathbf{1}_{n\times 1}$ vecteur à n lignes formées de 1

i.e. $\bar{\boldsymbol{x}} = (\bar{x}^1 \quad \ldots \quad \bar{x}^p)$ où $\bar{x}^j$ est la ***moyenne empirique*** de la variable j (de même s^j sera son ***écart-type***)

3. Résumés multidimensionnels

Espaces de représentation

En retranchant à chaque colonne sa moyenne, on obtient la **matrice des données centrées** :

$$\mathbf{X_C}^{\mathrm{T}} = [\boldsymbol{x}_1 - \bar{\boldsymbol{x}} \quad \cdots \quad \boldsymbol{x}_n - \bar{\boldsymbol{x}}] = \mathbf{X}^{\mathrm{T}} - \bar{\boldsymbol{x}}\mathbf{1}_{1\times n}$$
$$= \mathbf{X}^{\mathrm{T}}\left[\mathbf{I_n} - \frac{1}{n}\mathbf{1}_{n\times 1}(\mathbf{1}_{n\times 1})^{\mathrm{T}}\right] = \mathbf{X}^{\mathrm{T}}\left[\mathbf{I}_n - \frac{1}{n}\mathbf{1}_{n\times n}\right]$$

- où $\mathbf{1}_{n\times n}$ matrice formée de 1 et $\mathbf{I_n} - \frac{1}{n}\mathbf{1}_{n\times n}$ sont des matrices carrés de dim n ; la 1ère est de rang 1 et la 2nde de rang $n-1$.
- Ces matrices carrées sont *idempotentes* ($\mathbf{A}\times\mathbf{A} = \mathbf{A}$) et symétriques ($\mathbf{A}^{\mathrm{T}} = \mathbf{A}$) : ce sont des *matrices de projection*.

- En transposant la dernière équation, on a :

$$\mathbf{X_C} = \left[\mathbf{I}_n - \frac{1}{n}\mathbf{1}_{n\times n}\right]\mathbf{X}$$

3. Résumés multidimensionnels

Première transformation : matrice de dispersion

- *Matrice de variances-covariances*

$$\mathbf{S}_{p\times p} = \mathbf{X_C}^{\mathrm{T}}\mathbf{D}\mathbf{X_C} = \begin{bmatrix} s_1^2 & s_{12} & \cdots & \\ & s_2^2 & & \cdots \\ \vdots & & \ddots & \cdots \\ & & \cdots & s_p^2 \end{bmatrix}$$

où $\mathbf{D}$ est la **matrice carrée** $(n\times n)$ **diagonale des poids**.

- *Si toutes les obs ont la même importance*, elles ont **même *poids***, i.e. $1/n$. On a bien sûr : $\mathbf{D} = \mathbf{I}_n/n$
- *Quelquefois*, il peut être utile de leur donner des **poids** $\neq$, p_i pour l'obs i . Ces poids sont des nbs > 0 de somme à 1 :

$$\mathbf{D} = \begin{bmatrix} p_1 & 0 & \cdots & 0 & 0 \\ 0 & p_2 & 0 & \cdots & 0 \\ \vdots & & \ddots & & \cdots \\ 0 & 0 & \cdots & 0 & p_n \end{bmatrix} \text{ avec } \sum_{i=1}^{n} p_i = 1$$

- *Avec pondération*, le *vecteur moyenne* est alors calculé par :

$$\bar{\boldsymbol{x}} = \mathbf{X}^{\mathrm{T}}\mathbf{D}\mathbf{1}_{n\times 1} \in \mathbb{R}^p \quad \text{et} \quad \mathbf{X_C} = [\mathbf{I}_n - \mathbf{1}_{n\times n}\mathbf{D}]\mathbf{X}$$

3. Résumés multidimensionnels

Première transformation : matrice de dispersion

- *Matrice de variances-covariances*

- On a alors la **matrice centrée réduite** par :

$$\mathbf{X_{CR}} = \mathbf{X_C}\big(\mathbf{diag(S)}\big)^{-1/2}$$

- *Matrice des coeff. de corrélation linéaire* $\mathbf{R}$, carrée de dim p

$$\mathbf{R} = \mathbf{D}_{1/s}\mathbf{S}\mathbf{D}_{1/s} = \mathbf{X_{CR}}^{\mathrm{T}}\mathbf{D}\mathbf{X_{CR}}$$

Où

$\mathbf{D}_{1/s}$ est la matrice diagonale définie par $\mathbf{D}_{1/s} = \big(\mathbf{diag(S)}\big)^{-1/2}$.

- Sous R : avec 1/n-1 pour le calcul et non 1/n comme en AD
  - `var( ) ; cor( )`
  - `scale(x, center = TRUE, scale = TRUE)`

3. Résumés multidimensionnels

- *Matrice des coeff. de corrélation linéaire* $\mathbf{R}$, de dim p

$$\mathbf{R} = \mathbf{D}_{1/s}\mathbf{S}\mathbf{D}_{1/s} = \mathbf{X_{CR}}^{\mathrm{T}}\mathbf{D}\mathbf{X_{CR}}$$

- Le degré de liaison entre 2 variables quantitatives se mesure à l'aide du **coefficient de corrélation**.
  D'autant plus voisin de +1 ou –1 que la liaison est étroite.
- **Attention ! Une corrélation n'est pas forcément une relation de cause à** effet.
- **Attention ! L'absence de corrélation ne signifie pas que les variables ne** sont pas liées. Il peut exister des relations non linéaires.
- **Remarque sur** $\mathbf{S}$ et $\mathbf{R}$: matrices carrées d'ordre p, symétriques et def. positives $\Rightarrow$ valeurs propres ≥ 0.

3. Résumés multidimensionnels

- **Liaison entre 2 variables qualitatives : exemple** tiré de Analyses multidimensionnelles. INRA formation permanente. Janvier 2006. André Bouchier

  - Nous avons artificiellement découpé la population de l'INRA en 4 types professionnels :

    A pour administratifs
    I pour ingénieurs
    S pour scientifiques
    T pour techniciens

  - Nous avons croisé ce type professionnel avec le sexe des agents INRA (ramené à 50/50 pour faciliter la compréhension). Nous avons obtenu le **tableau de contingence** suivant :

	A	I	S	T	Total
F	15	9	2	24	50
H	0	5	15	30	50
Total	15	14	17	54	100

3. Résumés multidimensionnels

- **Liaison entre 2 variables qualitatives : exemple**

Exemple : homogénéité de 2 échantillons de la population de l'INRA

Tableau des effectifs

	A	I	S	T	Total
F	15	9	2	24	50
H	0	5	15	30	50
Total	15	14	17	54	100

Tableau des effectifs ajustés (fréquences théoriques)

	A	I	S	T	Total
F	7.5	7	8.5	27	50
H	7.5	7	8.5	27	50
Total	15	14	17	54	100

3. Résumés multidimensionnels

- **Liaison entre 2 variables qualitatives : exemple**

  - Le tableau des écarts à l'indépendance, (fréquences observées - fréquences théoriques) contient l'information « intéressante »

Tableau des écarts à l'indépendance
(fréq obs - fréq théorique)

	A	I	S	T	*Total*
F	7.5	+2	-6.5	-3	*0*
H	-7.5	-2	+6.5	+3	*0*
Total	*0*	*0*	*0*	*0*	*0*

3. Résumés multidimensionnels

- **Liaison entre 2 variables qualitatives : exemple**

  - On calcule la contribution du khi2 pour chaque cellule du tableau.

$$Khi2 = \frac{(FréqObservée - FréqThéorique)^2}{FréqThéorique}$$

Tableau des contributions au Khi2

	A	I	S	T
F	7.5	0.57	4.97	0.33
H	7.5	0.57	4.97	0.33

  - Le Khi2 est ici égal à 26.74. Mais ce chiffre n'a pas de signification en lui même. En effet, plus le tableau sera grand, plus le Khi2 sera élevé.

3. Résumés multidimensionnels

- **Liaison entre 2 variables qualitatives : exemple**

- On calcul les contributions en pourcentage de chaque cellule au Khi2

	A	I	S	T	Total
F	28,05	2,13	18,59	1,23	50%
H	28,05	2,13	18,59	1,23	50%
Total	56,10	4,26	37,17	2,47	100%

- On remarque par exemple que la contribution maximale est apportée par la catégorie A.

- Plus de 93% de l'inertie du tableau est due aux catégories A et S. Concrètement, ça veut dire quoi ?

3. Résumés multidimensionnels

Espaces de représentation

Interprétation **géométrique** des lignes et les colonnes du tableau $\mathbf{X} \in \mathcal{M}_{n\times p}$ par des points dans 2 espaces différents:

l'espace des individus et **l'espace des variables** :

- *les espaces de représentation* :
  - **celui des p variables,** de dim n, noté $\Re^n \subset \mathbb{R}^n$
    - Les p colonnes sont considérées comme p pts de **l'espace des variables** à n dimensions.

    2 variables sont proches si leurs n coordonnées sont très voisines (i.e. ces variables mesurent la même chose ou sont liées par une relation particulière).
  - **celui des n individus,** de dim p, noté $\Re^p \subset \mathbb{R}^p$
    - Les n lignes mis en colonne sont considérées comme n pts de **l'espace des individus** à p dimensions.

    2 points sont très proches si les p coord. de ces 2 pts sont très proches (mêmes valeurs pour les différentes variables).

3. Résumés multidimensionnels

L'espace des individus : regard sur les lignes, inertie

L'analyse du nuage de points utilise la notion fondamentale de **distance**.

- **En physique**, on munit l'espace des individus de la distance euclidienne classique:

$$d^2(\mathbf{x}_1, \mathbf{x}_2) = (x_1^1 - x_2^1)^2 + \cdots + \left(x_1^p - x_2^p\right)^2 = (\boldsymbol{x}_1 - \boldsymbol{x}_2)^{\mathrm{T}}(\boldsymbol{x}_1 - \boldsymbol{x}_2)$$

- **En statistique**, pb souvent un peu plus compliqué car les unités de mesure sont ≠ ; on peut avoir un mélange de mensurations, de poids et d'âge…Il faut donc **pondérer**, d'une certaine manière, les différents carrés de l'expression précédente.

On choisit $\mathbb{R}^p$ muni de la ***métrique*** $\mathbf{Q} \in \mathcal{M}_{p\times p}$ (matrice carrée, sym def. pos.), lui conférant une structure d'espace euclidien :

Produit scalaire : $\langle \boldsymbol{x}_1; \boldsymbol{x}_2 \rangle_{\mathbf{Q}} = \boldsymbol{x}_1{}^{\mathrm{T}}\mathbf{Q}\boldsymbol{x}_2$

Carré de la distance : $d(\boldsymbol{x}_1; \boldsymbol{x}_2) = (\boldsymbol{x}_1 - \boldsymbol{x}_2)^{\mathrm{T}}\mathbf{Q}(\boldsymbol{x}_1 - \boldsymbol{x}_2)$

$\mathbf{Q}$-*norme de* $\boldsymbol{x}_1$: $\|\boldsymbol{x}_1\|_{\mathbf{Q}}^2 = \boldsymbol{x}_1{}^{\mathrm{T}}\mathbf{Q}\boldsymbol{x}_1$

3. Résumés multidimensionnels

L'espace des individus : regard sur les lignes, inertie

- ***Matrice produit scalaire*** **entre individus** : $\boxed{\mathbf{W} = \mathbf{X}\mathbf{Q}\mathbf{X}^{\mathrm{T}}}$
  - $\mathbf{Q} = \mathbf{I}_p$: ***produit scalaire usuel.*** Les unités initiales de chaque variable sont conservées
    - revient à rendre dominantes les variables de plus grande variabilité.
  - $\mathbf{Q} = \left(\mathbf{diag}(\mathbf{S})\right)^{-1}$: ***métrique inverse des variances.***
    - Chaque var est divisée par son écart-type ; son importance devient indep. de sa dispersion.
    - **distances calculées** non plus sur $\mathbf{X}$, mais **sur la matrice centrée réduite** $\mathbf{X}_{\mathbf{CR}}$.
  - Utiliser une métrique diagonale quelconque $\mathbf{Q} = \mathbf{T}^{\mathrm{T}}\mathbf{T}$ car sym. pos.…

3. Résumés multidimensionnels

L'espace des individus : regard sur les lignes, inertie

Matrice produit scalaire entre individus : $\mathbf{W} = \mathbf{X}\mathbf{Q}\mathbf{X}^{\mathrm{T}}$

Propriété *(Factorisation de Cholesky)* : $\mathbf{Q}$ étant symétrique, def. positive, on peut écrire $\mathbf{Q} = \mathbf{T}^{\mathrm{T}}\mathbf{T}$ avec $\mathbf{T}$ matrice triangulaire sup., d'où

$$\langle \boldsymbol{x}_i ; \boldsymbol{x}_{i'} \rangle_{\mathbf{Q}} = (\mathbf{T}\boldsymbol{x}_i)^{\mathrm{T}}\mathbf{T}\boldsymbol{x}_{i'}$$

Tout se passe comme si on utilisait la métrique $\mathbf{Q} = \mathbf{I}_p$ sur le tableau $\mathbf{X}\mathbf{T}^{\mathrm{T}}$.

Utiliser la métrique $\mathbf{Q} = \left(\mathbf{diag}(\mathbf{S})\right)^{-1}$, sur le tableau centré $\mathbf{X}_{\mathrm{C}}$

$\Leftrightarrow$

utiliser la métrique $\mathbf{Q} = \mathbf{I}_p$ sur le tableau centré réduit $\mathbf{X}_{\mathrm{CR}}$.

3. Résumés multidimensionnels

L'espace des individus : regard sur les lignes, inertie

- *L'inertie totale du nuage de points* : I_g

$\Rightarrow$ *L'inertie* I_g *permet de* quantifier la *variabilité* contenue dans un tableau de données.

- On appelle **inertie la quantité d'information contenue dans un tableau de** données.
- Une **inertie nulle signifie que tous les individus sont presque identiques.**
- L'**inertie du nuage sera égale à la somme des variances des** p **variables.**
- Si les p variables sont centrées-réduites, l'**inertie** sera égale **à** p.

Soit sous forme mathématique !

3. Résumés multidimensionnels

L'espace des individus : regard sur les lignes, inertie

- *L'inertie totale du nuage de points* : I_g

Définition : L'**inertie totale**, notée I_g, est *la* ***moyenne des carrés des distances pondérées*** *des* n *pts au centre de gravité* $\overline{\boldsymbol{x}}$. Soit :

$$I_g = \sum_{i=1}^{n} p_i \|\boldsymbol{x}_{\mathrm{i}} - \overline{\boldsymbol{x}}\|_{\mathbf{Q}}^2 = tr\left(\sum_{i=1}^{n} p_i (\boldsymbol{x}_{\mathrm{i}} - \overline{\boldsymbol{x}})^{\mathrm{T}}\mathbf{Q}\,(\boldsymbol{x}_{\boldsymbol{i}} - \overline{\boldsymbol{x}})\right) = tr(\mathbf{QS})$$

- Cette quantité caractéristique du nuage mesure *l'éloignement des points par rapport à leur centre de gravité* :

Ex : I_g proche de $0 \Rightarrow$ les individus sont identiques ou presque et sont confondus avec leur centre de gravité.

3. Résumés multidimensionnels

L'espace des individus : regard sur les lignes, inertie

- *L'inertie totale du nuage de points* : I_g

Propriétés :

$$I_g = tr(\mathbf{QS}) = tr(\mathbf{SQ}) = tr(\mathbf{WD}) = tr(\mathbf{DW})$$

où

$\mathbf{W} = \mathbf{X}\mathbf{Q}\mathbf{X}^{\mathrm{T}}$, m*atrice produit scalaire* entre individus et $\mathbf{S} = \mathbf{X_C}^{\mathrm{T}}\mathbf{D}\mathbf{X_C}$, matrice des variances-covariances.

D'où si :

- $\mathbf{Q} = \mathbf{I}_p$, l'inertie est égale à la **somme des** p **variances**,
- $\mathbf{Q} = \left(\mathbf{diag}(\mathbf{S})\right)^{-1}$, l'inertie est égale à p, le nombre de variables.

3. Résumés multidimensionnels

L'espace des individus : regard sur les lignes, inertie

Matrice d'inertie

Définition :

Pour une métrique qcque $\mathbf{Q} \in \mathcal{M}_{p\times p}$, on appelle **matrice d'inertie de** $\mathbf{X}$, la matrice

$$\mathbf{S}_\mathrm{I} = \mathbf{X}^\mathrm{T}\mathbf{D}\mathbf{X}\mathbf{Q}$$

- Pour $\mathbf{Q} = \mathbf{I}_p$, on a : $\mathbf{S}_\mathrm{I} = \mathbf{S} = \mathbf{X}^\mathrm{T}\mathbf{D}\mathbf{X}$, matrice de variances-covariance, si $\mathbf{X} = \mathbf{X}_\mathbf{C}$ est centrée.
- *Interprétation de* $\mathbf{S}_\mathrm{I}$ *:*

$\mathbf{S}_\mathrm{I}$ est symétrique et en gal def. positive.

Elle définit une forme quadratique et pour $\boldsymbol{u}$ donné, tq $\boldsymbol{u}^\mathrm{T}\boldsymbol{u} = 1$, on a $\boldsymbol{u}^\mathrm{T}\mathbf{X}^\mathrm{T}\mathbf{D}\mathbf{X}\boldsymbol{u}$ qui vaut l'inertie du nuage projeté sur $\mathbb{R}\boldsymbol{u}$ (cf cours ACP, $\mathbf{X} = \mathbf{X}_\mathbf{C}$ et $\mathbf{Q} = \mathbf{I}_p$) .

3. Résumés multidimensionnels

L'espace des individus : regard sur les lignes, inertie

- *L'inertie totale du nuage de points* : ***Propriétés***

Propriété : L'inertie totale I_g est la moitié de la moyenne pondérée des carrés des distances entre les individus:

$$2I_g = \sum_{i=1}^{i=n}\sum_{i'} p_i\, p_{i'} \left\|\mathbf{x_i} - \mathbf{x_{i'}}\right\|_\mathbf{Q}^2 .$$

Dem : Soit N un nuage de points $\{(\mathbf{x}_i,p_i)_{i=1,\dots,n}\}$ de c.d.g. $\bar{\boldsymbol{x}}$

$$\left\|\mathbf{x_i} - \mathbf{x_{i'}}\right\|_\mathbf{Q}^2 = \left\|\mathbf{x_i} - \bar{\mathbf{x}} + \bar{\mathbf{x}} - \mathbf{x_{i'}}\right\|_\mathbf{Q}^2$$

$$= \left\|\mathbf{x_i} - \bar{\mathbf{x}}\right\|_\mathbf{Q}^2 + \left\|\bar{\mathbf{x}} - \mathbf{x_{i'}}\right\|_\mathbf{Q}^2 + 2\left\langle \mathbf{x_i} - \bar{\mathbf{x}}, \bar{\mathbf{x}} - \mathbf{x_{i'}} \right\rangle_\mathbf{Q}$$

D'où comme $\sum_i p_i\left(\mathbf{x_i} - \bar{\mathbf{x}}\right) = \mathbf{0}$

$$\sum_{i=1}^{i=n}\sum_{i'} p_i\, p_{i'} \left\|\mathbf{x_i} - \mathbf{x_{i'}}\right\|_\mathbf{Q}^2 = 2\sum_{i=1}^{n} p_i \left\|\mathbf{x_i} - \bar{\mathbf{x}}\right\|_\mathbf{Q}^2 = 2I_g .$$

3. Résumés multidimensionnels

L'espace des individus : regard sur les lignes, inertie

- *L'inertie totale du nuage de points* : ***Propriétés***

Propriété:

$I_g = tr(\mathbf{X}^\mathbf{T}\mathbf{D}\mathbf{X}\mathbf{Q}) = tr(\boldsymbol{S}_I) = ||\boldsymbol{X}||_{\boldsymbol{QxD}}$ norme de Hilbert-Schmidt de $\mathbf{X}$

$= tr(\mathbf{Q}\mathbf{S}) = tr(\mathbf{S}\mathbf{Q})$

$= tr(\mathbf{W}\mathbf{D}) = tr(\mathbf{D}\mathbf{W})$ si $\mathbf{X}$ est centrée

où

$\mathbf{W} = \mathbf{X}\mathbf{Q}\mathbf{X}^\mathrm{T}$, *matrice produit scalaire* entre individus et

$\mathbf{S} = \mathbf{X}^\mathrm{T}\mathbf{D}\mathbf{X}$, matrice des var.- cov.

D'où si :

- $\mathbf{Q} = \mathbf{I}_p$, l'inertie est égale à la **somme des** p **variances**,
- $\mathbf{Q} = \boldsymbol{diag}\{\mathbf{S}\}^{-1}$, l'inertie est égale à p, le nbre de variables.

dem : si $\mathbf{X}$ est centrée

3. Résumés multidimensionnels

L'espace des individus : regard sur les lignes, inertie

- *L'inertie totale du nuage de points* : ***Propriétés***

Dem : si $\mathbf{X}$ est centrée i.e. $\mathbf{X} = \mathbf{X}_\mathrm{C}$

$$I_g = \sum_{i=1}^{i=n} p_i \left\|\mathbf{x_i}\right\|_\mathbf{Q}^2 = \sum_{i=1}^{i=n} p_i \left(\mathbf{x_i}\right)^T \mathbf{Q}\left(\mathbf{x_i}\right) : \text{si } \lambda \text{ nb réel positif } tr\left(\lambda\right) = \lambda \text{ d'où}$$

$$= \sum_{i=1}^{i=n} tr\left(p_i \left(\mathbf{x_i}\right)^T \mathbf{Q}\left(\mathbf{x_i}\right)\right) = \sum_{i=1}^{i=n} tr\left(p_i \mathbf{Q}\left(\mathbf{x_i}\right)\left(\mathbf{x_i}\right)^T\right) : \text{or } \mathrm{tr}(\mathbf{AB}) = \mathrm{tr}(\mathbf{BA}) \text{ d'où ;}$$

$$= tr\left(\sum_{i=1}^{i=n} p_i \left(\mathbf{x_i}\right)^T \mathbf{Q}\left(\mathbf{x_i}\right)\right) = tr\left(\mathbf{Q}\mathbf{X}^\mathbf{T}\mathbf{D}\mathbf{X}\right) = tr\left(\mathbf{X}^\mathbf{T}\mathbf{D}\mathbf{X}\mathbf{Q}\right).$$

3. Résumés multidimensionnels

L'espace des individus : regard sur les lignes, inertie

- *L'inertie totale du nuage de points* : **Propriétés**

Théorème de Huyghens :

Si on note:

$$I_a = \sum_{i=1}^{n} p_i (\mathbf{x_i} - \mathbf{a})^T \mathbf{Q} (\mathbf{x_i} - \mathbf{a}) = \sum_{i=1}^{n} p_i \|\mathbf{x_i} - \mathbf{a}\|_{\mathbf{Q}}^2$$

l'inertie du nuage de points par rapport à un pt a quelconque
$\Rightarrow$ l'inertie totale du nuage de pts I_g vérifie :

$$\mathrm{I_a} = \mathrm{I_g} + \|\overline{\mathbf{x}} - \mathbf{a}\|_{\mathbf{Q}}^2$$

3. Résumés multidimensionnels

L'espace des individus : regard sur les lignes, inertie

- *L'inertie totale du nuage de points*

$\Rightarrow$ **projection des individus de $\mathbb{R}^p$ dans un ss-espace de dim $< p$**

Principe ACP (cf ch. 2) = obtenir une *représentation approchée* du nuage des n individus dans un ss-espace de faible dim $k < p$.

=> Pour y parvenir, **projection** des individus sur un ss-espace de dimension plus faible (= $k < p$), ss-espace de projection choisi selon le critère suivant:

Les distances en projection devront être le moins déformées possible.

3. Résumés multidimensionnels

L'espace des variables : regard sur les colonnes

- Pour étudier la proximité des variables, il faut munir cet espace d'une **métrique** ie trouver une **matrice carrée d'ordre *n* sym. def. pos.**
- On choisit la **matrice diagonale des poids D** (matrice *nxn*, sym. def. pos.), lui conférant une structure d'espace euclidien : $x^j \in \mathbb{R}^n$
- Nous supposons les x^j **centrées** :

Produit scalaire entre 2 variables : $\left\langle \mathbf{x}^j ; \mathbf{x}^k \right\rangle_{\mathbf{D}} = \mathbf{x}^{j^T} \mathbf{D} \mathbf{x}^k = \sum_{i=1}^{n} p_i x_i^k x_i^j$

Norme d'une variable : $\|\mathbf{x}^j\|_{\mathbf{D}}^2 = s_j^2$
ie la « longueur » d'une variable est égale à son écart-type.

Dans un espace euclidien, on définit l'angle θ entre 2 vecteurs par son cosinus de la façon suivante:

$$\cos(\theta_{jk}) = \frac{\left\langle \mathbf{x}^j ; \mathbf{x}^k \right\rangle_{\mathbf{D}}}{\|\mathbf{x}^j\|_{\mathbf{D}} \|\mathbf{x}^k\|_{\mathbf{D}}} = \frac{s_{jk}}{s_j s_k} = r_{jk}$$

Le cosinus de l'angle entre 2 variables centrées est égal à leur coefficient de corrélation linéaire.

3. Résumés multidimensionnels

Définition d'une étude

On appelle **étude** un **triplet** $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$ où

- $\mathbf{X}$ est **tableau de données** croisant individus et variables ;
- $\mathbf{Q}$ est la métrique permettant le calcul des distances entre individus ;
- $\mathbf{D}$ est la métrique des poids, permettant le calcul des distances entre variables.

↪ Une étude peut être caractérisée ou représentée par deux « **objets** » différents :

$$\mathbf{W} = \mathbf{X}\mathbf{Q}\mathbf{X}^{\mathrm{T}} \text{ ou } \mathbf{S} = \mathbf{X}^{\mathrm{T}}\mathbf{D}\mathbf{X}$$

3. Résumés multidimensionnels

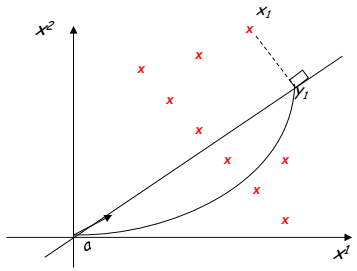
Transformation linéaire d'un tableau de données

- Les analyses que nous allons étudier ultérieurement consistent, dans la majorité des cas, à **trouver des combinaisons linéaires des p variables** définissant les colonnes de $\mathbf{X}$.
- Nous chercherons des **combinaisons linéaires** qui ont des **propriétés particulières**.
  - Ces combinaisons linéaires seront des transformations des vecteurs $\mathbf{x}^j (j = 1,\ldots, p)$ qui vont fournir de nouvelles variables.

3. Résumés multidimensionnels

Transformation linéaire d'un tableau de données

- A chaque vecteur d'observation $\boldsymbol{x} \in \mathbb{R}^p$, on peut associer un **axe** de l'espace des observations $\mathcal{R}^p$; une ***transformation linéaire*** consiste à **projeter les observations sur un nouvel axe de *vecteur unitaire*** $\boldsymbol{a} \in \mathbb{R}^p$.



3. Résumés multidimensionnels

Transformation linéaire d'un tableau de données

- Si on a i $(i = 1, \ldots, n)$ observations, la **coordonnée de l'obs i sur cet axe** est définie par :

$$y_i = \boldsymbol{a}^{\mathrm{T}}\mathbf{Q}(\boldsymbol{x}_i - \overline{\boldsymbol{x}}) = \langle \boldsymbol{a}; \boldsymbol{x}_i - \overline{\boldsymbol{x}} \rangle_{\mathbf{Q}} \in \mathbb{R}$$

- L'ensemble des n coordonnées par :

$$\boldsymbol{y} = \mathbf{X}\mathbf{Q}\boldsymbol{a} = \mathbf{X}\boldsymbol{u} = \sum_{j=1}^{p} \boldsymbol{x}^j u_j \in \mathbb{R}^n \text{ avec } \boldsymbol{u} = \mathbf{Q}\boldsymbol{a}$$

- Nous avons donc créé une ***nouvelle variable*** $\boldsymbol{y}$ avec :
  - un **axe** de vecteur unitaire $\boldsymbol{a}$,
  - un vecteur $\boldsymbol{y}$ de l'espace des variables, **composante principale**
  - une forme linéaire $\boldsymbol{u}$ qui est appelée ***facteur***.

3. Résumés multidimensionnels

Transformation linéaire d'un tableau de données

- Si nous faisons q **transformations de l'ensemble du tableau** définies par la matrice $\mathbf{A} = [\boldsymbol{a}^1 \quad |\ldots| \quad \boldsymbol{a}^q]$ tq :

$$\mathbf{Y}_{n\times q} = \mathbf{X}_{n\times p}\mathbf{Q}_{p\times p}\mathbf{A}_{p\times q}$$

- Supposons ici que $\mathbf{Q} = \mathbf{I}_p$
  - On peut montrer que la **matrice de dispersion** $\mathbf{S}_{\mathbf{Y}}$ de $\mathbf{Y}$ se déduit de celle de $\mathbf{X}$ par :

$$\mathbf{S}_{\mathbf{Y};q\times q} = \mathbf{A}^{\mathrm{T}}_{q\times p}\mathbf{S}_{\mathbf{X};p\times p}\mathbf{A}_{p\times q}$$

3. Résumés multidimensionnels

Transformation linéaire d'un tableau de données

- Il est souvent utile **d'imposer des conditions à la matrice de transformation** $\mathbf{A}$:
  - les q vecteurs sont ***unitaires*** : $\left\|\mathbf{a}^k\right\|^2 = 1; k = 1, \cdots, q$
  - les q vecteurs sont ***orthogonaux 2 à 2*** : $\left(\mathbf{a}^k\right)^{\mathrm{T}} \mathbf{a}^l = 0; k \neq l = 1, \cdots, q$

- La matrice $\mathbf{A}$ est une ***matrice orthogonale*** si elle est carrée (c'est-à-dire ici si $q = p$) et $\mathbf{A}\,\mathbf{A}^{\mathrm{T}} = \mathbf{A}^{\mathrm{T}}\mathbf{A} = \mathbf{I}_q$.

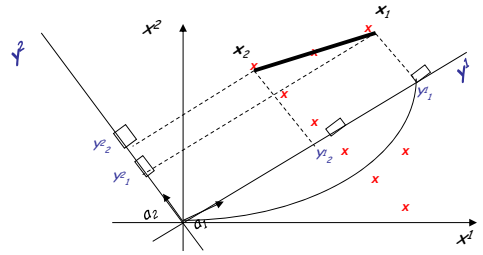
- quand $q = p$, la transformation est une simple ***rotation*** du repère de départ (avec $\mathbf{Q} = \mathbf{I}_p$) :
  - $\boldsymbol{x}_i \in \mathbb{R}^p$ transformée en $y_i = \mathbf{A}^{\mathrm{T}}\boldsymbol{x}_i$; $\forall i = 1, \ldots, n$
  - Chaque composante du nouveau vecteur est la ***projection*** de $\boldsymbol{x}_i$ sur les p vecteurs $\boldsymbol{a}^j$; $j = 1, \ldots, p$

3. Résumés multidimensionnels

Transformation linéaire d'un tableau de données

- ***Propriété intéressante et importante :***
  La distance entre 2 obs est invariante ds une transformation linéaire :

$$(\mathbf{y}_r - \mathbf{y}_s)^T(\mathbf{y}_r - \mathbf{y}_s) = (\mathbf{x}_r - \mathbf{x}_s)^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{A}(\mathbf{x}_r - \mathbf{x}_s) = (\mathbf{x}_r - \mathbf{x}_s)^{\mathrm{T}}(\mathbf{x}_r - \mathbf{x}_s)$$



3. Résumés multidimensionnels

Décomposition d'une matrice de données $\mathbf{X}$

La plupart des méthodes **d'Analyse Factorielle des Données** peuvent être présentées dans un cadre commun :

celui de **l'extension du théorème de la DVS au cadre d'espaces Euclidiens généraux**.

Elle correspond à la DVS du triplet $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$

3. Résumés multidimensionnels

Décomposition d'une matrice de données $\mathbf{X}$

- ***Rappel : Décomposition en valeurs singulières (DVS)***

Version « maigre » : Toute matrice $\mathbf{X}_{n\times p}$ peut être décomposée de la façon suivante :

$$\mathbf{X}_{n\times p} = \mathbf{U}_{n\times r}\mathbf{\Theta}_{r\times r}\left(\mathbf{V}_{p\times r}\right)^{\mathrm{T}}$$

Où

- $\mathbf{\Theta}_{r\times r} = \mathbf{\Lambda}^{1/2}$ matrice **diagonale** contenant les r **valeurs singulières** de la matrice $\mathbf{X}$; *i.e.* les racines carrées des v.p. **positives et non nulles** λ_i de $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ ou de $\mathbf{X}\mathbf{X}^{\mathrm{T}}$; $\lambda_1 \geq \lambda_2 \geq \cdots \lambda_r > 0$
- $r = rg(\mathbf{X}) \leq min(n, p)$. Si $r = p$, la matrice $\mathbf{X}$ est dite de ***plein rang***.
- $\mathbf{U}_{n\times r} = [\boldsymbol{u}^1 \quad \ldots \quad \boldsymbol{u}^r]$ unitaire $n \times r$ et tq $\boldsymbol{u}^i$ est vecteur propre de $\mathbf{X}\mathbf{X}^{\mathrm{T}}$ associé à la valeur propre non nulle λ_i

  $\Rightarrow$ *Vecteurs principaux de l'espace des colonnes*
- $\mathbf{V}_{p\times r} = [\boldsymbol{v}^1 \quad \ldots \quad \boldsymbol{v}^r]$ unitaire $p \times r$ et tq $\boldsymbol{v}^i$ est vecteur propre de $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ associé à la valeur propre non nulle λ_i

  $\Rightarrow$ *Vecteurs principaux de l'espace des lignes*
- **U et V sont des matrices colonnes-orthonormales** (i.e. matrices contenant des vecteurs colonnes unitaires et orthogonaux 2 à 2).

3. Résumés multidimensionnels

Décomposition d'une matrice données $\mathbf{X}$

Schéma de dualité

- $\mathbf{Q}$ est la matrice d'un produit scalaire de $\mathbb{R}^p$ (*L'espace des individus)*, soit une matrice carrée symétrique def. pos. qui définit la fonction :

$$(\mathbf{x},\mathbf{y})=\left(\begin{bmatrix}x_1\\ \vdots\\ x_p\end{bmatrix},\begin{bmatrix}y_1\\ \vdots\\ y_p\end{bmatrix}\right)\in R^p\times R^p\mapsto \mathbf{x}^{\mathrm{T}}\mathbf{Q}\mathbf{y}=\langle\mathbf{x},\mathbf{y}\rangle_{\mathbf{Q}}\in R$$

- $\mathbf{D}$ est la matrice d'un produit scalaire de $\mathbb{R}^n$ (*L'espace des variables)*, soit une matrice carrée symétrique def. pos. qui définit la fonction :

$$(\mathbf{x},\mathbf{y})=\left(\begin{bmatrix}x_1\\ \vdots\\ x_n\end{bmatrix},\begin{bmatrix}y_1\\ \vdots\\ y_n\end{bmatrix}\right)\in R^n\times R^n\mapsto \mathbf{x}^{\mathrm{T}}\mathbf{D}\mathbf{y}=\langle\mathbf{x},\mathbf{y}\rangle_{\mathbf{D}}\in R$$

- L'essentiel est que ces 4 matrices $\mathbf{X}$, $\mathbf{X}^{\mathrm{T}}$, $\mathbf{Q}$ et $\mathbf{D}$ s'assemblent car les produits $\mathbf{XQ}$, $\mathbf{QX}^{\mathrm{T}}$, $\mathbf{X}^{\mathrm{T}}\mathbf{D}$ et $\mathbf{DX}$ ont un sens. On les dispose dans un schéma dit ***schéma de dualité*** :

4. Dualité en analyse de données

Décomposition d'une matrice données

Schéma de dualité

Cf pour plus de détails http://pbil.univ-lyon1.fr/R/cours/stage3.pdf

- Idée d'origine de P. Cazes , il est appelé analyse générale par Greenacre (1984, Annexe A2 p. 346) ; introduit en écologie par Escoufier (1987).
- Un schéma est ***constitué de trois éléments*** donnant un triplet $(\mathbf{X},\mathbf{Q},\mathbf{D})$.
- $\mathbf{X}$ est une ***matrice de données*** en gl issue d'une matrice de données brutes $\widetilde{\mathbf{X}}$ à l'aide d'une transformation préalable. $\mathbf{X}$ a n lignes et p colonnes.
  - Les n lignes de $\mathbf{X}$ sont des vecteurs de $\mathbb{R}^p$ tandis que
  - les p colonnes de $\mathbf{X}$ sont des vecteurs de $\mathbb{R}^n$.

4. Dualité en analyse de données

Décomposition d'une matrice données

Schéma de dualité

- On les dispose dans un schéma dit ***schéma de dualité*** :

$$\begin{array}{ccc} R^p & \xrightarrow{\ \mathbf{Q}\ } & R^{p*} \\ \mathbf{X}^{\mathrm{T}}\uparrow & & \downarrow\mathbf{X} \\ R^{n*} & \xleftarrow[\ \mathbf{D}\ ] & R^n \end{array}$$

- Rigoureusement, $\mathbb{R}^{p*}$ est le dual de $\mathbb{R}^p$ (ensemble des applications linéaires de $\mathbb{R}^p$ dans $\mathbb{R}$), $\mathbb{R}^{n*}$ est le dual de $\mathbb{R}^n$ (ensemble des applications linéaires de $\mathbb{R}^n$ dans $\mathbb{R}$),
- $\mathbf{Q}$ est vue comme la matrice d'une application linéaire définie par $(\mathbf{Q}(\boldsymbol{x}))(\boldsymbol{y}) =< \boldsymbol{x},\boldsymbol{y} >_{\mathbf{Q}}$,
- $\mathbf{D}$ est vue comme la matrice d'une application linéaire définie par $(\mathbf{D}(\boldsymbol{x}))(\boldsymbol{y}) =< x,y >_{\mathbf{D}}$.

4. Dualité en analyse de données

Décomposition d'une matrice données

Schéma de dualité

- Pour l'utilisateur, la simplification :

$$(\mathbf{X},\mathbf{Q},\mathbf{D})\Leftrightarrow\begin{array}{ccc} [p] & \xrightarrow{\ \mathbf{Q}\ } & [p] \\ \mathbf{X}^{\mathrm{T}}\uparrow & & \downarrow\mathbf{X} \\ [n] & \xleftarrow[\ \mathbf{D}\ ] & [n] \end{array}$$

suffit pour se souvenir que les produits de matrices :

$\mathbf{X}^{\mathrm{T}}\mathbf{DXQ}$, $\mathbf{DXQX}^{\mathrm{T}}$, $\mathbf{XQX}^{\mathrm{T}}\mathbf{D}$ et $\mathbf{QX}^{\mathrm{T}}\mathbf{DX}$

ont un sens.

4. Dualité en analyse de données

Décomposition d'une matrice données

Schéma de dualité - DVS du triplet $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$

- On définit deux matrices $\mathbf{S} = \mathbf{X}^{\mathrm{T}}\mathbf{D}\mathbf{X}$ et $\mathbf{W} = \mathbf{X}\mathbf{Q}\mathbf{X}^{\mathrm{T}}$ qui s'insèrent par :

  $\mathbf{S}\mathbf{Q} = \mathbf{S}_{\mathbf{I}}$ et $\mathbf{W}\mathbf{D}$ sont les opérateurs d'Escoufier.
- Le schéma a des ***propriétés théoriques*** très gales qui prennent des significations propres lorsqu'on utilise un ensemble de paramètres originaux.
- En particulier les ACP classiques, l'analyse des correspondances (AFC), les analyses de correspondances non symétriques (ANSC), les analyses discriminantes (AFD), l'analyse canonique (AC), les analyses de co-inertie (ACO), les analyses sur variables instrumentales (ACPVI), l'analyse des correspondances multiples (ACM), l'analyse factorielle multiple (AFM) et diverses extensions sont des **conséquences directes des propriétés générales**.

4. Dualité en analyse de données

Décomposition d'une matrice de données $\mathbf{X}$

DVS de $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$

Soient les matrices réelles $\mathbf{X}_{n\times p}$ de rang r et les métriques $\mathbf{Q}_{p\times p}$ et $\mathbf{D}_{n\times n}$ de $\mathcal{R}^p$ et de $\mathcal{R}^n$. Il existe :

- Une matrice $\mathbf{U}_{p\times r} = [\boldsymbol{u}_1 \quad \dots \quad \boldsymbol{u}_r]$ $\mathbf{Q}$-orthonormée

les colonnes $\boldsymbol{u}_i$ sont les vecteurs propres de $\mathbf{X}^{\mathrm{T}}\mathbf{D}\mathbf{X}\mathbf{Q} = \mathbf{S}\mathbf{Q} = \mathbf{S}_{\mathrm{I}}$ associés aux valeurs propres $\lambda_1 \geq \lambda_2 \geq \cdots \lambda_r > 0$.

$\Longrightarrow$ *Vecteurs principaux de l'espace des indiv.*

- Une matrice $\mathbf{V}_{n\times r} = [\boldsymbol{v}_1 \quad \dots \quad \boldsymbol{v}_r]$ $\mathbf{D}$-orthonormée

les colonnes st les vecteurs propres de $\mathbf{X}\mathbf{Q}\mathbf{X}^{\mathrm{T}}\mathbf{D} = \mathbf{W}\mathbf{D}$ associés aux valeurs propres $\lambda_1 \geq \lambda_2 \geq \cdots \lambda_r > 0$

$\Longrightarrow$ *Vecteurs principaux de l'espace des variables.*

- Une matrice diagonale $\mathbf{\Theta}_{r\times r} = \mathbf{\Lambda}^{1/2}$ contenant les r **valeurs singulières** du triplet $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$ tq $\mathbf{X}$ se décompose en :

$$\mathbf{X}_{n\times p} = \mathbf{V}_{n\times r}\mathbf{\Theta}_{r\times r}\left(\mathbf{U}_{p\times r}\right)^{\mathrm{T}} = \sum_{i=1}^{r}\sqrt{\lambda_i}\,\boldsymbol{v}_i(\boldsymbol{u}_i)^{\mathrm{T}}$$

Ce qu'il faut retenir

- ***Les questions préalables à toute analyse :***
  - Les unités de mesure. Comment les données ont été acquises ?
  - La structure du tableau de données, le type des variables.
- ***Résumé d'une variable :***
  - graphiques,
  - paramètres numériques : position, dispersion, forme.
- ***Résumé d'un tableau de variables :***
  - *centre de gravité,*
  - *matrice de dispersion,*
  - *les deux espaces de représentation.*
- ***DVS d'une matrice de données.***

Exercices Td/TP ch I: calcul var, cor, DVS

Données supplémentaires

Données	Description
`Eaux1`	Corpus 20 eaux minérales décrites par 7 variables
`Eaux2`	Corpus données supplémentaires
`hemo_cir_quali.txt`	Corpus 136 sujets décrits par 24 variables
`Poumon.txt`	Corpus 72 patients décrits par 7 variables
`ChaZeb-a`	Corpus de 23 bovins (Charolais & Zébus) décrits par 6 variables pondérales.
`Kangourou`	Corpus de 151 kangourous des 2 sexes appartenant à 3 espèces décrits par 18 variables quantitatives.
`Loup`	Description de 43 crânes Chien/Loup par 6 variables. Identification d'un crane d'origine inconnue.

Travail à rendre par écrit en binôme …. `hemo_cir_quali.txt`

Données : hémachromatose

=> A rendre lors du dernier TP semaine 43 ou 44

Références

- L. Bellanger, R. Tomassone, *Exploration de données et méthodes statistiques : Data analysis & Data mining avec R. Collection Références Sciences, Editions* Ellipses, Paris, 2014.
- J.-M. Bouroche & G. Saporta, *L'analyse des données*. Presses Universitaires de France : Que sais-je ? 85, Paris, 1992.
- Site: www.math.univ-montp2.fr/~durand

Support intitulé « Elts de Calcul matriciel et d'Analyse Factorielle de Données »

- L. Lebart, A. Morineau, M. Piron, *Statistique exploratoire multidimensionnelle*. Dunod, Paris, 2006.
- J.-P. Nakache, J. Confais, *Approche pragmatique de la Classification.* Editions Technip, Paris, 2005.
- G. Saporta, *Probabilités, Analyse des données.* Editions Technip, Paris, 2006.
- R Foundation, http://www.r-project.org
- R. Tomassone, C. Dervin & J.-P. Masson. *Biométrie : modélisation de phénomènes biologiques*. Masson, Paris, 1993.