Algèbre linéaire, modèle gaussien et analyse de données. Notes de cours

Elisabeth Gassiat

Université Paris-Sud

Table des matières

1 Introduction.

On va étudier des tableaux de données :

$$(x_{i,j}) \quad i = 1, \ldots, n \quad j = 1, \ldots, p$$

$X_i = (x_{i,j})$ individu i, p caractères mesurés.

Représentation des tableaux: géométrie, donc algèbre linéaire.
Comprendre comment sont liés les caractères: lien avec les modèles probabilistes.

2 Espaces vectoriels $\mathbb{R}^n$.

2.1 Espaces et sous-espaces vectoriels.

Pour représenter en grande dimension, il faut des directions : les vecteurs; et des distances et angles, produit scalaire et norme.

Définition 1 *Un espace vectoriel réel E est un ensemble qui est muni de deux opérations: l'addition de deux vecteurs (éléments de l'espace vectoriel) et la multiplication par un nombre réel, pour lequel le résultat de ces opérations redonne un vecteur, et qui vérifient:*

- *E muni de son addition est un groupe abélien (addition associative, commutative, élément neutre 0, tout élément a un inverse)*
- *Multiplication par un réel qui est associative et distributive des deux côtés.*

$\mathbb{R}^n$ est un espace vectoriel. On note les vecteurs verticalement, et les opérations sont données par:
si $X = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$ et $Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$ alors $X + Y = \begin{pmatrix} x_1 + y_1 \\ \vdots \\ x_n + y_n \end{pmatrix}$
et si $\lambda \in \mathbb{R}$, alors $\lambda X = \begin{pmatrix} \lambda x_1 \\ \vdots \\ \lambda x_n \end{pmatrix}$.
Dans $\mathbb{R}^2$ représentation plane des vecteurs dans la base canonique.
Interprétation géométrique de l'addition et de la multiplication par un scalaire.

Définition 2 *F est un sous-espace vectoriel de $\mathbb{R}^n$ si il est non vide et si il est stable pour les deux opérations, c'est à dire*

$$\forall X \in F, \forall Y \in F, \forall \lambda \in \mathbb{R}, X + Y \in F \text{ et } \lambda X \in F.$$

Exemples et contre-exemples.

- La doite de pente -2, $F = \{\begin{pmatrix} x \\ y \end{pmatrix} \in \mathbb{R}^2 \ : \ y = -2x\}$, est un sous-espace vectoriel de $\mathbb{R}^2$.

– Dans $\mathbb{R}^3$, le plan $F = \{\begin{pmatrix} x \\ y \\ 0 \end{pmatrix}\}$, est un un sous-espace vectoriel de $\mathbb{R}^3$. On peut l'identifier à $\mathbb{R}^2$.
– Le carré $F = \{\begin{pmatrix} x \\ y \end{pmatrix} \in \mathbb{R}^2 \ : \ -1 \leq x \leq 1, -1 \leq y \leq 1\}$, n'est pas un sous-espace vectoriel de $\mathbb{R}^2$.

Propriétés.

1. Un sous espace vectoriel n'est jamais vide, il contient au moins le vecteur nul 0.
2. Le sous espace vectoriel de $\mathbb{R}^n$ réduit au vecteur nul est le plus petit sous espace vectoriel de $\mathbb{R}^n$. Le plus grand est $\mathbb{R}^n$ lui-même.
3. L'intersection de deux sous espaces vectoriels est un sous espace vectoriel. Attention: ce n'est pas le cas pour la réunion.
4. Si $m \leq n$, $\mathbb{R}^m$ est un sous espace vectoriel de $\mathbb{R}^n$ (avec la même identification que dans les exemples.).

Sous espaces vectoriels de $\mathbb{R}^2$.
En dehors de $\{0\}$ et de $\mathbb{R}^2$, les seuls sous-espaces vectoriels de R^2 sont les droites, c'est à dire les $F = \{\begin{pmatrix} x \\ y \end{pmatrix} \in \mathbb{R}^2 \ : \ ax + by = 0\}$ pour des réels fixés a et b.
En effet, si F est un sous-espace vectoriel de R^2 qui contient un vecteur non nul $U = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}$, alors F contient tous les vecteurs λU pour λ réel, et on peut montrer que F contient la droite $\{\begin{pmatrix} x \\ y \end{pmatrix} \in \mathbb{R}^2 \ : \ ax + by = 0\}$ avec $a = u_2$ et $b = -u_1$. Si F contient un autre vecteur non nul qui n'est pas élément de cette droite, on peut montrer que $F = \mathbb{R}^2$ (on le verra plus loin). Sinon, on dit que c'est la droite engendrée par U.

De manière générale, si un sous-espace vectoriel F contient un vecteur non nul U, il contient la droite engendrée par U.

Sous espaces vectoriels de $\mathbb{R}^3$.
En dehors de $\{0\}$ et de $\mathbb{R}^3$, les seuls sous-espaces vectoriels de R^2 sont les droites et les plans.

La droite engendrée par $U = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix}$ est l'ensemble des vecteurs

$X = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$ qui vérifient $\begin{cases} x_1 u_2 - x_2 u_1 = 0 \\ x_1 u_3 - x_3 u_1 = 0 \end{cases}$.

Si F est un sous-espace vectoriel qui contient deux vecteurs $U = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix}$ et $V = \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix}$ qui ne sont pas sur la même droite, on peut montrer que c'est un plan de la forme $F = \{ \begin{pmatrix} x \\ y \\ z \end{pmatrix} \in \mathbb{R}^3 \ : \ ax + by + cz = 0\}$ pour des réels fixés a et b (dépendant de U et V).
On voit que l'intersection de deux plans non confondus est une droite.

2.2 Combinaisons linéaires. Familles génératrices.

Définition 3 *Soient $X_1, \ldots, X_k$ des vecteurs de $\mathbb{R}^n$. On appelle combinaison linéaire de $X_1, \ldots, X_k$ tout vecteur de la forme $\alpha_1 X_1 + \ldots + \alpha_k X_k$ où $\alpha_1, \ldots, \alpha_k$ sont des réels.*

Si F est un sous espace vectoriel, toute combinaison linéaire de vecteurs de F est un vecteur de F.

Définition 4 *Si $X_1, \ldots, X_k$ sont des vecteurs de $\mathbb{R}^n$, on appelle espace engendré par $X_1, \ldots, X_k$ l'ensemble de toutes les combinaisons linéaires de $X_1, \ldots, X_k$. C'est un sous espace vectoriel de $\mathbb{R}^n$ (exercice: le démontrer).*

Exemple : droites de R^n; plans de $\mathbb{R}^n$.

Définition 5 *Soient $X_1, \ldots, X_k$ des vecteurs de $\mathbb{R}^n$ et F un sous espace vectoriel de $\mathbb{R}^n$. On dit que $X_1, \ldots, X_k$ est une famille génératrice de F (ou encore que $X_1, \ldots, X_k$ engendrent F) si tout vecteur de F est combinaison linéaire de $X_1, \ldots, X_k$.*

Exemples.
- $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$ et $\begin{pmatrix} 0 \\ 1 \end{pmatrix}$ engendrent $\mathbb{R}^2$.
- $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$ et $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$ engendrent $\mathbb{R}^2$.
- $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$ et $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$ et $\begin{pmatrix} 0 \\ 1 \end{pmatrix}$ engendrent $\mathbb{R}^2$.
- $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$ n'engendre pas $\mathbb{R}^2$.
- $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$ et $\begin{pmatrix} 3 \\ 3 \end{pmatrix}$ engendrent la droite de $\mathbb{R}^2$ déquation $x = y$.
- De manière générale les vecteurs $E_1, \ldots E_n$ définis par E_i a toutes ses coordonnées nulles sauf la i-ème engendre $\mathbb{R}^n$.

Sur les exemples on voit qu'on peut trouver différentes familles gńératrices.On cherche à trouver la famille la plus petite.

2.3 Indépendance linéaire.

Définition 6 *On dit que $X_1, \ldots, X_k$ forment une famille liée de vecteurs si au moins l'un des X_i est combinaison linéaire des autres.*
On dit que $X_1, \ldots, X_k$ forment une famille libre de vecteurs si ils ne forment pas une famille liée.

Proposition 1 *$X_1, \ldots, X_k$ forment une famille libre de vecteurs si et seulement si une combinaison linéaire de $X_1, \ldots, X_k$ ne peut être nulle que lorsque les coefficients sont tous nuls, c'est à dire:*

$$\alpha_1 X_1 + \ldots + \alpha_k X_k = 0 \Longrightarrow \alpha_1 = \ldots = \alpha_k = 0.$$

On dit aussi que les vecteurs $X_1, \ldots, X_k$ sont linéairement indépendants.

Preuve.
Soit $X_1, \ldots, X_k$ une famille de vecteurs. Si cette famille est liée, alors quitte à renuméroter les vecteurs on peut dire que X_1 est combinaison linéaire des autres, c'est à dire qu'il existe des réels $\beta_2, \ldots, \beta_k$ tels que $X_1 = \beta_2 X_2 + \ldots + \beta_k X_k$, et l'on a $\alpha_1 X_1 + \ldots + \alpha_k X_k = 0$ avec $\alpha_1 = 1$ et $\alpha_i = -\beta_i$ pour $i = 2, \ldots, k$, donc les vecteurs $X_1, \ldots, X_k$ ne sont pas linéairement indépendants.
Réciproquement si il existe des réels $\alpha_1, \ldots, \alpha_k$ qui ne sont pas tous nuls tels que, $\alpha_1 X_1 + \ldots + \alpha_k X_k = 0$, quitte à les renuméroter on peut supposer $\alpha_1 \neq 0$, et alors $X_1 = -\frac{\alpha_2}{\alpha_1} X_2 - \ldots - \frac{\alpha_k}{\alpha_1} X_k$, et la famille $X_1, \ldots, X_k$ est liée.

Exemples.

- $U = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$, $V = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$ et $W = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$ sont liés dans $\mathbb{R}^2$. En effet, $W = U - 2V$.
- $U = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ et $V = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$ sont linéairement indépendants dans $\mathbb{R}^2$, car si α et β sont des réels tels que $\alpha U + \beta V = 0$, alors $\begin{pmatrix} \alpha \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ \beta \end{pmatrix} = 0$, c'est à dire $\begin{pmatrix} \alpha \\ \beta \end{pmatrix} = 0$, et donc $\alpha = 0$ et $\beta = 0$.

2.4 Bases et dimension d'un espace vectoriel.

Définition 7 *Une famille de vecteurs de F est une base de F si elle est à la fois libre et génératrice de F.*

Exemples:

- Les vecteurs $E_1, \ldots E_n$ définis par E_i a toutes ses coordonnées nulles sauf la i-ème est une base de $\mathbb{R}^n$.
- Si X est un vecteur non nul de $\mathbb{R}^n$ et si F est la droite engendrée par X, alors X est une base de F.
- Si X et Y sont deux vecteurs non nuls de $\mathbb{R}^n$ qui ne sont pas alignés (c'est à dire que Y n'est pas élément de la droite engendrée par X), et si F est le plan engendrée par X et Y, alors X, Y est une base de F.

Remarque: il y a une infinité de bases.

Proposition 2 *Si $B_1, \ldots, B_m$ est une base de F, pour tout vecteur X de F il existe des réels uniques $\alpha_1, \ldots, \alpha_m$ tels que $X = \alpha_1 B_1 + \ldots + \alpha_m B_m$.*
On dit que X admet une décomposition unique sur B. Les réels $\alpha_1, \ldots, \alpha_m$ s'appellent les composantes (ou les coordonnées) de X dans B.

Preuve.
Puisque $B_1, \ldots, B_m$ est une famille génératrice de de F, si X est un vecteur de F, X est une combinaison linéaire de $B_1, \ldots, B_m$, donc il existe des réels $\alpha_1, \ldots, \alpha_m$ tels que $X = \alpha_1 B_1 + \ldots + \alpha_m B_m$.
Si des réels $\beta_1, \ldots, \beta_m$ sont tels que $X = \beta_1 B_1 + \ldots + \beta_m B_m$, alors on a

$$(\alpha_1 - \beta_1)B_1 + \ldots + (\alpha_m - \beta_m)B_m = 0,$$

et puisque la famille $B_1, \ldots, B_m$ est une famille libre, ceci implique que $\alpha_1 - \beta_1 = \ldots = \alpha_m - \beta_m = 0$, autrement dit que les β_i sont les mêmes que les α_i.

Théorème 1 *Si F est un sous espace vectoriel de $\mathbb{R}^n$, si $X_1, \ldots, X_p$ forme une famille génératrice de F et si $Y_1, \ldots, Y_m$ forme une famille libre de F, alors $p \geq m$.*

Preuve.
Puisque $X_1, \ldots, X_p$ forme une famille génératrice de F, il existe des réels $a_1, \ldots, a_p$ tels que $Y_1 = a_1 X_1 + \ldots + a_p X_p$. Au moins l'un des a_i est non nul, quitte à renuméroter on peut supposer que $a_1 \neq 0$. On a alors $X_1 = \frac{1}{a_1} Y_1 - \frac{a_2}{a_1} X_2 - \ldots - \frac{a_p}{a_1} X_p$, et l'on voit facilement que $Y_1, X_2, \ldots, X_p$ forme une famille génératrice de F.
Il existe donc des réels $b_1, \ldots, b_p$ tels que $Y_2 = b_1 Y_1 + b_2 X_2 + \ldots + b_p X_p$. Au moins l'un des b_i est non nul, il n'est pas possible que seul b_1 soit non nul (sinon Y_1, Y_2 formerait une famille liée), donc quitte à renuméroter on peut supposer que $b_2 \neq 0$. On a alors $X_2 = \frac{1}{b_2} Y_2 - \frac{b_1}{b_2} Y_1 - \frac{b_3}{b_2} X_3 - \ldots - \frac{b_p}{b_2} X_p$, et l'on voit facilement que $Y_1, Y_2, X_3, \ldots, X_p$ forme une famille génératrice de F.
Si $p < m$, en continuant ainsi, on arrive au fait que $Y_1, Y_2, \ldots, Y_{p-1}, X_p$ forme une famille génératrice de F, puis que $Y_1, Y_2, \ldots, Y_p$ forme une famille génératrice de F. $p+1 \leq m$, et l'on peut obtient alors que Y_{p+1} est combinaison linéaire de $Y_1, Y_2, \ldots, Y_p$, ce qui contredit le fait que $Y_1, \ldots, Y_m$ forme une

famille libre de F, et donc $p \geq m$.

Conséquence: Si F est un sous espace vectoriel de $\mathbb{R}^n$, toutes les bases de F ont le même nombre de vecteurs.
Pour le voir, prendre deux bases $B_1, \ldots, B_p$ et $C_1, \ldots, C_m$. En appliquant le théorème avec les B_i famille génératrice et les C_j famille libre, on obtient $p \geq m$, puis en prenant les B_i famille libre et les C_j famille génératrice on obtient $m \geq p$ et donc $p = m$.

Définition 8 *On appelle dimension d'un sous espace vectoriel F le nombre de vecteurs d'une base de F.*

Proposition 3 *Si F est un sous espace vectoriel de E, alors $dim(F) \leq dim(E)$ et $F = E$ si et seulement si $dim(F) = dim(E)$*

Preuve.
Si $dim(F) = m$ et si $B_1, \ldots, B_m$ est une base de F, alors puisque $F \subset E$, $B_1, \ldots, B_m$ forme une famille libre de E.
Si ensuite $dim(E) = m$, $B_1, \ldots, B_m$ est une base de E.

3 Outils de géométrie euclidienne.

3.1 Produit scalaire.

Définition 9 *Si* $X = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$ *et* $Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$ *alors le produit scalaire de X et Y est*

$$\langle X, Y \rangle = x_1 y_1 + \ldots + x_n y_n$$

Notations et calcul avec $\langle X, Y \rangle = X^T Y$, où $X^T = (x_1, \ldots, x_n)$.

Propriétés:

1. Symétrie : $\langle X, Y \rangle = \langle Y, X \rangle$.
2. Si α est un réel, $\langle \alpha X, Y \rangle = \langle X, \alpha Y \rangle = \alpha \langle X, Y \rangle$.
3. $\langle X + Y, Z \rangle = \langle X, Z \rangle + \langle Y, Z \rangle$
4. Bilinárité (se développe comme un produit), c'est à dire, si a, b, c, d sont des réels et X, Y, Z, U sont des vecteurs, alors $\langle aX + bY, cZ + dU \rangle = ac\langle X, Z \rangle + bc\langle Y, Z \rangle + ad\langle X, U \rangle + bd\langle Y, U \rangle$.
5. Si X ou Y est nul, $\langle X, Y \rangle = 0$, mais la réciproque est fausse.

3.2 Norme.

Introduction dans $\mathbb{R}^2$: calcul de la norme avec Pythagore.

Définition 10 *Si* $X \in \mathbb{R}^n$*, la norme de* X *est le réel*

$$\| X \| = \sqrt{\langle X, X \rangle}.$$

Soit si $X = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$, $\| X \| = \sqrt{x_1^2 + \ldots + x_n^2}$.

Propriétés:

1. La norme est un réel positif.
2. $\| X \| = 0$ si et seulement si X est le vecteur nul.
3. Si α est un réel, $\| \alpha X \| = |\alpha| \, \| X \|$.
4. $\| X + Y \|^2 = \| X \|^2 + \| Y \|^2 + 2\langle X, Y \rangle$. En effet, $\| X + Y \|^2 = \langle X + Y, X + Y \rangle$ et on développe.
5. Inégalité de Cauchy-Schwartz:

   $$|\langle X, Y \rangle| \leq \| X \| \| Y \|$$

   Preuve: Pour tout réel α, $\| X + \alpha Y \|^2$ est un réel positif, autrement dit en développant, $\| X \|^2 + \alpha^2 \| Y \|^2 + 2\alpha \langle X, Y \rangle$ est un polynôme de degré 2 en α qui est toujours positif, donc son discriminant est négatif, et son discriminant réduit est $\langle X, Y \rangle^2 - \| X \|^2 \| Y \|^2$.
6. Inégalité triangulaire:

   $$\| X + Y \| \leq \| X \| + \| Y \|$$

   Preuve: s'obtient à partir du développement de $\| X + Y \|^2$ en appliquant l'inégalité de Cauchy-Schwartz.

3.3 Orthogonalité.

Introduction angle (non orienté) entre deux vecteurs dans $\mathbb{R}^2$: calcul du cosinus de l'angle (non orienté) entre deux vecteurs.

Proposition 4 *Si* X *et* Y *sont deux vecteurs non nuls de* $\mathbb{R}^n$*, et si* θ *est l'angle non orienté entre les deux vecteurs, alors*

$$\cos\theta = \frac{\langle X, Y \rangle}{\| X \| \| Y \|}$$

Remarques: on retrouve Cauchy-Schwartz.
Orthogonaux ssi $\theta = \pi/2$ ou $\theta = 3\pi/2$.

Définition 11 *Deux vecteurs X et Y sont dits orthogonaux si et seulement si $\langle X, Y\rangle = 0$.*

Exemples.

- Dans $\mathbb{R}^2$, si $V = \begin{pmatrix} u_2 \\ -u_1 \end{pmatrix}$, l'ensemble des vecteurs orthogonaux à V est l'ensemble des vecteurs $\begin{pmatrix} x \\ y \end{pmatrix}$ tels que $u_2 x - u_1 y = 0$. C'est la droite engendrée par $U = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}$.
- Dans $\mathbb{R}^3$, si $W = \begin{pmatrix} a \\ b \\ c \end{pmatrix}$, l'ensemble des vecteurs orthogonaux à W est l'ensemble des vecteurs $\begin{pmatrix} x \\ y \\ z \end{pmatrix}$ tels que $ax+by+cz = 0$. On peut montrer (exercice) que c'est un sous espace vectoriel de dimension 2. $U = \begin{pmatrix} b \\ -a \\ 0 \end{pmatrix}$ et $V = \begin{pmatrix} c \\ 0 \\ -a \end{pmatrix}$ en sont deux vecteurs linéairement indépendants, et ce sous espace vectoriel est donc le plan engendré par U et V.

3.4 Bases orthonormées.

Définition 12 *On dit que $B_1, \ldots, B_m$ est une base orthonormée de F si et seulement si les B_i sont tous de norme 1 et sont orthogonaux deux à deux.*

Exemple: la base canonique de $\mathbb{R}^n$ est une base orthonormée de $\mathbb{R}^n$.

Proposition 5 *Un changement de base orthonormée conserve norme et produit scalaire: c'est à dire que si $B = (B_1, \ldots, B_m)$ est une base orthonormée de F, si X est un vecteur de F de coordonnées $\alpha_1, \ldots, \alpha_m$ dans cette base, et si Y est un vecteur de F de coordonnées $\beta_1, \ldots, \beta_m$ dans cette base, on peut calculer le produit scalaire avec les coordonnées dans la nouvelle base orthonormée:*

$$\langle X, Y\rangle = \alpha_1\beta_1 + \ldots + \alpha_m\beta_m.$$

Preuve :

$$\begin{aligned}\langle X, Y\rangle &= \langle\sum_{i=1}^{m}\alpha_i B_i, \sum_{j=1}^{m}\beta_j B_j\rangle \\ &= \sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\beta_j\langle B_i, B_j\rangle \\ &= \sum_{i=1}^{m}\alpha_i\beta_i\end{aligned}$$

car si $i \neq j$, $\langle B_i, B_j\rangle \neq 0$, et si $i = j$, $\langle B_i, B_j\rangle = 1$.

On peut remarquer que $\langle X, B_i\rangle = \alpha_i$, autrement dit que la décomposition d'un vecteur dans une base orthonormée s'écrit simplement

$$X = \sum_{i=1}^{m}\langle X, B_i\rangle B_i.$$

Proposition 6 *Si $B_1, \ldots, B_m$ est une base de F, on peut construire une base orthonormée de F $U_1, \ldots, U_m$ telle que pour tout $j \leq m$, l'espace engendré par $B_1, \ldots, B_j$ est égal à l'espace engendré par $U_1, \ldots, U_j$.*

Preuve:
On construit $U_1, \ldots, U_m$ progressivement par "orthonormalisation". On pose tout d'abord $U_1 = B_1/ \parallel B_1 \parallel$. L'espace engendré par B_1 est bien égal à l'espace engendré par U_1. On choisit ensuite $U_2 = a_1B_1 + a_2B_2$ de telle sorte que $\parallel U_2 \parallel = 1$ et $\langle U_1, U_2\rangle = 0$: on trouve (calcul) que

$$a_1 = -\frac{\langle B_1, B_2\rangle}{\parallel B_1 \parallel \sqrt{\parallel B_1 \parallel^2 \parallel B_2 \parallel^2 - (\langle B_1, B_2\rangle)^2}} \quad a_2 = \frac{\parallel B_1 \parallel}{\sqrt{\parallel B_1 \parallel^2 \parallel B_2 \parallel^2 - (\langle B_1, B_2\rangle)^2}}$$

L'espace engendré par B_1 et B_2 est alors bien égal à l'espace engendré par U_1 et U_2. On continue en choisissant U_3 combinaison linéaire de B_1, B_2 et B_3, etc...

3.5 Projection orthogonale

3.5.1 Projection orthogonale sur une droite.

Introduction dans $\mathbb{R}^2$, avec dessin et calcul du projeté (colinéaire avec le vecteur directeur de la droite + orthogonalité).

Proposition 7 *Dans $\mathbb{R}^n$, la projection orthogonale d'un vecteur Y sur la droite engendrée par un vecteur X est le vecteur*

$$\frac{\langle X, Y\rangle}{\parallel X \parallel^2} \cdot X.$$

Preuve:
Si l'on appelle $\Pi(Y)$ la projection orthogonale du vecteur Y sur la droite engendrée par le vecteur X, alors $\Pi(Y)$ est un vecteur colinéaire à X tel que $Y - \Pi(Y)$ est orthogonal à X, soit $\Pi(Y) = \alpha X$ pour un réel α, et $\langle X, Y - \Pi(Y)\rangle = 0$ donne $\langle X, Y\rangle = \alpha\langle X, X\rangle$.

Faire une projection orthogonale c'est faire des "moindres rectangles":

Proposition 8 *Dans $\mathbb{R}^n$, la projection orthogonale d'un vecteur Y sur la droite $\mathcal{D}$ engendrée par un vecteur X est le vecteur de $\mathcal{D}$ le plus proche de Y, c'est à dire le vecteur $\Pi(Y)$ tel que*

$$\Pi(Y) \in \mathcal{D} \textit{ et } \| Y - \Pi(Y) \| = \min\{\| Y - V \|, V \in \mathcal{D}\}.$$

Preuve:
Si V est un vecteur quelconque de $\mathcal{D}$, alors

$$\| Y - V \|^2 = \| Y - \Pi(Y) \|^2 + \| \Pi(Y) - V \|^2$$

car $Y - \Pi(Y)$ est orthogonal à V.

3.5.2 Projection orthogonale sur un sous espace-vectoriel.

Soit F un sous espace vectoriel de $\mathbb{R}^n$ de dimension m. On peut choisir une base orthonormée $B_1, \ldots B_n$ de $\mathbb{R}^n$ telle que $B_1, \ldots, B_m$ soit une base orthonormée de F.
Alors, pour tout vecteur Y de $\mathbb{R}^n$, la projection orthogonale $\Pi(Y)$ de Y sur F est

$$\Pi(Y) = \sum_{i=1}^{m} \langle Y, B_i\rangle B_i.$$

En effet, ce vecteur $\Pi(Y)$ est un vecteur de F tel que $Y - \Pi(Y)$ est orthogonal à tous les $B_1, \ldots, B_m$, donc à tous les vecteurs de F.
V De plus,

Proposition 9 *La projection orthogonale de Y sur F est le vecteur de F le plus proche de Y, c'est à dire le vecteur $\Pi(Y)$ tel que*

$$\Pi(Y) \in F \textit{ et } \| Y - \Pi(Y) \| = \min\{\| Y - V \|, V \in F\}.$$

La preuve est exactement la même que lorsque F est une droite en prenant V quelconque dans F.

4 Vecteurs gaussiens 1.

Introduction: pourquoi des vecteurs gaussiens multivariés et pourquoi des variables gaussiennes non indépendantes : parce que lorsqu'on étudiera des nuages de points, d'individus avec p caractères, on se demandera si les caractères sont indépendants ou non, et qu'une façon de modéliser est le modèle gaussien multidimensionnel.

4.1 Rappels

4.1.1 Lois gaussiennes sur $\mathbb{R}$.

La loi Gaussienne $\mathcal{N}(0;1)$ est la loi sur $\mathbb{R}$ de densité

$$\frac{1}{\sqrt{2\pi}}\exp(-\frac{x^2}{2}).$$

Elle est symétrique : si $X \sim \mathcal{N}(0;1)$, $-X \sim \mathcal{N}(0;1)$.
La loi Gaussienne $\mathcal{N}(m,\sigma^2)$ est la loi sur $\mathbb{R}$ de densité

$$\frac{1}{\sigma\sqrt{2\pi}}\exp(-\frac{(x-m)^2}{2\sigma^2}).$$

Lorsque U suit la loi $\mathcal{N}(0,1)$, alors σU suit la loi $\mathcal{N}(0,\sigma^2)$. De même $m+\sigma U$ suit la loi $\mathcal{N}(m,\sigma^2)$.
Réciproquement, si X suit la loi $\mathcal{N}(m,\sigma^2)$, $\frac{X-m}{\sigma}$ suit la loi $\mathcal{N}(0,1)$. (on appelle cela centrer et réduire).

4.1.2 Loi du χ^2.

Suposons que $U_1,\ldots,U_d$ sont des variables indépendantes ayant pour loi la loi normale centrée réduite $\mathcal{N}(0,1)$. On pose alors

$$Z = U_1^2 + \cdots + U_d^2.$$

La loi de Z s'appelle la loi du χ^2 à d degrés de liberté. On la note χ^2_d. Notons que Z est toujours positive (c'est une somme de carrés): c'est la norme au carré du vecteur $U = \begin{pmatrix} U_1 \\ \vdots \\ U_d \end{pmatrix}$ de $\mathbb{R}^d$.
On peut montrer que sa densité est (pour tout $x \geq 0$),

$$c_d x^{(d-2)/2} e^{-x/2}$$

(c_d est une constante choisie pour que $\int_0^\infty c_d x^{(d-2)/2}e^{-x/2}dx = 1$). Cette densité décroit donc rapidement lorsque x croît vers l'infini. En pratique, pour calculer des probabilités, si $d \leq 30$, on utilise les tables de valeurs numériques. Si $d \geq 30$, on peut utiliser l'approximation par une loi normale. En effet, on peut appliquer le Théorème central limite: Z/d est une moyenne empirique, et il suffit de calculer l'espérance et la variance des Y_i pour appliquer le théorème. On a:

$$E(U_1^2) = 1, \quad \mathrm{Var}(U_1^2) = 2,$$

donc le théorème central limite donne, pour d grand

$$\frac{Z-d}{\sqrt{2d}} \sim \mathcal{N}(0,1).$$

4.1.3 Variables indépendantes. Covariance entre deux variables réelles.

On dit que les variables à densité $X_1, \dots, X_n$ sont indépendantes si pour tous $x_1, \dots, x_n$,

$$\begin{aligned} & P(X_1 \in [x_1, x_1 + dx_1] \text{ et } \dots \text{ et } X_n \in [x_n, x_n + dx_n]) \\ = \quad & P(X_1 \in [x_1, x_1 + dx_1]) \times \dots \times P(X_n \in [x_n, x_n + dx_n]). \end{aligned}$$

On a vu aussi que deux variables X et Y sont indépendantes si et seulement si leur densité jointe est égale au produit des densités marginales de X et de Y. Lorsque l'on a plus de deux variables, on peut reprendre toutes les notions introduites pour un couple de variables et les étendre à un vecteur de variables aléatoires:

Un vecteur aléatoire d-dimensionnel est une variable aléatoire $X = \begin{pmatrix} X_1 \\ \vdots \\ X_d \end{pmatrix}$ à valeur dans $\mathbb{R}^d$.

Si l'ensemble des valeurs que peut prendre X est "continu", on peut caractériser sa loi par sa densité de probabilité $f(x_1, \dots, x_d)$, que l'on appelle densité jointe. On a:

$$f(x_1, \dots, x_d) \geq 0 \text{ et } \int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} f(x_1, \dots, x_d) dx_1 \dots dx_d = 1.$$

La loi permet de calculer les probabilités d'événements relatifs à X: si A est une partie de $\mathbb{R}^d$,

$$P(X \in A) = \int \dots \int_{(x_1, \dots, x_d) \in A} f(x_1, \dots, x_d) dx_1 \dots dx_d.$$

Si $X = \begin{pmatrix} X_1 \\ \vdots \\ X_d \end{pmatrix}$ est un vecteur aléatoire d-dimensionnel, chacun des X_i est une variable aléatoire. On appelle loi marginale de X_i la loi de la variable aléatoire

X_i.
On a alors: les variables $X_1, \ldots, X_d$ sont indépendantes si et seulement si leur densité jointe est égale au produit de leurs densités marginales.

La covariance entre deux variables réelles X et Y est le nombre

$$\mathrm{Cov}(X,Y) = E\left[(X - E(X))\,(Y - E(Y))\right] = E(XY) - E(X)E(Y).$$

On peut remarquer que si l'on prend $Y = X$, on obtient

$$\mathrm{Cov}(X,X) = \mathrm{Var}(X),$$

et que si l'on prend $Y = -X$, on obtient

$$\mathrm{Cov}(X,-X) = -\mathrm{Var}(X).$$

La corrélation entre X et Y est le nombre

$$Corr(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}.$$

Rappelons que l'on peut interpréter le coefficient de corrélation comme un indice de liaison affine entre deux variables car on a:

Proposition 10 *On a toujours*

$$-1 \leq Corr(X,Y) \leq 1.$$

$Corr(X,Y) = 1$ si et seulement si Y est une fonction affine de X de coefficient directeur positif, et $Corr(X,Y) = -1$ si et seulement si Y est une fonction affine de X de coefficient directeur négatif.

4.2 Vecteurs gaussiens.

Définition 13 *On dit que $X = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}$ est un vecteur gaussien n-dimensionnel si c'est un vecteur aléatoire tel que toute combinaison linéaire réelle des X_i est une variable gaussienne.*

Remarques:

1. Chacun des X_i est une variable gaussienne de loi $\mathcal{N}(m_i, \sigma_i^2)$.
2. Si $Y = \alpha_1 X_1 + \ldots + \alpha_n X_n$, on a:

$$E(Y) = \alpha_1 E(X_1) + \ldots + \alpha_n E(X_n) = \alpha_1 m_1 + \ldots + \alpha_n m_n,$$

$$Var(Y) = \sum_{i,j=1}^{n} \alpha_i \alpha_j Cov(X_i, X_j).$$

En effet,

$$\begin{aligned}
Var(Y) &= E\left[(\alpha_1 X_1 + \ldots + \alpha_n X_n) - E\left(\alpha_1 X_1 + \ldots + \alpha_n X_n\right)\right]^2 \\
&= E\left[\alpha_1 (X_1 - m_1) + \ldots + \alpha_n (X_n - m_n)\right]^2 \\
&= \sum_{i,j=1}^{n} \alpha_i \alpha_j E\left[(X_i - m_i)(X_j - m_j)^2\right].
\end{aligned}$$

Définition 14 *Si* $X = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}$ *est un vecteur gaussien n-dimensionnel, on appelle espérance de* X *le vecteur* $m = \begin{pmatrix} m_1 \\ \vdots \\ m_n \end{pmatrix}$ *où* $m_i = E(X_i)$.
On appelle matrice de variance de X *la matrice* Σ *telle que*

$$\Sigma_{i,j} = Cov(X_i, X_j).$$

On note alors $\mathcal{N}(m, \Sigma)$ *la loi du vecteur* X.

On peut noter que la corrélation de X_i et X_j est alors

$$Corr(X_i, X_j) = \frac{\Sigma_{i,j}}{\sqrt{\Sigma_{i,i}\Sigma_{j,j}}}.$$

Apprentissage du calcul matriciel, avec si $A = \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{pmatrix}$, si $Y = \langle A, X \rangle = A^T X$, alors $E(Y) = A^T m$ et $\mathrm{Var}(Y) = A^T \Sigma A$.

Remarques:

1. Σ est symétrique, c'est à dire pour tous i et j, $\Sigma_{i,j} = \Sigma_{j,i}$.
2. Σ est semi-définie positive: si A est un vecteur de $\mathbb{R}^n$, $A^T \Sigma A \geq 0$ puisque c'est la variance de $A^T X$.
3. Il se peut qu'il y ait des combinaisons linéaires des X_i de variance nulle, c'est à dire qu'il existe des coefficients $\alpha_1, \ldots, \alpha_n$ tel que $Var(\alpha_1 X_1 + \ldots + \alpha_n X_n) = 0$.
   Cela signifie que les variables $X_1 - m_1$, ..., $X_n - m_n$ sont presque surement dans l'hyperplan d'équation $\alpha_1 x_1 + \ldots + \alpha_n x_n = 0$, autrement dit que le vecteur $X - m$ appartient à cet hyperplan avec probabilité 1.
   Dans ce cas $X - m$ est avec probabilité dans un espace dimension $n - 1$, et la loi de X n'a pas de densité sur $\mathbb{R}^n$.

Proposition 11 *(Admis. Attention) Dans un vecteur gaussien, deux variables sont indépendantes si et seulement si leur covariance est nulle.*

4.3 Echantillon gaussien

Définition 15 *Un n-échantillon gaussien standard est un vecteur gaussien n-dimensionnel dont les coordonnées sont deux à deux indépendantes et de loi $\mathcal{N}(0,1)$.*

Remarques:

1. Si $X = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}$ est un n-échantillon gaussien standard, alors sa densité est
$$f(x_1, \ldots, x_n) = \frac{1}{(\sqrt{2\pi})^n} \exp(-\frac{x_1^2 + \ldots + x_n^2}{2}).$$

2. Soit $B_1, \ldots, B_n$ une base orthonormée de $\mathbb{R}^n$. Soit $X = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}$ un n-échantillon gaussien standard. Soit $Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$ les coordonnées de X dans la base orthonormée. Alors Y est un n-échantillon gaussien standard. En effet, on sait que $Y_i = \langle X, B_i \rangle$. Toute combinaison linéaire des Y_i est une combinaison linéaire des X_i, donc est une variable gaussienne. Donc Y est un vecteur gaussien n-dimensionnel. Pour montrer que c'est un n-échantillon gaussien, il suffit de montrer que les Y_i sont centrés, réduits, et deux à deux indépendants, donc d'après la proposition précédente, que les covariances de deux Y_i différents sont nulles. Dans la base canonique $E_1, \ldots, E_n$, soit A_i le vecteur des coordonnées de B_i, on a:
$$E(Y_i) = A_i^T m = 0$$
où m est l'espérance de X, donc le vecteur nul.
Si l'on note $A_i = \begin{pmatrix} A_{1,i} \\ \vdots \\ A_{n,i} \end{pmatrix}$, on a
$$Y_i = \sum_{k=1}^{n} A_{k,i} X_k,$$
donc
$$Y_i Y_j = \sum_{k=1}^{n} \sum_{l=1}^{n} A_{k,i} A_{l,j} X_k X_l,$$
et comme Y_i et Y_j sont centrés
$$Cov(Y_i, Y_j) = E(Y_i Y_j) = \sum_{k=1}^{n} \sum_{l=1}^{n} A_{k,i} A_{l,j} E(X_k X_l).$$

Puisque si $k \neq l$, $E(X_k X_l) = 0$ et si $k = l$, $E(X_k X_l) = 1$ on a

$$Cov(Y_i, Y_j) = \sum_{k=1}^{n} A_{k,i} A_{k,j} = \langle B_i, B_j \rangle$$

qui vaut 0 si $i \neq j$ et 1 si $i = j$.

Exemple dans $\mathbb{R}^2$.
Soit $X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ un vecteur gaussien de $\mathbb{R}^2$ de loi $\mathcal{N}(m, \Sigma)$ avec

$$m = \begin{pmatrix} m_1 \\ m_2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \sigma_1^2 & \rho \\ \rho & \sigma_1^2 \end{pmatrix}.$$

X est un 2-échantillon gaussien standard si $m_1 = m_2 = 0$, $\sigma_1 = \sigma_2 = 1$ et $\rho = 0$.
Si $Y = aX_1 + bX_2$ et $Z = cX_1 + dX_2$, alors $\begin{pmatrix} Y \\ Z \end{pmatrix}$ est un vecteur gaussien de $\mathbb{R}^2$ de loi $\mathcal{N}(\mu, \Lambda)$ avec

$$\begin{aligned}
\mu_1 &= am_1 + bm_2 \\
\mu_2 &= cm_1 + dm_2 \\
\Lambda_{1,1} &= a^2\sigma_1^2 + 2ab\rho + b^2\sigma_2^2 \\
\Lambda_{2,2} &= c^2\sigma_1^2 + 2cd\rho + d^2\sigma_2^2 \\
\Lambda_{1,2} &= ac\sigma_1^2 + (ad + bc)\rho + bd\sigma_2^2.
\end{aligned}$$

En particulier, si $\begin{pmatrix} a \\ b \end{pmatrix}$ et $\begin{pmatrix} c \\ d \end{pmatrix}$ forment une base orthonormée de $\mathbb{R}^2$, alors $\Lambda = \Sigma$: la transformation d'un vecteur gaussien par changement de base orthonormée ne change pas la "géométrie" de sa loi.

Soit maintenant $U = \begin{pmatrix} U_1 \\ U_2 \end{pmatrix}$ un vecteur de norme 1 et $\mathcal{D}$ la droite engendrée par U. Soit $P(X)$ la projection orthogonale de X sur $\mathcal{D}$. Puisque

$$P(X) = \langle X, U \rangle U,$$

$P(X)$ est une variable gaussienne de $\mathbb{R}$ de loi $\mathcal{N}(\langle m, U \rangle U, \sigma^2)$ avec $\sigma^2 = U_1^2\sigma_1^2 + 2U_1U_2\rho + U_2^2\sigma_2^2$, et si $m = 0$, $\rho = 0$ et $\sigma_1 = \sigma_2$, alors $P(X)$ suit la loi $\mathcal{N}0, \sigma_1^2)$.

5 Applications linéaires et matrices associées.

Introduction: Pour l'instant on a vu une matrice comme un tableau de chiffres avec des lignes et des colonnes. Dans ce cas on peut voir la matrice comme représentant des vecteurs, c'est à dire chaque colonne représente un vecteur, donc ses coordonnées dans la base canonique. On peut alors aussi, si on

change de base et que l'on met dans les colonnes les coordonnées des vecteurs dans la nouvelle base, dire que la matrice représente les vecteurs dans cette nouvelle base.
Une matrice peut donc être vue comme représentant des vecteurs, et il faut alors dire dans quelle base.
Une autre façon de voir les matrices est de les associer à des applications linéaires, c'est à dire à des transformations de l'espace vectoriel composant des translations et des dilatations. Cela permettra de comprendre les opérations que l'on peut faire subir à un tableau de données, et aussi de voir comment cela opère sur les modèles probabilistes.

5.1 Application linéaire.

Définition 16 *Soient E et F des espaces vectoriels. L'application f de E dans F est une application linéaire si et seulement si:*

$$\forall X, Y \in E, \;\; f(X+Y) = f(X) + f(Y)$$

$$\forall X \in E, \forall \alpha \in \mathbb{R} \;\; f(\alpha X) = \alpha f(X)$$

Exemples et contre-exemples:

- Si U est un vecteur fixé de E, l'application qui à tout vecteur X de E associe $f(X) = \langle U, X \rangle$ est une application linéaire de E dans $\mathbb{R}$.
- L'application de $\mathbb{R}^2$ dans $\mathbb{R}$ qui à tout vecteur X associe sa norme n'est pas une application linéaire.

Proposition 12 *Si f est une application linéaire de E dans F, alors l'image par f d'une combinaison linéaire d'éléments de E est la combinaison linéaire des images par f de ces éléments:*

$$\forall X_1, \ldots, X_p \in E, \forall \alpha_1, \ldots, \alpha_p \in \mathbb{R} \;\; f(\alpha_1 X_1 + \ldots + \alpha_p X_p) = \alpha_1 f(X_1) + \ldots + \alpha_p f(X_p).$$

se démontre par récurrence sur p à partir des propriétés de la définition.

Application: Si E est de dimension n, et si $B_1, \ldots, B_n$ est une base de E, alors tout vecteur X de E admet une décomposition unique sur cette base: $X = \alpha_1 B_1 + \ldots + \alpha_n B_n$, et si f est une application linéaire, l'image de X est la combinaison linéaire sur les images de la base : $f(X) = \alpha_1 f(B_1) + \ldots + \alpha_n f(B_n)$.
Donc:

Proposition 13 *Une application linéaire de E dans F est entièrement déterminée par la donnée des images des vecteurs d'une base de E.*

Exemple:

Soit f de $\mathbb{R}^3$ dans $\mathbb{R}^3$ donnée par les images de la base canonique:

$$f(E_1) = \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}, \quad f(E_2) = \begin{pmatrix} 3 \\ 1 \\ 2 \end{pmatrix}, \quad f(E_3) = \begin{pmatrix} 2 \\ 1 \\ 3 \end{pmatrix},$$

alors si $X = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = x_1 E_1 + x_2 E_2 + x_3 E_3$, alors $f(x) = \begin{pmatrix} x_1 + 3x_2 + 2x_3 \\ 2x_1 + x_2 + x_3 \\ x_1 + 2x_2 + 3x_3 \end{pmatrix}$.

5.2 Matrice associée à une application linéaire.

Soit f une application linéaire de E dans F, avec $dim(E) = m$ et $dim(F) = n$. On munit E de la base $B_E = (U_1, \ldots, U_m)$ et F de la base $B_F = (V_1, \ldots, V_n)$. Pour connaitre f il suffit de connaitre les composantes des m vecteurs $f(U_1)$, $\ldots$, $f(U_m)$ dans la base B_F. Donc f est caractérisée par $n \cdot m$ réels. Si

$$f(U_j) = A_{1,j} V_1 + \ldots + A_{n,j} V_n,$$

la matrice $A(f)$ représentant f dans les bases B_E et B_F est:

$$A(f) = \begin{pmatrix} A_{1,1} & A_{1,2} & \cdots & A_{1,j} & \cdots & A_{1,m} \\ A_{2,1} & A_{2,2} & \cdots & A_{2,j} & \cdots & A_{2,m} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ A_{n,1} & A_{n,2} & \cdots & A_{n,j} & \cdots & A_{n,m} \end{pmatrix}$$

$n \times m$-matrice: dans les colonnes les composantes des $f(U_j)$ dans la base B_F.

Si les coordonnées de X dans la base B_E sont $\begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix}$ et si les coordonnées de $Y = f(X)$ dans la base B_F sont $\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$, alors on a

$$f(X) = \sum_{j=1}^{m} x_j f(U_j) = \sum_{i=1}^{n} \left(\sum_{j=1}^{m} A_{i,j} x_j \right) V_i$$

soit pour tout $i = 1, \ldots, n$

$$y_i = \sum_{j=1}^{m} A_{i,j} x_j$$

ce qui s'écrit:

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} A_{1,1} & A_{1,2} & \cdots & A_{1,j} & \cdots & A_{1,m} \\ A_{2,1} & A_{2,2} & \cdots & A_{2,j} & \cdots & A_{2,m} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ A_{n,1} & A_{n,2} & \cdots & A_{n,j} & \cdots & A_{n,m} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix}$$

Exemples:

- Soit U est un vecteur fixé de E, et soit f l'application qui à tout vecteur X de E associe $f(X) = \langle U, X\rangle$. Si l'on munit E d'une base orthonormée $B_E = (B_1, \dots B_m)$, on sait que si $U_1, \dots, U_m$ sont les coordonnées de U dans cette base et si $X_1, \dots, X_m$ sont les coordonnées de X dans cette base, ona $f(X) = U_1X_1 + \dots U_mX_m$. Si l'on choisit dans $\mathbb{R}$ le vecteur de base (1), la matrice qui représente f dans les bases B_E et (1) est la matrice $1 \times m$ ("vecteur ligne") $(U_1, \dots U_m)$, soit le transposé de $\begin{pmatrix} U_1 \\ \vdots \\ U_m \end{pmatrix}$.
- Soit f l'application linéaire de $\mathbb{R}^3$ dans $\mathbb{R}^3$ telle que
$$f(\begin{pmatrix} x \\ y \\ z \end{pmatrix}) = \begin{pmatrix} x+y+z \\ x+2y+3z \\ 2x+3y+2z \end{pmatrix}.$$
Alors dans la base canonique la matrice représentant f est
$$\begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \\ 2 & 3 & 2 \end{pmatrix}.$$
Si l'on munit maintenant $\mathbb{R}^3$ de la base $B = (U_1, U_2, U_3)$ avec
$$U_1 = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}, \quad U_2 = \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}, \quad U_3 = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix},$$
on a $f(U_1) = f(E_1)+f(E_2) = 2E_1+3E_2+5E_3$, $f(U_2) = f(E_2)+f(E_3) = 2E_1+5E_2+5E_3$, $f(U_3) = f(E_1)+f(E_3) = 2E_1+4E_2+4E_3$, et la matrice représentant f dans la base B et la base canonique est
$$\begin{pmatrix} 2 & 2 & 2 \\ 3 & 5 & 4 \\ 5 & 5 & 4 \end{pmatrix}.$$
On a aussi $f(U_1) = 3U_2 + 2U_3$, $f(U_2) = U_1 + 4U_2 + U_3$, $f(U_3) = U_1 + 3E_2 + U_3$, et la matrice représentant f dans la base B est
$$\begin{pmatrix} 0 & 1 & 1 \\ 3 & 4 & 3 \\ 2 & 1 & 1 \end{pmatrix}.$$
Il est donc indispensable de préciser les deux bases (de départ et d'arrivée) dans lesquelles la matrice représente l'application!

5.3 Opérations.

5.3.1 Addition et multiplication par un réel.

Proposition 14 *La somme de deux applications linéaires de E dans F est une application linéaire de E dans F.*
La multiplication d'une application linéaire de E dans F par un réel est une application linéaire de E dans F.

Preuve:
Soient f et g deux applications linéaires de E dans F. Pour tous vecteurs X et Y de E, on a

$$\begin{aligned}(f+g)(X+Y) &= f(X+Y)+g(X+Y) \quad \text{par définition de la somme de deux fonctions}\\ &= f(X)+f(Y)+g(X)+g(Y) \quad \text{car f et g sont linéaires}\\ &= (f+g)(X)+(f+g)(Y)\end{aligned}$$

et de même pour tout réel α

$$\begin{aligned}(f+g)(\alpha X) &= f(\alpha X)+g(\alpha X)\\ &= \alpha f(X)+\alpha g(X)\\ &= \alpha(f+g)(X).\end{aligned}$$

Ensuite pour tout réel λ, pour tous vecteurs X et Y de E, on a

$$\begin{aligned}(\lambda f)(X+Y) &= \lambda f(X+Y) \quad \text{par définition de la multiplication d'une fonction par un réel}\\ &= \lambda(f(X)+f(Y)) \quad \text{car f est linéaire}\\ &= (\lambda f)(X)+(\lambda f)(Y)\end{aligned}$$

et de même pour tout réel α

$$\begin{aligned}(\lambda f)(\alpha X) &= \lambda f(\alpha X)\\ &= \lambda\alpha f(X)\\ &= \alpha(\lambda f)(X).\end{aligned}$$

On peut montrer :

Proposition 15 *L'ensemble $\mathcal{L}(E,F)$ des applications linéaires d'un espace vectoriel E dans un espace vectoriel F muni de l'addition des fonctions et de la multiplication par un réel est un espace vectoriel.*

Addition des matrices et multiplication par un réel.
Soient f et g deux applications linéaires de E dans F et A et B les matrices qui les représentent dans les bases B_E et B_F. Alors la matrice qui représente $f+g$ est la matrice $A+B$ obtenue en additionnant terme à terme.
De même si λ est un réel, la matrice qui représente λf est la matrice λA obtenue en multipliant chaque terme par λ.
On peut montrer :

Proposition 16 *L'ensemble $\mathcal{M}_{n,m}$ des matrices $n \times m$ muni de l'addition et dela multiplication par un réel est un espace vectoriel.*

5.3.2 Composition des applications et multiplication des matrices.

Proposition 17 *La composition d'une application linéaire de E dans F et d'une application linéaire de F dans G est une application linéaire de E dans G.*

Preuve.
Soit f une application de E dans F et g une application de F dans G. L'application composée $g \circ f$ associe à tout vecteur X de E le vecteur de G $(g \circ f)(X) = g[f(X)]$. Soient X et Y deux vecteurs quelconques de E. On a

$$\begin{aligned} (g \circ f)(X+Y) &= g[f(X+Y)] \quad \text{par définition de la composition des fonctions} \\ &= g[f(X)+f(Y)] \quad \text{car f est linéaire} \\ &= g[f(X)]+g[f(Y)] \quad \text{car f est linéaire} \\ &= (g \circ f)(X)+(g \circ f)(Y). \end{aligned}$$

De même pour tout réel α

$$\begin{aligned} (g \circ f)(\alpha X) &= g[f(\alpha X)] \\ &= g[\alpha f(X)] \\ &= \alpha g[f(X)] \\ &= \alpha (g \circ f)(X). \end{aligned}$$

Si $dim(E) = m$, $dim(F) = n$ et $dim(G) = p$, si $B_E = (U_1, \ldots, U_m)$, $B_F = (V_1, \ldots, V_n)$, $B_G = (W_1, \ldots, W_p)$ sont des bases de E, F, et G, si A est la matrice $n \times m$ qui représente f dans les bases B_E et B_F, et B la matrice $p \times n$ qui représente g dans les bases B_F et B_G, soit C la matrice $p \times m$ qui représente $g \circ f$ dans les bases B_E et B_G. Alors les colonnes de C sont les coordonnées dans B_G des images des vecteurs de B_E par $g \circ f$.

$$\begin{aligned} (g \circ f)(U_k) &= g\left(\sum_{j=1}^{n} A_{j,k} V_j\right) \\ &= \sum_{j=1}^{n} A_{j,k} g(V_j) \\ &= \sum_{j=1}^{n} A_{j,k} \sum_{i=1}^{p} B_{i,j} W_i \end{aligned}$$

ce qui montre que le terme d'indice i, k de C est $\sum_{j=1}^{n} B_{i,j} A_{j,k}$.
Produit de matrices.

On appelle alors produit de B par A et on note BA la matrice C ci-dessus. Autrement dit, si B est une matrice $p \times n$ et A une matrice $n \times m$, BA est la matrice $p \times m$ de terme général

$$(BA)_{i,k} = \sum_{j=1}^{n} B_{i,j} A_{j,k}.$$

Exemple:
Soit E de dimension 2 de base $B_E = (U_1, U_2)$. Soit f l'application linéaire donnée par $f(U_1) = U_2$ et $f(U_2) = -U_1 + 4U_2$. Soit g l'application linéaire donnée par $g(U_1) = U_1 - U_2$ et $g(U_2) = 2U_1 - 4U_2$. Dans B_E la matrice représentant f est $A = \begin{pmatrix} 0 & -1 \\ 1 & 4 \end{pmatrix}$, la matrice représentant g est $B = \begin{pmatrix} 1 & 2 \\ -1 & -4 \end{pmatrix}$.
La matrice représentant $g \circ f$ est

$$BA = \begin{pmatrix} 1 & 2 \\ -1 & -4 \end{pmatrix} \begin{pmatrix} 0 & -1 \\ 1 & 4 \end{pmatrix} = \begin{pmatrix} 2 & 7 \\ -4 & 17 \end{pmatrix}$$

et la matrice représentant $f \circ g$ est

$$AB = \begin{pmatrix} 0 & -1 \\ 1 & 4 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ -1 & -4 \end{pmatrix} = \begin{pmatrix} 1 & 4 \\ -3 & -14 \end{pmatrix}.$$

Sur cet exemple, on voit que le produit de matrice n'est pas commutatif.

5.3.3 Transposition

Soit A une matrice $m \times n$. La matrice A^T est la matrice $n \times m$ dont les colonnes sont les lignes de A, autrement dit dont le terme général est donné par

$$\left(A^T\right)_{i,j} = A_{j,i}.$$

On a

$$\left(A^T\right)^T = A,$$

et l'on vérifie facilement que si B est une matrice $p \times n$ et A une matrice $n \times m$, $(BA)^T$ est une matrice $m \times p$ et

$$(BA)^T = A^T B^T.$$

En effet, pour tout $i = 1, \ldots, m$ et $j = 1, \ldots, p$:

$$\begin{aligned} \left((BA)^T\right)_{i,j} &= (BA)_{j,i} \\ &= \sum_{k=1}^{n} B_{j,k} A_{k,i} \\ &= \sum_{k=1}^{n} (A^T)_{i,k} (B^T)_{k,j} \\ &= \left(A^T B^T\right)_{i,j} \end{aligned}$$

Si l'on interprète A comme représentant une application linéaire f de $\mathbb{R}^n$ dans $\mathbb{R}^m$ alors A^T représente une application linéaire f^* de $\mathbb{R}^m$ dans $\mathbb{R}^n$ qui vérifie:

$$\forall X \in \mathbb{R}^n, \forall Y \in \mathbb{R}^m, \ \langle f(X), Y \rangle = \langle X, f^*(Y) \rangle.$$

En effet, $\langle f(X), Y \rangle = (AX)^T Y = X^T A^T Y = \langle X, f^*(Y) \rangle$.

5.4 Image et noyau.

Définition 17 *Si f est une application linéaire de E dans F, l'image de f, notée $Im(f)$ est l'ensemble des vecteurs $f(X)$ où X parcourt E:*

$$Im(f) = \{f(X) \ : \ X \in E\} = \{Y \in F \ : \ \exists X \in E, Y = f(X)\}.$$

Le noyau de f, noté $Ker(f)$, est l'ensemble des vecteurs X dont l'image par f est le vecteur nul de F:

$$Ker(f) = \{X \in E \ : \ f(X) = 0\}.$$

Exemples:

- Soit f de $\mathbb{R}^3$ dans $\mathbb{R}^3$ donnée par $f\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} x+y+2z \\ x+2y+3z \\ 2x+3y+5z \end{pmatrix}$.

  $Im(f)$ est l'ensemble des vecteurs de la forme

  $$\begin{pmatrix} x+y+2z \\ x+2y+3z \\ 2x+3y+5z \end{pmatrix} = x\begin{pmatrix} 1 \\ 1 \\ 2 \end{pmatrix} + y\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} + z\begin{pmatrix} 2 \\ 3 \\ 5 \end{pmatrix} = (x+z)\begin{pmatrix} 1 \\ 1 \\ 2 \end{pmatrix} + (y+z)\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}.$$

  $Im(f)$ est donc engendré par $\begin{pmatrix} 1 \\ 1 \\ 2 \end{pmatrix}$ et $\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$ qui sont linéairement indépendants, donc $Im(f)$ est un plan de $\mathbb{R}^3$.

  $Ker(f)$ est l'ensemble des vecteurs $\begin{pmatrix} x \\ y \\ z \end{pmatrix}$ tels que

  $$\begin{cases} x+y+2z = 0 \\ x+2y+3z = 0 \\ 2x+3y+5z = 0 \end{cases}$$

  soit $x = y = -z$, c'est la droite engendrée par $\begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix}$.

- Soit g de $\mathbb{R}^3$ dans $\mathbb{R}^2$ donnée par $g\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} x-y+z \\ 2x-z \end{pmatrix}$.

  $Im(g)$ est l'ensemble des vecteurs de la forme

  $$\begin{pmatrix} x-y+z \\ 2x-z \end{pmatrix} = x\begin{pmatrix} 1 \\ 2 \end{pmatrix} + y\begin{pmatrix} -1 \\ 0 \end{pmatrix} + z\begin{pmatrix} 1 \\ -1 \end{pmatrix} = (x-\frac{z}{2})\begin{pmatrix} 1 \\ 2 \end{pmatrix} + (y-\frac{3z}{2})\begin{pmatrix} -1 \\ 0 \end{pmatrix}.$$

  $Im(g)$ est donc engendré par $\begin{pmatrix} 1 \\ 2 \end{pmatrix}$ et $\begin{pmatrix} -1 \\ 0 \end{pmatrix}$ qui sont linéairement indépendants, donc $Im(g) = \mathbb{R}^2$.

$Ker(g)$ est l'ensemble des vecteurs $\begin{pmatrix} x \\ y \\ z \end{pmatrix}$ tels que

$$\begin{cases} x - y + z = 0 \\ 2x - z = 0 \end{cases}$$

soit $z = 2x$ et $y = 3x$, c'est la droite engendrée par $\begin{pmatrix} 1 \\ 3 \\ 2 \end{pmatrix}$.

Proposition 18 *Si f est une application linéaire de E dans F, $Im(f)$ est un sous espace vectoriel de F et $Ker(f)$ est un sous espace vectoriel de E.*

Preuve:
Si Y_1 et Y_2 sont deux éléments de $Im(f)$, il existe X_1 et X_2 vecteurs de E tels que $f(X_1) = Y_1$ et $f(X_2) = Y_2$. On a alors $Y_1 + Y_2 = f(X_1) + f(X_2)$, et donc $(Y_1 + Y_2) \in Im(f)$.
Si maintenant α est un réel, $\alpha Y_1 = \alpha f(X_1) = f(\alpha X_1)$ et donc $\alpha Y_1 \in Im(f)$. $Im(f)$ est donc un sous-espace vectoriel de F.
De même, si X_1 et X_2 sont deux éléments de $Ker(f)$, alors $f(X_1) = 0$ et $f(X_2) = 0$. On a alors $0 = f(X_1) + f(X_2) = f(X_1 + X_2)$, et donc $(X_1 + X_2) \in Ker(f)$.
Si maintenant α est un réel, $f(\alpha X_1) = \alpha f(X_1) = 0$ et donc $\alpha X_1 \in Ker(f)$. $Ker(f)$ est donc un sous-espace vectoriel de E.

Théorème 2 *Si f est une application linéaire de E dans F, alors*

$$dim(Im(f)) + dim(Ker(f)) = dim(E)$$

Preuve:
Soit f une application linéaire de E dans F, soit p la dimension de E et m la dimension de $Ker(f)$. Soit $B_1, \ldots, B_p$ une base de E telle que $B_1, \ldots B_m$ soit une base de $Ker(f)$.
Pour tous réels $\alpha_1, \ldots, \alpha_p$, $f(\alpha_1 B_1 + \ldots + \alpha_p B_p) = \alpha_1 f(B_1) + \ldots + \alpha_p f(B_p) = \alpha_{m+1} f(B_{m+1}) + \ldots + \alpha_p f(B_p)$ ($= 0$ si $p = m$).
Si Y est dans $Im(f)$, il existe un vecteur X de E tel que $f(X) = Y$, et donc des réels $\alpha_{m+1}, \ldots, \alpha_p$ tels que $Y = \alpha_{m+1} f(B_{m+1}) + \ldots + \alpha_p f(B_p)$. Donc $f(B_{m+1}), \ldots, f(B_p)$ engendrent $Im(f)$. Si ces vecteurs étaient liés, il existerait des réels $\beta_{m+1}, \ldots, \beta_p$ tels que $\beta_{m+1} f(B_{m+1}) + \ldots + \beta_p f(B_p) = 0$, donc tels que $f(\beta_{m+1} B_{m+1} + \ldots + \beta_p B_p) = 0$, soit tels que $\beta_{m+1} B_{m+1} + \ldots + \beta_p B_p \in Ker(f)$, donc combinaison linéaire de $B_1, \ldots, B_m$, ce qui est impossible puisque $B_1, \ldots, B_p$ est une base de E. Donc $f(B_{m+1}), \ldots, f(B_p)$ forme une base de $Im(f)$, et la dimension de $Im(f)$ est $p - m$.

Définition 18 *On appelle rang d'une application linéaire f (ou rang d'une matrice A) la dimension de l'image de f (ou de l'application linéaire que représente A)*

Remarques:

1. Une matrice peut toujours représenter une application linéaire (prendre les bases canoniques). La dimension de l'image ne dépend pas du choix de la base !
2. Pour une matrice A $n \times m$, on a $rg(A) \leq m$ (pas plus de m colonnes) et $rg(A) \leq n$ (pas plus de n vecteurs linéairement indépendants en dimension n).
3. $rg(A) = rg(A^T)$ (résultat admis): donc on peut aussi regarder combien de lignes sont linéairement indépendantes.

Reprise des exemples:

- La matrice représentant f est
$$A = \begin{pmatrix} 1 & 1 & 2 \\ 1 & 2 & 3 \\ 2 & 3 & 5 \end{pmatrix}$$
et l'on voit $rg(A) = 2$, donc $dim(Ker(f)) = 1$.
- La matrice représentant g est
$$B = \begin{pmatrix} 1 & -1 & 1 \\ 2 & 0 & -1 \end{pmatrix}$$
et l'on voit $rg(A) = 2$, donc $dim(Ker(f)) = 1$.

Définition 19 *Une application f de E dans F est dite:*

- *injective si et seulement si tout élément de F a au plus un antécédent dans E, ou encore si et seulement si: $(X \in E, X' \in E, f(X) = f(X')) \Longrightarrow (X = X')$,*
- *surjective si et seulement si tout élément de F a au moins un antécédent dans E, c'est à dire si et seulement si $Im(f) = F$.*
- *bijective si elle est injective et surjective, soit si tout élément de F a exactement un antécédent dans E.*

Remarque: Si une application est injective, alors on peut l' inverser: il existe une fonction f^{-1} de F dans E telle que $f^{-1} \circ f = I_E$ et $f \circ f^{-1} = I_F$.

Proposition 19

1. *f est injective si et seulement si $Ker(f) = \{0\}$.*
2. *f est injective si et seulement si $Rang(f) = dim(E)$.*
3. *f est surjective si et seulement si $Rang(f) = dim(F)$.*
4. *f est bijective si et seulement si $Rang(f) = dim(F) = dim(E)$.*

Preuve:

1. Si f est injective, alors 0 est l'unique antcédent de 0, donc $Ker(f) = \{0\}$. Réciproquement, supposons $Ker(f) = \{0\}$. Soient X et X' tels que $f(X) = f(X')$. Alors $f(X - X') = f(X) - f(X') = 0$, donc $(X - X') \in Ker(f)$, donc $X - X' = 0$, et f est injective.
2. f est injective si et seulement si $Ker(f) = \{0\}$, donc si et seulement si $dim(Ker(f)) = 0$ donc si et seulement si $Rang(f) = dim(E)$ par le théorème des dimensions.
3. f est surjective si et seulement si $Im(f) = F$, donc si et seulement si $Rang(f) = dim(F)$ puisque $Im(f)$ est un sous-espace vectoriel de F.
4. En appliquant 2. et 3., f est bijective si et seulement f est injective et surjective donc si et seulement si $Rang(f) = dim(F) = dim(E)$.

Exemples:
Dans les exemples précédents: f n'est ni injective ni surjective; g est surjective et pas injective.

5.5 Endomorphismes et matrices carrées.

Définition 20 *On appelle endomorphisme de E une application linéaire de E dans E. La matrice associée est carrée $n \times n$ si $dim(E) = n$.*

Proposition 20 *Un endomorphisme de E est bijectif si et seulement si il est injectif ou encore si et seulement si il est surjectif.*

Si f est un endomorphisme de E bijectif, alors son inverse f^{-1} existe. Si E est muni d'une base et que A est la matrice $n \times n$ qui représente f, la matrice B qui représente f^{-1} vérifie, du fait que le produit des matrices représente la composition des applications: $A \cdot B = B \cdot A = I_n$; on la note A^{-1}: c'est l'inverse de A pour la multiplication des matrices.

Exemple:
Soit f de $\mathbb{R}^3$ dans $\mathbb{R}^3$ donnée par $f\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} x+y+z \\ x+2y+3z \\ 2x+3y+2z \end{pmatrix}$.

$Ker(f)$ est l'ensemble des vecteurs $\begin{pmatrix} x \\ y \\ z \end{pmatrix}$ tels que

$$\begin{cases} x+y+z=0 \\ x+2y+3z=0 \\ 2x+3y+2z=0 \end{cases}$$

soit $x = y = z = 0$, donc f est injective, donc bijective.
Soit $\begin{pmatrix} x' \\ y' \\ z' \end{pmatrix} = f\begin{pmatrix} x \\ y \\ z \end{pmatrix}$, donc tel que

$$\left\{\begin{array}{l} x' = x + y + z \\ y' = x + 2y + 3z \\ z' = 2x + 3y + 2z \end{array}\right.$$

qui donne

$$\left\{\begin{array}{l} x = \frac{1}{2}(5x' - y' - z') \\ y = \frac{1}{2}(-4x' + 2z') \\ z = \frac{1}{2}(x' + y' - z') \end{array}\right.$$

et l'inverse de f est donnée par $f^{-1}\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} \frac{1}{2}(5x - y - z) \\ \frac{1}{2}(-4x + 2z) \\ \frac{1}{2}(x + y - z) \end{pmatrix}$.
L'inverse de la matrice $\begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \\ 2 & 3 & 2 \end{pmatrix}$ est la matrice $\begin{pmatrix} \frac{5}{2} & \frac{-1}{2} & \frac{-1}{2} \\ -2 & 0 & 1 \\ \frac{1}{2} & \frac{1}{2} & \frac{-1}{2} \end{pmatrix}$.

5.6 Changement de base.

Si $B = (B_1, \ldots, B_n)$ est une base de $\mathbb{R}^n$, si le vecteur X a pour coordonnées $\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$ dans la base $E = (E_1, \ldots, E_n)$, l'application f qui à X associe ses coordonnées dans la base B est une application linéaire de $\mathbb{R}^n$ dans $\mathbb{R}^n$, de matrice représentative dans les bases E et B la matrice A dont les colonnes sont les coordonnées des E_j dans la base B.
En effet, soit $A_{1,j}, \ldots, A_{n,j}$ les coordonnées de E_j dans la base B. On a

$$X = \sum_{j=1}^{n} x_j E_j = \sum_{j=1}^{n} x_j \sum_{i=1}^{n} A_{i,j} B_i$$

et donc, si $f(X) = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$, c'est à dire si $X = \sum_{i=1}^{n} y_i B_i$, on a pour tout i $y_i = \sum_{j=1}^{n} A_{i,j} x_j$, et donc

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} A_{1,1} & \ldots & A_{1,n} \\ \vdots & \vdots & \vdots \\ A_{n,1} & \ldots & A_{n,n} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}.$$

Par ailleurs, réciproquement, l'application qui associe pour tout vecteur X ses coordonnées dans la base E à ses coordonnées dans la base B est l'application linéaire inverse de f, et la matrice qui la représente dans les bases B et E est la matrice dont les colonnes sont les coordonnées des B_j dans la base E. Mais cette matrice est la matrice inverse de A, donc :

Proposition 21 *Si une matrice A a pour colonnes les coordonnées des vecteurs d'une base E dans une autre base B, alors A est inversible et A^{-1} a pour colonnes les coordonnées des vecteurs de B dans la base E.*

Exemple:

Soit B la base de $\mathbb{R}^2$ formée par les vecteurs $B_1 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$ et $B_2 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$. On a $E_1 = \frac{1}{3}(B_1 + 2B_2)$ et $E_2 = \frac{1}{3}(B_1 - B_2)$. Donc l'inverse de la matrice $A = \begin{pmatrix} \frac{1}{3} & \frac{1}{3} \\ \frac{2}{3} & -\frac{1}{3} \end{pmatrix}$ est la matrice $A^{-1} = \begin{pmatrix} 1 & 1 \\ 2 & -1 \end{pmatrix}$.

Si maintenant les deux bases sont orthonormées, et si l'on calcule $A^T A$, à la place i, j:

$$\left(A^T A\right)_{i,j} = \sum_{k=1}^{n} (A^T)_{i,k} A_{k,j} = \sum_{k=1}^{n} A_{k,i} A_{k,j} = \langle E_i, E_j \rangle$$

car la base B est orthonormée (puisque son calcul ne change pas si on prend les coordonnées dans une base orthonormée), et comme la base E est orthonormée, $(A^T A)_{i,j} = 1$ si $i = j$ et $(A^T A)_{i,j} = 0$ si $i \neq j$, donc $A^T A = I_n$, donc :

Proposition 22 *Si A est une matrice $n \times n$ dont les colonnes sont les coordonnées d'une base orthonormée dans une autre base orthonormée, alors $A^{-1} = A^T$, et donc $A^T A = A A^T = I_n$.*

5.7 Projection orthogonale (bis).

Soit D une droite de $\mathbb{R}^n$, engendrée par le vecteur U de norme 1.
Pour tout vecteur X de $\mathbb{R}^n$, la projection orthogonale de X sur D est:

$$P(X) = \langle X, U \rangle U$$

et la fonction P qui à X associe sa projection orthogonale $P(X)$ est une application linéaire de $\mathbb{R}^n$ dans D.
On peut aussi considérer P comme une application linéaire de $\mathbb{R}^n$ dans $\mathbb{R}^n$.

De manière générale, si E est un sous espace vectoriel de $\mathbb{R}^n$, définissons l'orthogonal de E par

$$E^\perp = \{X \in \mathbb{R}^n \ : \forall U \in E, \langle X, U \rangle = 0\}.$$

C'est un sous-espace vectoriel de $\mathbb{R}^n$ (exercice : le démontrer), et on a:

Proposition 23 *La fonction P qui à tout vecteur X associe sa projection orthogonale $P(X)$ sur E est une application linéaire de $\mathbb{R}^n$ dans $\mathbb{R}^n$ dont l'image est E et dont le noyau est $E^\perp$.*

Preuve:
Soit p la dimension de E, et soit $B = (B_1, \dots, B_n)$ une base orthonormée de $\mathbb{R}^n$ telle que $B_1, \dots, B_p$ soit une base de E. Alors $B_{p+1}, \dots, B_n$ est une base orthonormée de $E^\perp$. En effet, pour tout X de $\mathbb{R}^n$, la décomposition de X sur la base B est

$$X = \sum_{i=1}^{n} \langle X, B_i \rangle B_i.$$

Si $X \in E^\perp$, pour tout $i = 1, \dots, p$, $B_i \in E$ et $\langle X, B_i \rangle = 0$, donc X est combinaison linéaire de $B_{p+1}, \dots, B_n$ qui forment donc un système générateur de $E^\perp$. Comme par ailleurs ils sont linéairement indépendants, ils forment une base de $E^\perp$, qui est orthonormée.
Maintenant $P(X)$ est le vecteur de E tel que $X - P(X)$ est rothogonal à E, autrement dit tel que $(X - P(X)) \in E^\perp$. On obtient donc facilement (exercice: le démontrer) que

$$P(X) = \sum_{i=p+1}^{n} \langle X, B_i \rangle B_i.$$

On voit donc facilement (exercice: le démontrer) que P est une application linéaire, et que $P(X) = 0$ si et seulement si $\langle X, B_1 \rangle = \dots = \langle X, B_p \rangle = 0$, donc $Ker(P)$ est l'ensemble des vecteurs du sous espace engendré par $B_{p+1}, \dots, B_n$, c'est à dire $E^\perp$.
Il est clair que $Im(P) \subset E$. Par ailleurs tout vecteur de E est sa propre projection orthogonale, donc $Im(P) = E$.

Si A est la matrice qui représente la projection P dans la base $B_1, \dots, B_n$, on a

$$A = \begin{pmatrix} & & & 0 & \dots & 0 \\ & I_p & & \vdots & \dots & \vdots \\ & & & 0 & \dots & 0 \\ 0 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \dots & \dots & \vdots & \dots & \vdots \\ 0 & \dots & 0 & 0 & \dots & 0 \end{pmatrix}$$

Si par contre on veut représenter P dans une autre base, on verra cela plus loin.

6 Vecteurs gaussiens 2.

6.1 Application linéaire et vecteur gaussien.

Proposition 24 *Soit X un vecteur gaussien de $\mathbb{R}^n$ de loi $\mathcal{N}_n(m, \Sigma)$, et soit f une application linéaire de $\mathbb{R}^n$ dans $\mathbb{R}^p$. Alors $f(X)$ est un vecteur gaussien de $\mathbb{R}^p$, et si A est la matrice $p \times n$ qui repésente f dans les bases canoniques, alors la loi de $f(X) = AX$ est $\mathcal{N}_p(Am, A\Sigma A^T)$.*

Preuve:

Si $X = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}$, si $f(X) = Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_p \end{pmatrix}$ (coordonnées dans la base canonique), on a $X = A \cdot Y$. Toute combinaison linéaire des Y_i est une combinaison linéaire des X_i, et est donc une variable aléatoire réelle gaussienne, donc Y est un vecteur gaussien p-dimensionnel.
Si $A = (A_{i,j})$, on a

$$E(Y_i) = E\left(\sum_{j=1}^{n} A_{i,j} X_j\right) = \sum_{j=1}^{n} A_{i,j} E(X_j) = \sum_{j=1}^{n} A_{i,j} m_j = (A \cdot m)_i,$$

donc $E(Y) = Am$.

$$\begin{aligned} Cov(Y_i, Y_j) &= Cov\left(\sum_{k=1}^{n} A_{i,k} X_k, \sum_{l=1}^{n} A_{j,l} X_l\right) \\ &= \sum_{k=1}^{n} \sum_{l=1}^{n} A_{i,k} A_{j,l} Cov(X_k, X_l) \\ &= \sum_{k=1}^{n} \sum_{l=1}^{n} A_{i,k} \Sigma_{k,l} A^T_{l,j} \\ &= \left(A \cdot \Sigma \cdot A^T\right)_{i,j}, \end{aligned}$$

donc la matrice de variance de Y est $A \cdot \Sigma \cdot A^T$.

Remarques:

- Le calcul d'espérance et de covariance est valable en dehors du cadre gaussien, c'est à dire que si X est un vecteur aléatoire n-dimensionnel et si A est une matrice $p \times n$, AX est un vecteur aléatoire p-dimensionnel d'espérance $E(AX) = AE(X)$ et de matrice de variance $Var(AX) = AVar(X)A^T$.
- On retrouve le fait que si l'on part d'un n-échantillon gaussien et qu'on fait le changement de coordonnées dans une base orthonormée, on retrouve un n-échantillon gaussien. En effet, si $X \sim \mathcal{N}_n(0, I_n)$, si Y sont les coordonnées de X dans une base orthonormée, alors $Y = AX$ où A est la matrice dont les colonnes sont les coordonnées des vecteurs canoniques dans le

nouvelle base. On a alors que $Y \sim \mathcal{N}_n(0, AA^T)$, et $AA^T = I_n$ puisque les bases sont orthonormées.

Théorème 3 (Théorème de Cochran.)
Soit X un n-échantillon gaussien de $\mathcal{N}(0,\sigma^2)$, c'est à dire $X \sim \mathcal{N}_n(0,\sigma^2 I_n)$.

1. *Les composantes de X dans toute base orthonormée de $\mathbb{R}^n$ forment un n-échantillon gaussien de $\mathcal{N}(0,\sigma^2)$.*
2. *Si $E_1, \ldots, E_p$ sont p espaces vectoriels de $\mathbb{R}^n$ deux à deux orthogonaux de dimensions $r_1, \ldots, r_p$ et si $P_1(X), \ldots, P_p(X)$ sont les projections orthogonales de X sur $E_1, \ldots, E_p$, alors les vecteurs aléatoires $P_1(X), \ldots, P_p(X)$ sont indépendants et les variables aléatoires $\| P_1(X) \|^2, \ldots, \| P_1(X) \|^2$ sont indépendantes de loi $\sigma^2\chi^2(r_1), \ldots, \sigma^2\chi^2(r_p)$.*

Preuve:

1. Est vu dans la remarque précédente.
2. Soit $B = B_1, \ldots, B_n$ une base orthonormée de $\mathbb{R}^n$ telle que $B_1, \ldots, B_{r_1}$ soit une base orthonormée de E_1, $B_{r_1+1}, \ldots, B_{r_1+r_2}$ soit une base orthonormée de E_2, ..., $B_{r_1+\ldots+r_{p-1}+1}, \ldots, B_{r_1+\ldots+r_p}$ soit une base orthonormée de E_p, soit Y les coordonnées de X dans cette base, alors $Y \sim \mathcal{N}_n(0,\sigma^2 I_n)$, et

$$P_1(X) = Y_1 B_1 + \ldots + Y_{r_1} B_{r_1},$$

$$P_2(X) = Y_{r_1+1} B_{r_1+1} + \ldots, Y_{r_1+r_2} B_{r_1+r_2}, \ldots,$$

$$P_p(X) = Y_{r_1+\ldots+r_{p-1}+1} B_{r_1+\ldots+r_{p-1}+1} + \ldots + Y_{r_1+\ldots+r_p} B_{r_1+\ldots+r_p}.$$

Les Y_i étant deux à deux indépendants, on obtient que les $P_i(X)$ sont deux à deux indépendants. Par ailleurs

$$\| P_1(X) \|^2 = \sum_{i=1}^{r_1} Y_i^2,$$

c'est la somme des carrés de r_1 gaussiennes centrées de variance σ^2, et suit donc la loi $\sigma^2\chi^2(r_1)$. La preuve est la même pour les autres.

Exemple: moyenne et variance empirique:
Soient $X_1, \ldots, X_n$ des variables indépendantes de loi $\mathcal{N}(m,\sigma^2)$, autrement dit $X = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix} \sim \mathcal{N}_n(m\mathbf{1}_n, \sigma^2 I_n)$, où $\mathbf{1}_n = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$. On a $X - m\mathbf{1}_n \sim \mathcal{N}_n(0,\sigma^2 I_n)$. Soit E_1 la droite engendrée par $\mathbf{1}_n$, et E_2 le sous-espace vectoriel orthogonal dans $\mathbb{R}^n$. Alors $r_1 = 1$ et $r_2 = n-1$. La projection de $X - m\mathbf{1}_n$ sur E_1 est le vecteur

$$P_1(X - m\mathbf{1}_n) = (\bar{X} - m)\mathbf{1}_n = \left(\frac{\sum_{i=1}^n X_i}{n} - m \right) \mathbf{1}_n,$$

et

$$P_2(X - m\mathbf{1}_n) = (X - m\mathbf{1}_n) - P_1(X - m\mathbf{1}_n) = X - P_1(X) = \begin{pmatrix} X_1 - \bar{X} \\ \vdots \\ X_n - \bar{X} \end{pmatrix}.$$

Donc aussi

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2 = \frac{1}{n-1} \| P_2(X - m\mathbf{1}_n) \|^2 .$$

Donc $(n-1)s^2 \sim \sigma^2\chi^2(n-1)$, et comme $P_1(X - m\mathbf{1}_n)$ et $P_2(X - m\mathbf{1}_n)$ sont indépendants, $\bar{X}$ et s^2 sont indépendants.

6.2 Modèle linéaire gaussien.

6.2.1 Cadre général.

Définition 21 *Si Z_1 suit une loi $\chi^2(n_1)$, Z_2 suit une loi $\chi^2(n_2)$, si Z_1 et Z_2 sont indépendantes, alors*

$$F = \frac{Z_1/n_1}{Z_2/n_2}$$

suit la loi de Fisher à n_1 et n_2 degrés de liberté not'ee $F(n_1, n_2)$.

On considère les n variables aléatoires

$$X_i = m_i + \sigma\varepsilon_i, \quad i = 1, \dots, n$$

où les ε_i forment un n-échantillon gaussien, autrement dit si $X = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}$ et

$m = \begin{pmatrix} m_1 \\ \vdots \\ m_n \end{pmatrix}$, $X \sim \mathcal{N}(m, \sigma^2 I_n)$.

Le cadre général des modèles linéaires est celui où l'on ne connait pas m et σ^2, mais on sait que m est dans un sous-espace vectoriel connu E, et l'on veut tester l'appartenance à des sous-espaces vectoriels plus petits. Par exemple, on a deux échantillons gaussiens de même variance, et l'on veut tester l'égalité des moyennes. Cela revient à considérer que E est l'ensemble des vecteurs dont les n_1 premières coordonnées sont égales entre elles, et les $n_2 = n - n_1$ autres sont égales entre elles, puis que l'on veut tester si m est dans F, l'ensemble des vecteurs dont toutes les coordonnées sont égales entre elles. (E est alors un sous-espace vectoriel de dimension 2, et F un sous-espace vectoriel de E de dimension 1).

De manière générale, si E est de dimension k, et si P_E est la projection orthogonale sur E, on a, en appliquant le théorème de Cochran:

Théorème 4 *$P_E(X)$ est un estimateur sans biais de m (c'est à dire d'espérance m) indépendant de $X - P_E(X)$,*

$$S^2 = \frac{1}{n-k} \parallel X - P_E(X) \parallel^2$$

est un estimateur sans biais de σ^2, $\parallel P_E(X) - m \parallel^2$ a pour loi $\sigma^2\chi^2(k)$ et $\parallel X - P_E(X) \parallel^2$ a pour loi $\sigma^2\chi^2(n-k)$.

Si F est un sous-espace vectoriel de E de dimension l, et si l'on pense que m est dans F, il vaut mieux estimer m par $P_F(X)$. Si maintenant on veut tester H_0 : "$m \in F$" contre H_1 : "$m \in E$ mais $m \notin F$", on peut se dire que $P_E(X) - P_F(X)$ est une bonne variable de test, car si H_0 est vraie elle a tendance à être proche de 0, alors que si H_1 est vraie elle a tendance à être loin de 0. On mesure l'écart au vecteur nul par $\parallel P_E(X) - P_F(X) \parallel^2$. $P_E(X) - P_F(X)$ est la projection de X sur le sous-espace vectoriel de E qui est orthogonal à F, donc $\parallel P_E(X) - P_F(X) \parallel^2$ suit la loi $\sigma^2\chi^2(k-l)$. Comme σ^2 est inconnu, on divise par son estimateur S^2, qui se trouve être indépendant par le théorème de Cochran. On a donc obtenu:

Théorème 5 *Si F est un sous-espace vectoriel de E de dimension l, on teste H_0 : "$m \in F$" contre H_1 : "$m \in E$ mais $m \notin F$" à l'aide de*

$$\frac{\parallel P_E(X) - P_F(X) \parallel^2 /(k-l)}{\parallel X - P_E(X) \parallel^2 /(n-k)}$$

qui suit la loi $F(k-l, n-k)$ sous H_0 et qui a tendance à être plus grande sous H_1.

6.2.2 Analyse de la variance

L'analyse de la variance est la comparaison de p échantillons gaussiens indépendants de même variance. Autrement dit, on suppose que l'on a un n_1-échantillon X^1 de $\mathcal{N}(m_1, \sigma^2)$, un n_2-échantillon X^2 de $\mathcal{N}(m_2, \sigma^2)$, ..., un n_p-échantillon X^p de $\mathcal{N}(m_p, \sigma^2)$, avec $n_1 + n_2 + \ldots + n_p = n$. Cela rentre dans le cadre général du modèle linéaire, avec

$$X = \begin{pmatrix} X^1 \\ X^2 \\ \vdots \\ X^p \end{pmatrix}, \quad m = \begin{pmatrix} m_1 \mathbf{1}_{n_1} \\ m_2 \mathbf{1}_{n_2} \\ \vdots \\ m_p \mathbf{1}_{n_p} \end{pmatrix},$$

et donc E le sous-espace vectoriel de $\mathbb{R}^n$ de dimension p engendré par les vecteurs

$$\begin{pmatrix} 0 \\ \vdots \\ \\ m_i \mathbf{1}_{n_i} \\ 0 \\ \vdots \end{pmatrix}.$$

Alors

$$P_E(X) = \begin{pmatrix} \bar{X^1}\mathbf{1}_{n_1} \\ \bar{X^2}\mathbf{1}_{n_2} \\ \vdots \\ \bar{X^p}\mathbf{1}_{n_p} \end{pmatrix}, \quad \text{où } \bar{X^i} = \frac{1}{n_i}\sum_{j=1}^{n_i} X_j^i,$$

et

$$\| X - P_E(X) \|^2 = \sum_{i=1}^{p}\sum_{j=1}^{n_i} \left(X_j^i - \bar{X^i}\right)^2.$$

Si l'on veut tester l'homogénéité, c'est à dire H_0: "$m_1 = \ldots = m_p$" contre H_1: "c'est faux", on teste l'appartenance de m à F la droite engendrée par $\mathbf{1}_n$. On a

$$P_F(X) = \bar{X}\mathbf{1}_n = \left(\frac{1}{n}\sum_{i=1}^{p}\sum_{j=1}^{n_i} X_j^i\right)\mathbf{1}_n, \quad \| P_E(X) - P_F(X) \|^2 = \sum_{i=1}^{p}\sum_{j=1}^{n_i}\left(\bar{X^i} - \bar{X}\right)^2$$

Donc pour effectuer le test, on utilise la variable

$$\frac{\sum_{i=1}^{p} n_i \left(\bar{X^i} - \bar{X}\right)^2 / (p-1)}{\sum_{i=1}^{p}\sum_{j=1}^{n_i}\left(X_j^i - \bar{X^i}\right)^2 / (n-p)}$$

qui, sous H_0, suit la loi $F(p-1, n-p)$, et sous H_1 est plus grande.

6.2.3 Régression linéaire.

On veut expliquer la moyenne par une combinaison linéaire de p variables explicatives $\Phi^1, \ldots, \Phi^p$, c'est à dire que l'on suppose que les vecteurs Φ^j de $\mathbb{R}^n$ étant fixés, m est dans l'espace E engendré par les Φ^j, autrement dit qu'il existe des réels $a_1, \ldots, a_p$ tels que pour tout $i = 1, \ldots, n$

$$m_i = \sum_{j=1}^{p} a_j \Phi_i^j.$$

Il s'agit donc de calculer la projection orthogonale de X sur E.

Soit M la matrice $n \times p$ dont les colonnes sont les vecteurs Φ^j. On supposera que M est de rang p, ce qui signifie que les Φ^j sont linéairement indépendants, ou encore que E est de dimension p. Alors le vecteur $P_E(X)$ est tel que pour un vecteur $\hat{a}$ de $\mathbb{R}^p$, $\hat{a} = \begin{pmatrix} \hat{a}_1 \\ \vdots \\ \hat{a}_p \end{pmatrix}$, on a

$$P_E(X) = M\hat{a}, \quad \forall b \in \mathbb{R}^p, \ \langle Mb, X - M\hat{a}\rangle = 0,$$

soit pour tout b de $\mathbb{R}^p$, $b^T[M^TX - M^TM\hat{a}] = 0$, ce qui est vrai si $M^TM\hat{a} = M^TX$.

Maintenant si M est de rang p, $Mb = 0$ si et seulement si $b = 0$, donc $b^T M^T M b = 0$ si et seulement si $b = 0$, donc la matrice $p \times p$ $M^T M$ est de rang p, donc inversible. Donc

$$\hat{a} = \left(M^T M\right)^{-1} M^T X,$$

et l'on a:

Proposition 25 *Si E est le sous-espace vectoriel de $\mathbb{R}^n$ engendré par les vecteurs linéairement indépendants $\Phi^1, \ldots, \Phi^p$, si M est la matrice $n \times p$ dont les colonnes sont les vecteurs Φ^j, alors pour tout vecteur X de $\mathbb{R}^n$, la projection orthogonale de X sur E est le vecteur*

$$P_E(X) = M \left(M^T M\right)^{-1} M^T X.$$

Si l'on veut tester le fait que certaines des variables explicatives n'interviennent pas, on définit le sous-espace vectoriel F de E engendré par les Φ^j sauf ceux qui sont censés ne pas intervenir, et l'on peut calculer P_F de la même façon, avec la matrice M_F où l'on a enlevé les Φ^j qui sont censés ne pas intervenir.

7 Inversion des matrices. Systèmes linéaires.

7.1 Systèmes linéaires.

Ecriture générale d'un système linéaire de n équations avec m inconnues:

$$\begin{cases} a_{1,1}x_1 + a_{1,2}x_2 + \cdots + a_{1,j}x_j + \cdots + a_{1,m}x_m = b_1 \\ a_{2,1}x_1 + a_{2,2}x_2 + \cdots + a_{2,j}x_j + \cdots + a_{2,m}x_m = b_2 \\ \cdots\cdots\cdots \\ a_{n,1}x_1 + a_{n,2}x_2 + \cdots + a_{n,j}x_j + \cdots + a_{n,m}x_m = b_n \end{cases}$$

Un tel système peut avoir: une infinité de solutions, une solution unique, ou pas de solution.

Un tel système s'écrit matriciellement:

$$AX = B,$$

où $X = \begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix}$ est le vecteur inconnu de $\mathbb{R}^m$, $B = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix} \in \mathbb{R}^n$, et

$A = \begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,j} & \cdots & a_{1,m} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,j} & \cdots & a_{2,m} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{n,1} & a_{n,2} & \cdots & a_{n,j} & \cdots & a_{n,m} \end{pmatrix}$ est une matrice $n \times m$.

Commentaires:

1. Pour qu'il existe une solution, il faut que B soit dans l'image de A.
2. Pour qu'il y ait au plus une solution, il faut que le noyau de A soit réduit au vecteur nul (car si il y a deux solutions différentes X et X', alors $AX = AX'$ et $X - X'$ est dans le noyau de A.)
3. Si $m = n$, pour qu'il y ait une et une seule solution, il faut et il suffit que A soit inversible, auquel cas $X = A^{-1}B$.

Soit I_n la matrice identité $n \times n$.

Définition 22 *(Rappel). Une matrice $n \times n$ A est inversible si et seulement si il existe une matrice $n \times n$ B telle que $AB = BA = I_n$. B est alors l'inverse de A et notée A^{-1}.*

Propriétés des inverses:

1. Si A est inversible, A^{-1} l'est aussi et son inverse est A.
   Preuve: immédiate à partir de la définition.
2. Si A et B sont deux matrices inversibles $n \times n$, alors AB est inversible et $(AB)^{-1} = B^{-1}A^{-1}$.
   Preuve: on a $(B^{-1}A^{-1})AB = B^{-1}(A^{-1}A)B = B^{-1}I_nB = B^{-1}B = I_n$.

3. Si A est inversible, A^T l'est aussi et son inverse est $(A^{-1})^T$.
   Preuve: $(A^{-1})^T A^T = (AA^{-1})^T = I_n^T = I_n$.

Remarques: Revoir les inverses déjà vues au cours précédent.
On va voir une méthode générale pour décider si une matrice est inversible, et, si elle l'est, calculer son inverse.

7.2 Déterminants et inversion des matrices carrées.

On va donner toutes les notions et les résultats dans le cas général, mais sans preuve. A admettre.

Déterminants des matrices 1×1: une matrice 1×1 est un réel, son déterminant est ce réel.
Déterminants des matrices 2×2: une matrice $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ a pour déterminant $det(A) = ad - bc$.
Déterminant des matrices $n \times n$:

Soit $A = \begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,j} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,j} & \cdots & a_{2,n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{n,1} & a_{n,2} & \cdots & a_{n,j} & \cdots & a_{n,n} \end{pmatrix}$.

On note $A_{i,j}$ la matrice obtenue en supprimant la i-ème ligne et la j-ème colonne de A.

Le déterminant de A est le réel (développement selon la colonne j):

$$det(A) = \sum_{i=1}^{n} (-1)^{i+j} a_{i,j} det(A_{i,j}),$$

ou encore le réel (développement selon la ligne i):

$$det(A) = \sum_{j=1}^{n} (-1)^{i+j} a_{i,j} det(A_{i,j}),$$

Exemples: Soit $A = \begin{pmatrix} 1 & 2 & -1 \\ -2 & 3 & 1 \\ 0 & 4 & -1 \end{pmatrix}$. On a (en développant selon la première ligne)

$$det(A) = 1 \begin{vmatrix} 3 & 1 \\ 4 & -1 \end{vmatrix} - 2 \begin{vmatrix} -2 & 1 \\ 0 & -1 \end{vmatrix} - 1 \begin{vmatrix} -2 & 3 \\ 0 & 4 \end{vmatrix} = -3,$$

ou encore, en en développant selon la première colonne:

$$det(A) = 1 \begin{vmatrix} 3 & 1 \\ 4 & -1 \end{vmatrix} + 2 \begin{vmatrix} 2 & -1 \\ 4 & -1 \end{vmatrix} = -3.$$

Soit $B = \begin{pmatrix} 1 & 2 & 3 & 1 \\ -1 & 0 & 1 & 2 \\ 2 & 0 & -1 & 0 \\ 3 & 0 & -3 & 1 \end{pmatrix}$. On a (en développant selon la deuxième colonne)

$$det(B) = -2 \begin{vmatrix} -1 & 1 & 2 \\ 2 & -1 & 0 \\ 3 & -3 & 1 \end{vmatrix}$$

et en développant selon la dernière colonne

$$det(B) = -2 \left(2 \begin{vmatrix} 2 & -1 \\ 3 & -3 \end{vmatrix} + 1 \begin{vmatrix} -1 & 1 \\ 2 & -1 \end{vmatrix} \right) = 14.$$

Propriétés des déterminants:

1. $det(A^T) = det(A)$.
   Preuve: en supprimant la i-ème ligne et la j-ème colonne de A^T, on supprime la j-ème ligne et la i-ème colonne de A et on transpose, autrement dit $(A^T)_{i,j}$ est la matrice $(A_{j,i})^T$. En développant selon la i-ème ligne,

   $$det(A^T) = \sum_{i=1}^{n} (-1)^{i+j} (A^T)_{i,j} det((A^T)_{i,j}) = \sum_{i=1}^{n} (-1)^{i+j} (A)_{j,i} det((A_{j,i})^T),$$

   et par récurrence on obtient que c'est égal à $det(A)$ développé selon la j-ème colonne.

2. Le déterminant d'une matrice diagonale est égal au produit des valeurs diagonales: $det\,(Diag(\lambda_1, \lambda_2, \ldots, \lambda_{n-1}\lambda_n)) = det \begin{pmatrix} \lambda_1 & 0 & \ldots & \ldots & 0 \\ 0 & \lambda_2 & 0 & \ldots & 0 \\ 0 & \ldots & \ldots & \ldots & 0 \\ 0 & \ldots & \ldots & \lambda_{n-1} & 0 \\ 0 & 0 & \ldots & 0 & \lambda_n \end{pmatrix} = \lambda_1 \lambda_2 \cdots \lambda_{n-1} \lambda_n$.
   Preuve: par récurrence en développant selon la première colonne.
   En particulier, $det(I_n) = 1$;

3. Le déterminant d'une matrice triangulaire est égal au produit des valeurs diagonales: $det \begin{pmatrix} A_{1,1} & A_{1,2} & \ldots & \ldots & A_{1,n} \\ 0 & A_{2,2} & A_{2,3} & \ldots & A_{2,n} \\ 0 & \ldots & \ldots & \ldots & \ldots \\ 0 & \ldots & 0 & A_{n-1,n-1} & A_{n-1,n} \\ 0 & 0 & \ldots & 0 & A_{n,n} \end{pmatrix} = A_{1,1} A_{2,2} \cdots A_{n-1,n-1} A_{n,n}$.
   Preuve: par récurrence en développant selon la première colonne.
   Pour les matrices diagonales inférieures, on applique le point 1.

4. Si A et B sont deux matrices $n \times n$, $det(AB) = det(BA) = det(A)det(B)$ (admis).

5. $det(A^{-1}) = 1/det(A)$.
   Preuve: $det(A^{-1}A) = det(I_n) = 1$ et $det(A^{-1}A) = det(A^{-1})det(A)$ par le point précédent.

Proposition 26 *Une matrice $n \times n$ est inversible si et seulement si son déterminant est non nul.*

Inverse des matrices 1×1: une matrice 1×1 est un réel, son inverse est l'inverse de ce réel.
Inverse des matrices 2×2: une matrice $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ a pour inverse $A^{-1} = \frac{1}{ad-bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$.
Inverse des matrices $n \times n$:
On appelle cofacteur (i,j) de la matrice A le réel $Cof(A)_{i,j} = (-1)^{i+j} det(A_{i,j})$, où $A_{i,j}$ est la matrice obtenue en supprimant la i-ème ligne et la j-ème colonne de A. On appelle matrice des cofacteurs la matrice $Cof(A)$. On a alors

$$A^{-1} = \frac{1}{det(A)} \left(Cof(A)\right)^T, \quad \left(A^{-1}\right)_{i,j} = \frac{(-1)^{i+j} det(A_{j,i})}{det(A)}.$$

Exemple:
Soit $A = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \\ 0 & 1 & -1 \end{pmatrix}$. On a $det(A) = 6$, $Cof(A) = \begin{pmatrix} -4 & 2 & 2 \\ 5 & -1 & -1 \\ -7 & 5 & -1 \end{pmatrix}$, donc $A^{-1} = \frac{1}{6} \begin{pmatrix} -4 & 5 & -7 \\ 2 & -1 & 5 \\ 2 & -1 & -1 \end{pmatrix}$.

8 Valeurs propres et vecteurs propres.

8.1 Généralités

Définition 23 *Soit A une matrice $n \times n$. On dit que λ est valeur propre de A si il existe un vecteur de $\mathbb{R}^n$ U non nul tel que $AU = \lambda U$. On dit alors que U est vecteur propre de A associé à la valeur propre λ.*

Exemples:
- Pour tout vecteur U de $\mathbb{R}^n$, $I_n U = U$, donc tout vecteur de $\mathbb{R}^n$ est vecteur propre de I_n associé à la valeur propre 1.
- Si P est la projection orthogonale sur le sous-espace vectoriel E de $\mathbb{R}^n$, on a pour tout U vecteur de $E^\perp$, $P(U) = 0$ et pour tout U vecteur de E, $P(U) = U$. Donc si A est la matrice qui représente P dans la base canonique, tous les vecteurs de $E^\perp$ sont vecteurs propres de A associés à

la valeur propre 0, et tous les vecteurs de E sont vecteurs propres de A associés à la valeur propre 1.

- Si $D = Diag(D_1, \dots, D_n)$ est une matrice diagonale, chaque vecteur canonique E_i est vecteur propre de D associé à la valeur propre D_i.

Remarque: on peut parler de valeur propre et vecteur propre associée à une application linéaire. Alors si A représente l'application linéaire, ses valeurs propres et vecteurs propres associés sont les mêmes pourvu d'écrire les vecteurs dans les bases coorespondant à la représentation.

Proposition 27 *Si λ est valeur propre de A, l'ensemble E_λ des vecteurs propres de A associés à λ est un sous-espace vectoriel de $\mathbb{R}^n$.*

Preuve:
Soit λ valeur propre de A. Si $U \in E_\lambda$, pour tout réel non nul α, $A(\alpha U) = \alpha AU = \alpha\lambda U$ car U est vecteur propre de A associé à la valeur propre λ, soit $A(\alpha U) = \lambda(\alpha U)$ et $\alpha U \in E_\lambda$.
Si de plus $V \in E_\lambda$, $A(U+V) = AU + AV = \lambda U + \lambda V = \lambda(U+V)$ et $U + V \in E_\lambda$.

Proposition 28 *Les valeurs propres de A sont les solutions en λ de l'équation*

$$det(A - \lambda I_n) = 0,$$

et les vecteurs propres correspondants sont les éléments du noyau de $A - \lambda I_n$.

Preuve:
Si λ est valeur propre, alors il existe un vecteur non nul U tel que $(A-\lambda I_n)U = 0$, donc $A - \lambda I_n$ n'est pas inversible. Réciproquement, si $det(A-\lambda I_n) = 0$, $A - \lambda I_n$ n'est pas inversible et il existe un vecteur non nul U tel que $(A - \lambda I_n)U = 0$.
Si λ est valeur propre, l'ensemble des vecteurs propres associés sont tous les vecteurs U vérifiant $(A - \lambda I_n)U = 0$, donc $E_\lambda = Ker(A - \lambda I_n)$.

Remarque: 0 est valeur propre de A si et seulement si A n'est pas inversible.

Exemples:

- Soit $A = \begin{pmatrix} 1 & -3 \\ -2 & 2 \end{pmatrix}$. $A - \lambda I_2 = \begin{pmatrix} 1-\lambda & -3 \\ -2 & 2-\lambda \end{pmatrix}$, $det(A - \lambda I_2) = (1-\lambda)(2-\lambda) - 6 = \lambda^2 - 3\lambda - 4$. L'équation $\lambda^2 - 3\lambda - 4 = 0$ a pour solutions $\lambda = -1$ et $\lambda = 4$.
  Les vecteurs propres associés à la valeur propre -1 sont ceux qui vérifient $\begin{pmatrix} 2 & -3 \\ -2 & 3 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = 0$, soit $2x - 3y = 0$, E_{-1} est la droite engendrée par $\begin{pmatrix} 3 \\ 2 \end{pmatrix}$.

Les vecteurs propres associés à la valeur propre 4 sont ceux qui vérifient $\begin{pmatrix} -3 & -3 \\ -2 & -2 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = 0$, soit $-3x - 3y = 0$, E_4 est la droite engendrée par $\begin{pmatrix} 1 \\ -1 \end{pmatrix}$.

- Soit $A = \begin{pmatrix} 1 & -3 \\ 0 & 1 \end{pmatrix}$. $A - \lambda I_2 = \begin{pmatrix} 1-\lambda & -3 \\ 0 & 1-\lambda \end{pmatrix}$, $det(A - \lambda I_2) = (1-\lambda)^2$. L'équation $(1-\lambda)^2 = 0$ a pour seule solution $\lambda = 1$.
  Les vecteurs propres associés à la valeur propre 1 sont ceux qui vérifient $\begin{pmatrix} 0 & -3 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = 0$, soit $-3y = 0$, E_1 est la droite engendrée par $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$.
- Soit $A = \begin{pmatrix} 1 & 2 \\ -1 & 2 \end{pmatrix}$. $A - \lambda I_2 = \begin{pmatrix} 1-\lambda & -3 \\ -2 & 2-\lambda \end{pmatrix}$, $det(A - \lambda I_2) = (1-\lambda)(2-\lambda) + 2 = \lambda^2 - 3\lambda + 4$. L'équation $\lambda^2 - 3\lambda + 4 = 0$ n'a pas de solution et A n'a pas de valeurs propres.

8.2 Diagonalisation.

Définition 24 *On dit que A est diagonalisable si elle possède n vecteurs propres $U_1, \dots, U_n$ linéairement indṕendants.*

Remarque:
Si A possède n vecteurs propre $U_1, \dots U_n$ linéairement indépendants associés aux valeurs propres $\lambda_1, \dots \lambda_n$, si A représente une application linéaire f dans la base canonique, on a que pour $i = 1, \dots, n$, $f(U_i) = \lambda_i U_i$. Maintenant $U = (U_1, \dots, U_n)$ est une base de $\mathbb{R}^n$, et la matrice qui représente f dans la base U est $Diag(\lambda_1, \dots, \lambda_n)$.
Autrement dit, l'intérêt d'avoir une base de vecteurs propres est que par un changement de base on se ramène à une matrice diagonale.

Exemples et contre-exemples :
Dans les exemples précédents, seul la première matrice est diagonalisable.

Proposition 29 *(Essentiellement admise sauf l'aspect BON).*
Une matrice symétrique est diagonalisable dans une base orthonormée, c'est à dire qu'elle admet n vecteurs propres formant une base orthonormée de $\mathbb{R}^n$.

Preuve:
Soit A une matrice symétrique. On admettra qu'elle possède une base de vecteurs propres. Si U et V sont des vecteurs propres de A associés à des valeurs propres distinctes λ et μ, alors $U^T AV = \mu U^T V = (A^T U)^T V = (AU)^T V = \lambda U^T V$, donc $(\mu - \lambda)\langle U, V\rangle = 0$, donc $\langle U, V\rangle = 0$. Donc des vecteurs propres de A associés à des valeurs propres distinctes sont orthogonaux. On peut alors les

normaliser.
Si maintenant on regroupe tous les vecteurs propres linéairement indépendants associés à la même valeur propre, ils forment une base de l'espace des vecteurs propres associés, et on peut orthonormaliser cette base en restant dans l'espace.

8.3 Vecteurs gaussiens 3.

Théorème 6 *Soit X de loi $\mathcal{N}_p(0, \Sigma)$. Il existe $k \leq p$ et une application linéaire b de $\mathbb{R}^p$ dans $\mathbb{R}^k$ telle que $b(X)$ soit un k-échantillon de $\mathcal{N}(0,1)$.*

Preuve:
Σ est une matrice symétrique, elle est donc diagonalisable dans une base orthonormée. Soit $U_1, \ldots, U_p$ cette base orthonormée. Les valeurs propres associées sont λ_i, et comme Σ est semi-définie positive, on a $U_i^T \Sigma U_i \geq 0$, soit $\lambda_i U_i^T U_i \geq 0$, soit donc $\lambda_i \geq 0$: les valeurs propres sont positives ou nulles.
Si il y en a k non nulles, quitte à renuméroter on peut dire que ce sont les k premières. Posons pour $i = 1, \ldots, k$,

$$V_i = \frac{1}{\sqrt{\lambda_i}} U_i,$$

B la matrice $k \times p$ qui a pour lignes les V_i, et b l'application linéaire associée dans les bases canoniques de $\mathbb{R}^k$ et $\mathbb{R}^n$.
Alors $BX \sim \mathcal{N}_k(0, B\Sigma B^T)$. Mais ΣB^T a pour colonnes les vecteurs $\sqrt{\lambda_i} U_i$, et donc le terme i,j de $B\Sigma B^T$ est le produit scalaire de V_i et $\sqrt{\lambda_j} U_j$, soit 0 si $i \neq j$ et 1 si $i = j$, donc $BX \sim \mathcal{N}_k(0, I_k)$.

On peut alors montrer que si Σ est inversible, la densité de la loi $\mathcal{N}_p(0, \Sigma)$ est

$$\frac{1}{(2\pi det(\Sigma))^{n/2}} \exp\left(-\frac{1}{2} X^T \Sigma^{-1} X\right) \quad \text{avec } X = \begin{pmatrix} x_1 \\ \vdots \\ x_p \end{pmatrix}.$$

Diagonaliser une matrice de covariance, c'est chercher les combinaisons linéaires des variables qui sont non corrélées (indépendantes dans le cas gaussien).

En effet, si on pose

$$u_i(X) = \langle U_i, X \rangle$$

on a

$$Cov\left(u_i(X); u_j(X)\right) = U_i^T Var(X) U_j = U_i^T \left(\lambda_j U_j\right) = 0$$

si $i \neq j$, et

$$Var\left(u_i(X)\right) = U_i^T Var(X) U_i = U_i^T \left(\lambda_i U_i\right) = \lambda_i.$$

Si l'on pose

$$v_i(X) = \frac{u_i(X)}{\sqrt{Var(u_i(X))}} = \langle V_i, X \rangle,$$

on a $Var(v_i(X)) = 1$ et $Cov(u_i(X); u_j(X)) = 0$ si $i \neq j$: les variables $v_i(X)$, combinaisons linéaires des coordonnées de X, sont non corrélées et de variance 1.

9 Analyse en Composantes Principales (ACP) et lien avec le modèle linéaire.

On a un tableau X de n lignes et p colonnes, coorespondant aux mesures de p variables sur n individus. Autrement dit les lignes $X_i = (X_{i,1}, \ldots, X_{i,p})$ pour $i = 1, \ldots, n$ représentent ces mesures pour l'individu i, à valeurs dans $\mathbb{R}^p$. Les colonnes X^j représentent les n mesures de la variable j sur tous les individus.
L'ensemble des points X_i de $\mathbb{R}^p$ est le nuage des points individus, et l'ensemble des points X^j de $\mathbb{R}^n$ est le nuage des points variables.
On souhaite réduire la dimension p (pour le nuage des points individus) ou la dimension n (pour le nuage des points variables) tout en conservant le maximum d'information, et comprendre le lien entre les individus et les variables.
Remarque: on fait de la géométrie affine. Un vecteur qui relie deux points a pour coordonnées les différences des coordonnées d'un point. Un point peut être vu comme l'origine plus un vecteur. Un point plus un vecteur donne un point. Pour définir un espace affine, on a besoin d'un point et d'un espace vectoriel.

9.1 Analyse du nuage des individus.

9.1.1 Cas p=2: droite des moindres rectangles.

On considère le cas $p = 2$, et l'on a donc n points $X_i = (X_{,1}, X_{i,2})$ dans le plan.
Le tableau de données est

$$X = \begin{pmatrix} X_{1,1} & X_{1,2} \\ X_{2,1} & X_{2,2} \\ \vdots & \vdots \\ X_{n,1} & X_{n,2} \end{pmatrix}.$$

On cherche la droite qui s'ajuste le mieux au nuage de points, c'est à dire qui rend le plus faible possible la somme des carrés des distances des points à la droite.
Pour définir une droite $\mathcal{D}$ on a besoin d'un point et d'une direction. Soit M le point et U la direction (vecteur de norme 1). La projection d'un point Z sur cette droite $\mathcal{D}$ est le point

$$M + P_{\mathcal{D}}(Z - M) = M + \langle Z - M, U \rangle U.$$

La distance du point à la droite est la norme du vecteur qui relie Z à sa projection, donc la norme du vecteur $M + P_{\mathcal{D}}(Z - M) - Z$, soit du vecteur

$P_{\mathcal{D}}(Z-M)-(Z-M)$. Maintenant

$$\begin{aligned} \| P_{\mathcal{D}}(Z-M)-(Z-M) \|^2 &= \| P_{\mathcal{D}}(Z-M) \|^2 + \| (Z-M) \|^2 - 2\langle P_{\mathcal{D}}(Z-M), Z-M\rangle \\ &= \| (Z-M) \|^2 - (\langle Z-M, U\rangle)^2 , \end{aligned}$$

donc la somme des carrés des distances des points à la droite, ce que l'on veut minimiser, est

$$\sum_{i=1}^{n} \| (X_i - M) \|^2 - \sum_{i=1}^{n} (\langle X_i - M, U\rangle)^2 .$$

On veut minimiser en le point M et la direction U; on minimise d'abord en M, et l'on obtient (pour le voir écrire en fonction des deux coordonnées, et l'on minimise coordonnée par coordonnée):

$$M = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i ,$$

"centre de gravité" du nuage de points si on leur donne tous la même masse. On minimise ensuite en U, ce qui revient à maximiser en U:

$$\sum_{i=1}^{n} (\langle X_i - \bar{X}, U\rangle)^2 = \sum_{i=1}^{n} U^T (\langle X_i - \bar{X})^T (\langle X_i - \bar{X})U$$

puisque les vecteurs X_i sont notés en ligne, il s'agit donc de maximiser $U^T SU$, avec $S == \sum_{i=1}^{n} (\langle X_i - \bar{X})^T (\langle X_i - \bar{X})$ matrice 2×2 sous la contrainte que U soit de norme 1. On peut remarquer que S/n est la matrice de variance empirique des deux variables X^1 et X^2. IL faut noter aussi que $U^T SU$ est la somme des carrés des longueurs des projections des $X_i - \bar{X}$ sur la droite; autrement dit la droite qui s'ajuste le mieux au nuage de points est celle qui rend l'inertie du nuage projeté la plus grande. Ou autrement dit: qui sépare le mieux les points projetés. Maintenant, S est symétrique donc diagonalisable dans une base orthonormée. Ses valeurs propres sont positives car pour tout vecteur U, $U^T SU \geq 0$. Soit U_1, U_2 la base de vecteurs propres et $\lambda_1 \geq \lambda_2$ les valeurs propres. On écrit $U = \alpha_1 U_1 + \alpha_2 U_2$ avec $\alpha_1^2 + \alpha_2^2 = 1$, et

$$\begin{aligned} U^T SU &= u^T (\alpha_1 \lambda_1 U_1 + \alpha_2 \lambda_2 U_2 \\ &= \lambda_1 \alpha_1^2 + \lambda_2 \alpha_2^2 . \end{aligned}$$

Cette quantité est maximum en prenant $\alpha_1 = 1$ et $\alpha_2 = 0$.
Donc **la droite qui s'ajuste le mieux au nuage de points est la droite passant par $\bar{X}$ et de direction le vecteur propre de S de plus grande valeur propre.**
On peut noter aussi que si $\lambda_2 = 0$, cela signifie que tous les points X_i sont alignés sur une droite qui est celle des moindres rectangles. En général, la représentation sur la droite est d'autant plus fidèle au nuage que λ_1 est grande.

Interprétation en modèle linéaire:
On suppose que les X_i sont les observations bruitées de vecteurs m_i, de deux coordonnées représentant chacune l'effet moyen de la variable sur l'individu, soit

$$X_i = m_i + \epsilon_i,$$

les $\epsilon_{i,j}$, $i = 1, \dots, n$, $j = 1, 2$, étant indépendantes de même variance σ^2.

On pose $m_i = \mu + a_i$, μ et a_i vecteurs de $\mathbb{R}^p$; μ est la partie qui ne varie pas selon les individus, a_i est la partie due à l'individu; mais on suppose que les a_i sont tous sur une même droite D, ce qui signifie que les deux variables ont un effet proportionnel, l'effet variant selon l'individu mais pas le coefficient de proportionnalité.
On va estimer μ, et U, vecteur directeur de D de norme 1, ainsi que les a_i par moindres carrés (ce qui équivaut au maximum de vraisemblance dans le cas gaussien).
En appliquant les méthodes du modèle linéaire (exercice: le faire; c'est à dire identifier E etc...), lorsque U est fixé, on obtient comme estimateurs (dans $\mathbb{R}^2$):

$$\widehat{\mu} = \bar{X}, \quad \widehat{a}_i = \langle X_i - \bar{X}, U \rangle U.$$

On a alors que les moindres carrés à minimiser pour estimer U sont:

$$\begin{aligned} \sum_{i=1}^{n} \| X_i - \widehat{\mu} - \widehat{a}_i \|^2 &= \sum_{i=1}^{n} \| X_i - \bar{X} - \langle X_i - \bar{X}, U \rangle U \|^2 \\ &= \sum_{i=1}^{n} \| X_i - \bar{X} \|^2 - \sum_{i=1}^{n} \left(\langle X_i - \bar{X}, U \rangle \right)^2 . \end{aligned}$$

Cela revient à maximiser $U^T SU$, S, on est ramené au problème précédent..
Autrement dit, chercher la droite qui s'ajuste le mieux au nuage de points c'est la même chose qu'estimer par moindres carrés la droite qui contient toutes les moyennes des effets des variables dans un modèle linéaire.

Remarque: ce n'est par contre pas la même chose que faire une régression linéaire d'une variables sur l'autre. Dans une régression linéaire la droite des “moindres carrés” n'est pas la droite des “moindres rectangles”.

9.1.2 Cas général.

On a donc n points dans un espace de dimension p. On cherche des espaces de plus faible dimension q qui s'ajustent le mieux au nuage de points.
On gńéralise le résultat précédent de la façon suivante:

Théorème 7 *Le sous-espace affine de dimension q qui s'ajuste le mieux au nuage de points est celui qui passe par $\bar{X}$ et dont les vecteurs directeurs sont les $U_1, \dots, U_q$, orthogonaux et normés, vecteurs propres de S, associés aux q plus* *grandes valeurs propres de S, où*

$$S = \sum_{i=1}^{n} \left(X_i - \bar{X}\right)\left(X_i - \bar{X}\right)^T = \sum_{i=1}^{n} X_i X_i^T - n\bar{X}\bar{X}^T .$$

Ici, $\bar{X}$ est le vecteur de $\mathbb{R}^p$ moyenne des individus pour chaque variable:

$$\bar{X} = \left(\bar{X}_1, \ldots, \bar{X}_p\right), \ \bar{X}_j = \frac{1}{n}\sum_{i=1}^{n} X_{i,j} .$$

On remarque que S/n est la matrice de variance empirique des p variables X^1, …, X^p calculée sur les n individus.

Eléments de preuve:
On cherche l'espace affine $\mathcal{V}$ en le caractérisant par un point M et une base orthonormée $B_1, \ldots B_q$. La projection d'un point Z sur $\mathcal{V}$ est le point

$$M + P_{\mathcal{V}}\left(Z - M\right) = M + \sum_{i=1}^{q} \langle Z - M, B_i\rangle B_i .$$

La distance du point à $\mathcal{V}$ est la norme du vecteur qui relie Z à sa projection, donc la norme du vecteur $M + P_{\mathcal{V}}(Z-M) - Z$, soit du vecteur $P_{\mathcal{V}}(Z-M) - (Z-M)$. Maintenant

$$\begin{aligned} \| P_{\mathcal{V}}\left(Z - M\right) - \left(Z - M\right) \|^2 &= \| P_{\mathcal{V}}\left(Z - M\right) \|^2 + \| \left(Z - M\right) \|^2 - 2\langle P_{\mathcal{V}}\left(Z - M\right), Z - M\rangle \\ &= \| \left(Z - M\right) \|^2 - \sum_{i=1}^{q} \left(\langle Z - M, B_i\rangle\right)^2 , \end{aligned}$$

donc la somme des carrés des distances des points à l'espace, ce que l'on veut minimiser, est

$$\sum_{i=1}^{n} \| \left(X_i - M\right) \|^2 - \sum_{i=1}^{n}\sum_{j=1}^{q} \left(\langle X_i - M, B_j\rangle\right)^2 .$$

On veut minimiser en le point M et les vecteurs orthogonaux et normés direction B_j; on minimise d'abord en M, et l'on obtient (pour le voir écrire en fonction des coordonnées, et l'on minimise coordonnée par coordonnée):

$$M = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i ,$$

"centre de gravité" du nuage de points si on leur donne tous la même masse. On minimise ensuite en $B_1, \ldots, B_q$, ce qui revient à maximiser en $B_1, \ldots, B_q$:

$$\sum_{i=1}^{n}\sum_{j=1}^{q} \left(\langle X_i - M, B_j\rangle\right)^2 = \sum_{j=1}^{q} B_j^T \sum_{i=1}^{n} ((\langle X_i - \bar{X})((\langle X_i - \bar{X})^T B_j ,$$

il s'agit donc de maximiser $\sum_{j=1}^q B_j^T SB_j$. S est symétrique, donc diagonalisable dans une base orthonormée, et ses valeurs propres sont ≥ 0. Soit $U_1, \ldots, U_p$ la base orthonormée de vecteurs propres de S, et soit $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$ les valeurs propres associées. En écrivant les B_i sur la base des U_j:

$$B_i = \sum_{j=1}^p b_{i,j} U_j$$

on obtient

$$\begin{aligned}
\sum_{j=1}^q B_j S B_j^T &= \sum_{j=1}^q \left(\sum_{k=1}^p b_{j,k} U_k \right) S \left(\sum_{l=1}^p b_{j,l} U_l^T \right) \\
&= \sum_{j=1}^q \sum_{k=1}^p \sum_{l=1}^p b_{j,k} b_{j,l} \lambda_l U_k U_l^T \\
&= \sum_{j=1}^q \sum_{k=1}^p b_{j,k}^2 \lambda_k
\end{aligned}$$

car la base $U_1, \ldots, U_p$ est orthonormée.
Pour $j = 1$,

$$\sum_{k=1}^p b_{1,k}^2 \lambda_k \leq \lambda_1 \sum_{k=1}^p b_{1,k}^2 = \lambda_1,$$

et cette valeur est atteinte en prenant $B_1 = U_1$. Si on cherche ensuite B_2 orthogonal à U_1 (donc $b_{2,1} = 0$), on a:

$$\sum_{k=2}^p b_{2,k}^2 \lambda_k \leq \lambda_2 \sum_{k=2}^p b_{2,k}^2 = \lambda_2,$$

et cette valeur est atteinte en prenant $B_2 = U_2$.
Si on continue ainsi on obtient $B_1 = U_1$, ..., $B_q = U_q$. On a alors

$$\sum_{j=1}^q U_j S U_j^T = \sum_{j=1}^q \lambda_j.$$

Noter que le fait que cela donne le maximisant n'est pas prouvé totalement, il faut voir que cette procédure récursive est optimale. Elle l'est...
Remarque importante: On a jusqu'à présent calculé les distances en faisant la somme des carrés des différences de coordonnées. Mais si les variables sont d'échelle très diverse, cela donne un poids plus important aux variables de grande échelle. On peut souhaiter le corriger en centrant et normalisant les individus variables par variable, c'est à dire avec des individus Y_i de coordonnées

$$Y_{i,j} = \frac{X_{i,j} - \bar{X}_j}{\sqrt{n} r_j}, \ r_j^2 = \frac{1}{n} \sum_{i=1}^n \left(X_{i,j} - \bar{X}_j \right)^2.$$

Cela revient ensuite à diagonaliser, au lieu de S, la matrice C des corrélations empiriques des variables,

$$C_{j,k} = \frac{1}{n}\sum_{i=1}^{n} \frac{(X_{i,j} - \bar{X}_j)(X_{i,k} - \bar{X}_k)}{r_j r_k}.$$

Les vecteurs propres de S (ou C) sont appelés les **axes factoriels.** L'intertie totale du facteur j est λ_j.
Sur l'axe factoriel U_j, la coordonnée de la projection de X_i est $\psi_j = \langle X_i, U_j \rangle$. On voit donc que si l'on range en n lignes les individus (vecteurs lignes de dimension p), et que X est la matrice du tableau de données centrées et normalisées (c'est à dire que les variables X^1,...,X^p sont de moyenne empirique nulle (sur les individus) et de variance empirique 1 (sur les individus)), alors $S = X^T X$, et le vecteur de $\mathbb{R}^n$

$$\Psi_j = XU_j$$

est un facteur combinaison linéaire des variables initiales, dont les coordonnées sont les coordonnées de chaque individu sur le j-ème axe factoriel.
Interprétation: modèle linéaire. Si on suppose que les X_i sont les observations bruitées de vecteurs m_i, soit

$$X_i = m_i + \epsilon_i,$$

les $\epsilon_{i,j}$, $i = 1, \ldots, n$, $j = 1, \ldots, p$, étant indépendantes de même variance σ^2.

On pose $m_i = \mu + a_i$, μ et a_i vecteurs de $\mathbb{R}^p$; μ est la partie qui ne varie pas selon les individus, a_i est la partie due à l'individu; mais on suppose que les a_i sont tous dans le même sous- espace vectoriel $\mathcal{V}$ de $\mathbb{R}^p$ de dimension q.

Si on estime μ, $\mathcal{V}$ et les a_i par moindres carrés (ce qui équivaut au maximum de vraisemblance dans le cas gaussien), on s'aperçoit comme pour $p = 2$ que l'estimateur du sous-espace vectoriel de dimension q est le sous-espace engendré par les q directions propres associées aux q plus grandes valeurs propres de S.
Remarque: le fait de supposer que tous les $\epsilon_{i,j}$ ont même variance montre l'importance du choix de la métrique, et justifie encore de faire une analyse sur la matrice des corrélations empiriques.

9.2 Analyse du nuage des points-variables.

On considère maintenant les colonnes de X qui sont des vecteurs de $\mathbb{R}^n$ qui représentent les variables avec leurs coordonnées sur chaque individu. Si on travaille avec le tableau centré et normalisé, on voit que chaque vecteur variable est de norme 1, et que le produit scalaire des deux vecteurs-variable colonne j et k est $C_{j,k}$. On a alors que la distance euclidienne entre ces deux vecteurs-variable est

$$d^2(j,k) = 2\,(1 - C_{j,k})\,.$$

Les proximités entre points-variables s'interprètent donc en termes de corrélations:

- Deux variables fortement corrélées sont très proches l'une de l'autre ou au contraire les plus éloignées possibles selon que la relation linéaire qui les relie est de signe positif ou négatif.
- Deux variables non corrélées sont à distance moyenne.

Tous les points-variables sont sur la sphère unité, appelée la sphère des corrélations.
Si maintenant on veut faire l'ajustement du nuage des points-variables comme pour les individus, on va chercher des espaces passant par l'origine, et les plans d'ajustement couperont la sphère suivant des cercles de rayon 1, appelés "cercles des corrélations".
Pour chercher les espaces d'ajustement passant par l'origine, en suivant l'analyse précédente on voit que cela remplace X par X^T (ici on ne centre pas les variables), et qu'il s'agit de diagonaliser la matrice $n \times n$ XX^T.
On peut vérifier que les vecteurs propres (de norme 1) de XX^T sont

$$V_j = \frac{1}{\sqrt{\lambda_j}} X U_j = \frac{1}{\sqrt{\lambda_j}} \Psi_j$$

pour la valeur propre λ_j. Les autres valeurs propres sont nulles car XX^T et $X^T X$ ont même rang. Ces variables artificielles sont appelées **composantes principales.**
Les coordonnées dans $\mathbb{R}^n$ de la j-ième variable sont

$$\phi_j = X^T V_j = \frac{1}{\sqrt{\lambda_j}} X^T X U_j = \sqrt{\lambda_j} U_j.$$

On a aussi

$$\phi_j = \frac{1}{\sqrt{\lambda_j}} X^T \psi_j$$

et la coordonnée factorielles d'un point-variable sur l'axe V_j est le coefficient de corrélation de cette variables avec le facteur ψ_j.

9.3 Représentations et interprétation.

Les deux nuages ne sont pas dans le même espace, on ne peut les représenter simultanément.
Les proximités entre variables s'interprètent en termes de corrélations. Dans le cercle des corrélations, les variables fortement corrélées avec un axe vont contribuer à la définition de cet axe. On s'intéresse aux variables proches du cercle des corrélations, et l'on interprétera les composantes principales en fonction des regroupements de certaines de ces variables et de l'opposition avec d'autres.
Dans le nuage des individus, les proximités entre individus s'interprètent en termes de similitude de comportement vis à vis des variables.
Si l'on considère des "directions de variables", on représente souvent simultanément les individus et des vecteurs de variables. Dans $\mathbb{R}^p$, on a les axes factoriels U_j, et les anciens axes unitaires E_j direction de la variable j. La projection deE_k

sur U_j a pour coordonnée $U_{j,k}$. La représentation simultanée des vecteurs de variables avec les individus est l'écrasement du repère orthonormé d'origine sur le plan factoriel. La direction d'une variable définit des zones pour les individus: d'un côté ceux qui prennent des valeurs fortes pour cette variable, et à l'opposé ceux qui prennent des valeurs faibles. On s'intéressera à l'éloignement des individus dans la direction de la variable. A l'intersection des axes se trouvent les valeurs moyennes de toutes les variables.
Rappelons que le point-variable X^k a pour coordonnée $\phi_{j,k} = \sqrt{\lambda_j} U_{j,k}$. Les points-variables et les directions-variables diffèrent par une dilatation sur chaque axe. Si dans les plan les deux valeurs propres sont voisines, alors*la représentation des vecteurs-variables ressemble à celle des points-variables.

10 Analyse Factorielle des Correspondances (AFC) et lien avec les modèles discrets.

On étudie maintenant des variables nominales, c'est à dire que l'on a des individus sur lesquels on observe des facteurs qui ne prennent pas des valeurs continues mais des valeurs discrètes, par exemple qualitatives (comme la couleur des yeux). Dans l'AFC, on considère deux facteurs et l'on souhaite étudier les relations entre les deux facteurs.
Les données sont donc pour chaque individu i la valeur $X_{i,1}$ du premier facteur et la valeur $X_{i,2}$ du deuxième facteur. Si le premier facteur a p niveaux et le deuxième facteur m niveaux, on peut considérer que les valeurs de $X_{i,1}$ vont de 1 à p et les valeurs de $X_{i,2}$ vont de 1 à m. A partir de ces données, on peut construire un tableau de contingence N avec p lignes et m colonnes, où si l'on a n individus

$$N_{k,l} = \sum_{i=1}^{n} \mathbf{1}_{X_{i,1}=k \text{ et } X_{i,2}=l},$$

c'est à dire $N_{i,j}$ est le nombre de fois où le premier facteur est k et le deuxième est l.
On note $p_{k,l}$ la probabilité d'observer simultanément le niveau k du premier facteur et le niveau l du deuxième facteur, autrement dit si l'on suppose que les $(X_{i,1}, X_{i,1})$ sont indépendantes et de même loi on a

$$p_{k,l} = P\left(X_{i,1} = k \text{ et } X_{i,2} = l\right).$$

Soit P la matrice $p \times m$ de terme général $P = (p_{k,l})$. On s'intéresse aux liens éventuels entre des niveaux des deux facteurs.
Par exemple si les deux facteurs sont indépendants, on a pour tout k et l

$$p_{k,l} = p_{k,+} p_{+,l},$$

avec

$$p_{k,+} = \sum_{l=1}^{m} p_{k,l}, \quad p_{+,l} = \sum_{k=1}^{p} p_{k,l}.$$

Si l'on note R le vecteur de $\mathbb{R}^p$ de coordonnées $p_{k,+}$ et S le vecteur de $\mathbb{R}^m$ de coordonnées $p_{+,l}$, cela s'écrit

$$P = RS^T$$

On a pour tout vecteur U de $\mathbb{R}^m$, $PU = RS^T U = \langle S, U \rangle R$, donc $PU = 0$ si et seulement si U est orthogonal à S, donc le noyau de P (comme application linéaire), donc le noyau de P est l'orthogonal de la droite engendrée par S, et est donc de dimension $m - 1$. Par le théorème des dimensions, la matrice P est de rang 1.
A l'opposé, si les niveaux des facteurs sont totalement liés,

- Si $p = m$, si le niveau du premier facteur détermine le niveau du deuxième facteur, en ordonnant convenablement les niveaux des facteurs, $P = I_p$, P est de rang maximal p.

– Plus généralement, si $p \leq m$, si on peut partitionner les niveaux du deuxième facteur de telle sorte que, si l'on est dans un niveau du premier facteur, alors le groupe de niveaux du deuxième facteur dans lequel on est est déterminé par la partition, alors en ordonnant convenablement les niveaux des facteurs P est constituée de p blocs lignes en diagonale, et la matrice P est de rang maximum égal à p.

Entre ces deux cas extrêmes, le rang de P indique comment on peut grouper des pseudos-niveaux de facteurs pour avoir des groupes liés. On cherche si P se groupe par blocs (ce n'est pas toujours possible).
Rechercher ces liens est le but de l'AFC; pour cela on teste le rang de P à l'aide d'un test du χ^2 construit à l'aide de N, on interprète quels sont les niveaux liés entre eux à l'aide de valeurs propres, vecteurs propres, et de représentations graphiques.

10.1 Tests du χ^2 pour le rang de P.

Commençons par le χ^2 d'ajustement. Soit Z une variable pouvant prendre q valeurs distinctes, et $Z_1, \ldots, Z_n$ n variables i.i.d. de même loi que Z. Soient Q et M les vecteurs de R^q de coordonnées

$$Q_k = P\left(Z = k\right), \; M_k = \sum_{i=1}^{n} \mathbf{1}_{Z_i = k}, \; k = 1, \ldots, q.$$

Notons que

$$E\left(\mathbf{1}_{Z_i=k}\right) = Q_k, \;\; Var\left(\mathbf{1}_{Z_i=k}\right) = Q_k(1 - Q_k)$$

et si $k \neq l$,

$$Cov\left(\mathbf{1}_{Z_i=k}, \mathbf{1}_{Z_i=l}\right) = E\left(\mathbf{1}_{Z_i=k}\mathbf{1}_{Z_i=l}\right) - Q_k Q_l = -Q_k Q_l,$$

autrement dit, M/n est la moyenne empirique des vecteurs aléatoires de $\mathbb{R}^q$ de coordonnées $\mathbf{1}_{Z_i=k}$, d'espérance Q et de matrice de variance Σ, avec $\Sigma_{k,l} = Q_k(1 - Q_k)$ si $k = l$ et $\Sigma_{k,l} = -Q_k Q_l$ si $k \neq l$. Pour approximer la loi de M quand n est grand ,on utilise le théorème de limite centrale multidimensionnel:

Théorème 8 *Si $Y_1, \ldots, Y_n$ sont des vecteurs aléatoires de $\mathbb{R}^q$, indépendants, de même loi, d'espérance m et de matrice de variance $\mathcal{V}$, alors quand n est grand, on a l'approximation*

$$\sqrt{n}\left(\frac{\sum_{i=1}^{n} Y_i}{n} - m\right) \approx \mathcal{N}_q(0, \mathcal{V}).$$

En appliquant cela aux Y_i vecteurs aléatoires de $\mathbb{R}^q$ de coordonnées $\mathbf{1}_{Z_i=k}$, on obtient

$$\sqrt{n}\left(\frac{1}{n}M - Q\right) \approx \mathcal{N}_q(0, \Sigma).$$

Soit U le vecteur de coordonnées $U_k = \sqrt{Q_k}$, et soit W celui de coordonnées

$$W_k = \sqrt{n}\left(\frac{\frac{1}{n}M_k - Q_k}{U_k}\right),$$

W est obtenu à partir de $\sqrt{n}\left(\frac{1}{n}M - Q\right)$ en divisant chaque coordonnée par le nombre U_k, donc par une application linéaire. Donc W est pour n grand approximativement gaussien, et l'on a

$$W \approx \mathcal{N}_q(0, \Gamma), \quad \Gamma = I_q - UU^T.$$

Or U est de norme 1, et l'on voit facilement que Γ est la matrice de projection orthogonale sur l'orthogonal de la droite engendrée par U, qui est un sous-espace vectoriel de $\mathbb{R}^q$ de dimension $q-1$. En effet, pour tout vecteur V orthogonal à U, $\Gamma V = V$, et tout vecteur colinéaire à U, $\Gamma V = V$.
Si l'on fait le changement de base orthonormée qui diagonalise Γ, et que l'on écrit tout dans cette nouvelle base, on a

$$\tilde{W} \approx \mathcal{N}(0, \tilde{\Gamma})), \ \tilde{\Gamma} = \begin{pmatrix} I_{q-1} & 0 \\ 0 & 0 \end{pmatrix}.$$

On peut aussi noter que W est orthogonal à U, car $\sum_{k=1}^q W_k U_k = 0$, donc $\tilde{W}$ a pour dernière coordonnée 0. Comme de plus $\| W \|^2 = \| \tilde{W} \|^2$ (le changement de base est orthonormé), on a par le théorème de Cochran que $\| W \|^2 \approx \chi^2(q-1)$, autrement dit que pour n grand,

$$\sum_{k=1}^{q} \frac{(M_k - nQ_k)^2}{nQ_k} \approx \chi^2(q-1),$$

qui est la variable de test du χ^2 d'ajustement.
Si l'on applique cela à la variable discrète qui est le croisement de deux facteurs, soit $q = p \cdot m$ et $Z_i = (X_{i,1}, X_{i,2})$, et au tableau de contingence N défini dans l'introduction, on a:

Proposition 30 *Pour n grand,*

$$\sum_{k=1}^{p} \sum_{l=1}^{m} \frac{(N_{k,l} - np_{k,l})^2}{np_{k,l}} \approx \chi^2(m \cdot p - 1).$$

Ce théorème permet de tester l'ajustement du tableau de contingence à une matrice (loi jointe des deux facteurs) donnée. Si l'on veut tester le rang de P, on va utiliser le résultat suivant (admis, non démontré):
Truc du chi-deux asymptotique:
Si E est un sous-espace vectoriel de dimension r de l'espace vectoriel des matrices $p \times m$, et si l'on estime P sous l'hypothèse $P \in E$ par $(\hat{p}_{k,l})$ qui minimise en $(p_{k,l})$

$$\sum_{k=1}^{p} \sum_{l=1}^{m} \frac{(N_{k,l} - np_{k,l})^2}{np_{k,l}} \approx \chi^2(m \cdot p - 1).$$

sous la contrainte $(p_{k,l}) \in P$, alors pour n grand,

$$\sum_{k=1}^{p} \sum_{l=1}^{m} \frac{(N_{k,l} - n\hat{p}_{k,l})^2}{n\hat{p}_{k,l}} \approx \chi^2(m \cdot p - 1 - r).$$

Si E est le sous-espace des matrices P de la forme $P = RS^T$, avec $\sum_{k=1}^{p} R_k = 1$ et $\sum_{l=1}^{m} S_l = 1$, E est de dimension $(p-1)+(m-1)$, et en appliquant le "Truc du chi-deux asymptotique", on obtient

Théorème 9 *Notons $N_{k,+} = \sum_{l=1}^{m} N_{k,l}$ et $N_{+,l} = \sum_{k=1}^{p} N_{k,l}$. Si les deux facteurs sont indépendants (c'est à dire si les variables $X_{i,1}$ et $X_{i,2}$ sont indépendantes), alors pour n grand*

$$\sum_{k=1}^{p}\sum_{l=1}^{m} \frac{\left(N_{k,l} - \frac{N_{k,+}N_{+,l}}{n}\right)^2}{\frac{N_{k,+}N_{+,l}}{n}} = n\sum_{j=2}^{q} \mu_j \approx \chi^2\left((m-1)\cdot(p-1)\right).$$

qui permet de tester l'indépendance des deux facteurs.
Dans ce théorème, les μ_j sont les valeurs propres de $(N^*)^T N^*$, voir la section suivante.
A l'aide de ces valeurs propres, en utilisant le "Truc du chi-deux asymptotique", de manière générale, on peut tester le rang de P par:

Théorème 10 *Si le rang de P vaut h, alors pour n grand*

$$n\sum_{j=h+1}^{q} \mu_j \approx \chi^2\left((m-h)\cdot(p-h)\right).$$

Si l'on retient que le rang de P est h, cela signifie qu'on peut décomposer pour chaque facteur l'espace de ses niveaux en h groupes de combinaisons linéaires des niveaux que l'on peut associer entre facteurs. On interpète ces groupes en regardant les vecteurs propres (représentant les groupes) selon les niveaux de facteur qui les composent de manière essentielle (ce n'est pas toujours possible, mais c'est ce qu'on cherche).

10.2 Composantes principales du tableau de contingence.

Comme pour l'ACP, on commence par "normaliser" les données pour comparer les niveaux des deux facteurs. Soit $N_{k,+} = \sum_{l=1}^{m} N_{k,l}$ et $N_{+,l} = \sum_{k=1}^{p} N_{k,l}$, et notons

$$D_p = Diag(N_{k,+}),\ D_m = Diag(N_{+,l}).$$

La matrice $D_p^{-1}N$ a pour terme général $N_{k,l}/N_{k,+}$: chaque ligne donne les "profils" des niveaux du deuxième facteur conditionnellement au niveau k du premier facteur, c'est à dire la fréquence de chaque niveau du deuxième facteur dans la population qui a le niveau k pour le premier facteur.
La matrice ND_m^{-1} a pour terme général $N_{k,l}/N_{+,l}$: chaque colonne donne les "profils" des niveaux du premier facteur conditionnellement au niveau l du deuxième facteur, c'est à dire la fréquence de chaque niveau du premier facteur dans la population qui a le niveau l pour le deuxième facteur.
On note

$$N^* = D_p^{-1/2} N D_m^{-1/2},$$

qui a pour terme général $N_{k,l}/\sqrt{N_{k,+}N_{+,l}}$.
On effectue l'ACP avec donnée N^*, donc les individus sont les modalités du premier facteur et les variables les modalités du deuxième facteur. Il s'agit donc de diagonaliser $(N^*)^T N^*$. Soient V_1, ..., V_m les vecteurs propres, associés aux valeurs propres $\mu_1 \geq \ldots \geq \mu_m$.
De même, on effectue l'ACP avec donnée $(N^*)^T$, donc les individus sont les modalités du deuxième facteur et les variables les modalités du premier facteur. Il s'agit donc de diagonaliser $N^*(N^*)^T$. Soient U_1, ..., U_p les vecteurs propres, associés aux valeurs propres $\mu_1 \geq \ldots \geq \mu_p$. On sait que ce sont les mêmes valeurs propres non nulles, mettons q valeurs propres non nulles ($q = rg(N^*)$), et que les V_j et U_i sont liés par

$$U_k = \frac{1}{\sqrt{\mu_k}} N^* V_k, \; V_k = \frac{1}{\sqrt{\mu_k}} (N^*)^T U_k.$$

Remarquons que si l'on note $\tilde{V}_k = D_m^{1/2} V_k$, on a

$$N^T D_p^{-1} N D_m^{-1} \tilde{V}_k = \mu_k \tilde{V}_k,$$

donc les ${}_m u_k$ sont aussi valeur propre de $A = N^T D_p^{-1} N D_m^{-1}$, de terme général

$$A_{k,l} = \frac{\sum_{j=1}^p N_{j,k} N_{j,l}}{N_{j,+} N_{+,l}}.$$

Chaque colonne de A est constituée de termes positifs de somme totale égale à 1, ce qui permet de déduire que

$$1 \geq \mu_1 \geq \mu_2 \geq \ldots \geq \mu_q > 0.$$

Par ailleurs, on vérifie facilement que $\mu_1 = 1$ avec V_1 et U_1 donnés par

$$(V_1)_l = \sqrt{N_{+,l}}, \; l = 1, \ldots, m, \; (U_1)_k = \sqrt{N_{k,+}}, \; k = 1, \ldots, p.$$

10.3 Interprétations et représentations graphiques.

En général on fait une représentation graphique dans les deux premiers axes factoriels ne correspondant pas à la valeur propre 1, c'est à dire dans (V_2, V_3) pour les lignes de N^*, et dans (U_2, U_3) pour les colonnes de N^*.
Traditionnellement, on effectue ces représentations dans les mêmes axes: dans le cas où, quitte à permuter des lignes et des colonnes, N peut se mettre dans une forme quasi diagonale par blocs, les niveaux des facteurs associés seront proches dans la représentation graphique.
L'inertie (la valeur propre) d'un axe mesure la liaison qu'il et en évidence. Une valeur propre proche de 1 ndique une dichotomie : pour chaque variable, deux groupes de modalités séparent le nuage de points. rest à interpéter si possible ces groupes de modalités en groupes de niveaux des facteurs.
Si deux valeurs propres sont proches de1, les modalités se décomposent en trois

groupes.
On peut interpéter la représentation comme la projection sur les axes principaux des deux nuages de profils, dans une métrique qui n'est plus la métrique euclidienne.