Analyse statistique mutlivariée

Antoine Gournay

Institut de Mathématiques,
Université de Neuchâtel
Suisse

Septembre, 2012

Notes de Cours

Table des matières

Introduction

La présente version des notes n'est encore qu'un brouillon (encore incomplet, probablement pas assez cohérent et riche en erreurs). Dessins, exemples et discussions viennent souvent à manquer ; l'index est incomplet. L'honorable lecteur est ainsi prié de transmettre à l'auteur les erreurs qu'il y trouvera, les références qui lui semblent faire défaut dans l'index, et les passages qui lui semblent trop confus ou trop peu illustrés (d'exemples ou de dessins).

Puisque ce cours s'adresse à des biologistes, il n'est pas attendu que le lecteur s'intéresse aux démonstrations. Elles sont présentées ici surtout parce qu'il y a assez peu de textes dans la littérature qui explique l'ACP [*PCA*, en anglais] et ses multiples variantes (comme l'ARD [*RDA*, en anglais]) de manière unifiée. Certains appendices dépassent largement ce que le lecteur est supposé connaître (*e.g.* les multiplicateurs de Lagrange).

Les interprétations géométriques ou physiques sont, de l'avis de l'auteur, cependant souvent utiles pour mieux comprendre les principes (et les limites !) des méthodes présentées.

Les buts de ce (court) cours sont

Chapitre 1 : Rappeler/introduire les matrices et exprimer les quantités usuelles de statistiques descriptives dans ce langage.

Chapitre 2 : Introduire une méthode de réduction (méthode R) : l'analyse en composante principale (ACP, *PCA* en anglais) et de l'analyse de redondance (ARD, *RDA* en anglais).

Chapitre 3 : Introduire une méthode de classification (méthode Q) :

Le texte des démonstrations ne contrastant pas beaucoup avec le reste, un ■ en marque la fin. Pour les mêmes raisons, un ★ termine le texte des définitions, un ♣ celui des exemples et un ♠ celui des remarques.

Voici quelques notations qui seront employées tout au long de ce texte :

$\mathbb{Z}$ pour l'ensemble des entiers (positifs, négatifs ou nul). $\mathbb{R}$ pour l'ensemble réel. Quelques notations qui seront utilisées pour des sous-ensembles de ceux-ci sont $\mathbb{Z}_{\geq a}$ (des entiers supérieurs ou égaux à a), $\mathbb{R}_{<a}$ ou $]-\infty, a[$ (des réels strictement inférieurs à a), $]a,b]$ (pour les réels strictement plus grand que a et plus petits ou égaux à b), etc...

Lorsqu'un symbole qui n'a pas été préalablement défini apparaît dans une égalité $:=$ c'est qu'il s'agit là de sa définition.

Finalement, quelques mots sur certaines abréviations :

1. *i.e.* (lat. *id est*) signifie plus ou moins "c'est la même chose que", "de manière équivalente", "en d'autres mots", ...

2. *e.g.* (lat. *exempli gracia*) signifie par exemple “par exemple”.
3. *c’àd.* (fr. c’est-à-dire) est un peu comme *i.e.* mais en moins chic.
4. “ou (équiv.)” pour préciser que la conjonction “ou” qui précède est en fait une équivalence mais que l’auteur n’a pas la motivation nécessaire pour la démontrer.
5. “... X (resp. Y) ... Z (resp. W) ...” signifie qu’on peut lire une première fois la définition/le théorème avec X et Z, puis une seconde fois mais avec Y et W au lieu de X et Z.
6. *mutatis mutandis* = “ce qui devait être changé ayant été changé”, une phrase très utile, comme le lecteur peut s’en douter, car il faut deviner ce qui doit être changé.
7. *ceteris paribus* [*sic stantibus*] = “toute chose égale par ailleurs”, pour bien spécifier qu’on change une quantité, mais que toutes les autres restent égales.

L’auteur aimerait remercier Béatrice de Tilière qui lui a donné les notes de cours des années précédentes (notes dont il s’est fortement inspirées). Un grand remerciement à Radu Slobodeanu (le conseiller en statistques de l’université) qui a pris le temps de lui fournir les jolis dessins et de lui expliquer certaines constructions.

Chapitre 1

Statistiques descriptives multivariées et Matrices

1.1 Généralités et vocabulaire

"La statistique" := méthode qui consiste à observer et étudier une/plusieurs propriétés communes chez un groupe d'être, de choses ou d'entités.

"Une statistique" := un nombre calculé à partir d'une population (d'être, de choses, ou d'entités).

"Population" := la collection (d'être, de choses, ou d'entités) ayant des propriétés communes. Terme hérité d'une des première applications de la statistique, la démographie ; *e.g.* un ensemble parcelles de terrain étudiées, une population d'insectes, l'ensemble des plantes d'une espèce donnée, une population d'humains.

"Individu" := élément de la population ; *e.g.* une parcelle, un insecte, une plante, un humain.

"Variable" := une des propriétés commune aux individus que l'on souhaite étudier. Peut-être :

— qualitative : appréciation de la parcelle, l'état de santé de l'insecte, couleur des pétales, appartenance religieuse

— quantitative [numérique] continues [pouvant prendre n'importe quelle valeur réelle] : le taux d'acidité du sol, la longueur de l'insecte, la longueur de la tige, l'indice de masse corporelle.

— quantitative [numérique] discrète [dès qu'il y a un saut minimum obligatoire entre deux valeur successives, *e.g.* les nombres entiers] : la somme (sur tous les jours) du nombre de vaches présentes sur la parcelle, l'âge de l'insecte (en jours), le nombre de pétales sur la fleur, le nombre d'année d'études (réussies) depuis la petite école.

La différence entre discrète est continue peut-être plus floue qu'il n'y paraît : les instruments de mesures ne sont pas d'une précision infinie, si les unités changent pour que 1 unité soit la précision de l'instrument, on se retrouve soudainement avec une valeur entière. Il s'agit donc plutôt d'une distinction sur la nature de la propriété mesurée.

"Échantillon" := individus de la population sur laquelle les mesures ont été faites.

"Données" := l'ensemble des valeurs de variables mesurées sur les individus de la population.

La collecte des données est une étape clef très subtile. Il faut s'assurer que l'échantillon pris est

représentatif de toute la population étudiée.

Exemple 1.1.1. Dans un sondage téléphonique il y a un biais naturel : les personnes qui sont souvent (ou longtemps) à la maison et qui ont le téléphone seront essentiellement les seules à être interrogées. Pourtant, elles ne représentent pas toute la population.

Un exemple plus simple : dans son école Toto demande à ses camarades le nombre de frères et sœurs qu'ils ont. Il trouve un nombre bien supérieur au taux de natalité officiel. La raison est qu'il ne fait que la moyenne des familles qui ont des enfants, et laisse pour contre toutes les familles avec 0 enfants. ♣

Deux directions en statistique :

1. Statistique descriptive : son but est de décrire, *c'àd.* de résumer ou représenter par des statistiques les données disponibles quand elles sont nombreuses. Questions types :
   - **a.** Représentation graphique.
   - **b.** Paramètres de position et dispersion.
   - **c.** Divers question liées aux grand jeux de données.
2. Statistique inférentielle : les données sont considérées incomplètes et elle a pour but de tenter de retrouver l'information sur la population initiale. La prémisse est que chaque mesure est une variable aléatoire suivant la loi de probabilité de la population. Questions types :
   - **a.** Estimations de paramètres.
   - **b.** Intervalles de confiance.
   - **c.** Tests d'hypothèse.
   - **d.** Modélisation (*e.g.* régression linéaire).

Ensuite, la statistique peut être :
- univariée : il n'y a qu'une seule variable qui rentre en jeu.
- multivariée : plusieurs variables rentrent en ligne de compte.

Exemple 1.1.2. Pour faire un rapide exemple, une statistique qui a eu un certain effet dans un pays qui restera anonyme est que la proportion de criminels chez les "étrangers" (*i.e.* les personnes ne possédant pas la citoyenneté) étaient bien plus élevée que chez les citoyens. Il s'adonne que les deux populations étaient complètement différentes, même d'un point de vue très superficielle : âge et sexe. En effet, si les populations avaient les mêmes proportions de personnes d'âge et de sexe donnés la situation se renverse : l'étranger moyen est moins criminel que le citoyen moyen. Seulement, l'étranger moyen appartient plus souvent à un "type" d'individu (homme entre 18 et 30 ans) qui est le plus souvent condamné dans la cours criminelle. ♣

Ce cours est un court cours de **Statistique descriptive multivariée**. Ce thème est très vaste. Une première étape consiste toujours à regarder les paramètre de position et de dispersion. Puis les méthodes se séparent en deux groupes :

1. Les méthodes factorielles, d'ordination ou de réduction ou encore méthodes R. Elles cherchent à réduire le nombre de variables en un petit nombre qui concentre toute l'information. L'analyse en composantes principales (ACP, en anglais *PCA*) est faite pour les variables quantitatives tandis que l'analyse [factorielle] de correspondance (AFC ou, en anglais, *CA*) s'occupe de variables qualitatives. C'est une méthode d'ordination "sans contraintes" (*unconstrained*). Elle sera complétée par l'analyse de redondance (*redundancy analysis* ou *RDA*, en anglais) une méthode d'ordination "avec contrainte" (*constrained*). Cette dernière est utilisée pour savoir si un large jeu de donnée pourrait "expliqué" un autre, *c'àd.* si un certains nombre de variables pourraient avoir une corrélation sur un autre groupe de variables.
2. Les méthodes de classifications ou méthodes Q. On commencera par une méthode de regroupement (ou agrégation, *clustering* en anglais) dite "non-supervisée" (*unsupervised*). Elle sera utilisée pour à réduire le nombre d'individus en formant des groupes le plus homogène possible. Ensuite, on tentera de faire une analyse discriminante (*discriminant analysis*) : étant donné un groupement pré-établi, on cherche à distinguer entre les groupes, ou associer une nouvelle observation à/caractériser un des groupes.

Une remarque d'ordre générale :

Remarque 1.1.3. Quelque soit l'étude que vous faites, il est bon de garder en tête ces quelques conseils généraux. Notez que ceux-ci peuvent (et devraient) être considéré AVANT de commencer l'étude.

- S'assurer de ne pas avoir de biais systématique.
- S'assurer d'une certaine homogénéité des "individus".
- Avoir le plus d'individus possibles. [1]
- Une fois les grandes lignes de l'expérience dessinées (mais avant de la réaliser), s'assurer d'avoir un moyen statistique approprié pour traiter les données et tester l'hypothèse qu'on cherche à établir ou infirmer. ♠

De nombreuses études qui sont relatées dans les journaux utiliser des populations au final très petites. Par exemple, l'auteur a vu des journaliste qui tente d'inférer à partir d'une étude sur le comportement de quelques 250 couples vivant dans un état rural des États-Unis[-d'Amérique] des principes généraux qui font qu'une relation dure. D'un point de vue scientifique c'est très bancal. D'une part, l'étude a eu le bon goût de tenter de prendre une population homogène (se restreindre à un état). Mais évidemment, ceci ne permet pas de conclure grand chose sur les couples chez les Lolos noirs, les Inuits, les Hollandais ou même des habitants d'une grande ville du même pays (vu la quantité notable de différence au niveau de l'environnement social). D'autre part, même à l'intérieur de l'état, il n'est pas clair que 250 couples soit un nombre suffisamment grand pour permettre de représenter toutes les variations possibles.

D'autre part, il arrive aussi régulièrement que des personnes se lancent dans des expérience, récoltent des données qui semblent utiles, puis une fois tout terminer tente de trouver une analyse

1. D'un point de vue statistique, 1000 individus c'est peu ; bien sûr, il y a des limites techniques ou financières qui peuvent forcer à en considérer moins.

statistique qui colle. Il se peut très bien que l'expérience s'avère non-significative parce que les données recueillies sont trop redondantes ou qu'on ait oublié de recueillir certaines données qui auraient pu être utiles ou qu'on ai oublié de mené d'autres tests en parallèle. Ainsi, il est chaudement recommandé de demandé l'avis d'un statisticien (ou conseiller en statistiques avant de commencer l'expérience).

1.2 Algèbre linéaire et matrices

Définition 1.2.1. Une **matrice** X de taille $n \times p$ est un tableau de nombre réels a n lignes et p colonnes :

$$X = \begin{pmatrix} x_{1,1} & x_{1,2} & x_{1,3} & \cdots & x_{1,(p-1)} & x_{1,p} \\ x_{2,1} & x_{2,2} & x_{2,3} & \cdots & x_{2,(p-1)} & x_{2,p} \\ x_{3,1} & x_{3,2} & x_{3,3} & \cdots & x_{3,(p-1)} & x_{3,p} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{(n-1),1} & x_{(n-1),2} & x_{(n-1),3} & \cdots & x_{(n-1),(p-1)} & x_{(n-1),p} \\ x_{n,1} & x_{n,n} & x_{n,3} & \cdots & x_{n,(p-1)} & x_{n,p} \end{pmatrix}$$

$$= \begin{pmatrix} x_{11} & x_{12} & x_{13} & \cdots & x_{1(p-1)} & x_{1p} \\ x_{21} & x_{22} & x_{23} & \cdots & x_{2(p-1)} & x_{2p} \\ x_{31} & x_{32} & x_{33} & \cdots & x_{3(p-1)} & x_{3p} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{(n-1)1} & x_{(n-1)2} & x_{(n-1)3} & \cdots & x_{(n-1)(p-1)} & x_{(n-1)p} \\ x_{n1} & x_{nn} & x_{n3} & \cdots & x_{n(p-1)} & x_{np} \end{pmatrix}$$

Les virgules étant omises lorsque qu'il n'y a pas ambiguïté. L'élément sur la $i^{\text{ème}}$ ligne et $j^{\text{ème}}$ colonne du tableau est aussi noté $(X)_{ij}$ ou x_{ij}.

Un **vecteur** colonne de taille n, noté $\mathbf{x}$, est une matrice de taille $n \times 1$. Un vecteur ligne de taille p, noté $\underline{x}$, est une matrice de taille $1 \times p$. Par abus de langage, pour noter les éléments d'un vecteur colonne, on écrit $x_i := (\mathbf{x})_{i1}$ et similairement pour un vecteur ligne, $x_i := (\underline{x})_{1i}$. ★

Pour la lisibilité, les entrées d'un vecteur ligne sont souvent séparées par des virgules : $\underline{x} = (x_1, x_2, \ldots, x_p)$.

1.2.i Vecteurs

L'espace des vecteurs (colonnes ou lignes) de taille n est noté $\mathbb{R}^n$.

Géométriquement, un vecteur indique la position dans les coordonnées cartésiennes. Si un repère est formé avec des axes (ici, les axes seront des variables quantitatives mesurées, *e.g.* la hauteur de la tige, le nombre de pétale sur la fleur, et le diamètre de la fleur), alors le vecteur servira à donner la position de l'individu dans cet "espace" : sa première coordonnée est la hauteur de la tige, etc…

Il est commode de voir un vecteur comme un déplacement depuis l'origine (*i.e.* le vecteur dont toutes les coordonnées sont nulles, *i.e.* le point de base du repère). Dans ce sens, la somme de deux vecteurs et la somme des déplacements. C'est le sens originel du mot vecteur : il représente un

déplacement (en ligne droite) entre deux points (ici entre l'origine et le point dont les coordonnées sont décrites dans le vecteur).

Définition 1.2.2. Ceci motive deux opérations usuelles sur les vecteurs :

1. L'addition : $\mathbf{x}$ et $\mathbf{y}$ deux vecteurs de même taille, alors $(\mathbf{x}+\mathbf{y})_i = x_i + y_i$.
2. Multiplication par un scalaire : si X est une matrice et $r \in \mathbb{R}$ un nombre réel (dans ce cours scalaire [2] n'est qu'un mot savant pour dire "nombre réel"), alors $(r\mathbf{x})_i = rx_i$.

Si un vecteur $\mathbf{x}$ peut être obtenu à partir d'autres vecteurs (disons $\mathbf{q}_{(1)}, \mathbf{q}_{(2)}, \ldots$ et $\mathbf{q}_{(k)}$) à partir des deux opérations ci-dessus, *i.e.*

$$\mathbf{x} = r_1\mathbf{q}_{(1)} + r_2\mathbf{q}_{(2)} + \ldots + r_k\mathbf{q}_{(k)},$$

alors $\mathbf{x}$ est dit une **combinaison linéaire** de $\mathbf{q}_{(1)}, \mathbf{q}_{(2)}, \ldots$ et $\mathbf{q}_{(k)}$. ★

Exemple 1.2.3. Les vecteurs $\begin{pmatrix}1\\2\\3\end{pmatrix}$ et $\begin{pmatrix}1\\0\\1\end{pmatrix}$ ont pour somme $\begin{pmatrix}2\\2\\4\end{pmatrix}$.

Tout vecteur $\begin{pmatrix}x_1\\x_2\end{pmatrix} \in \mathbb{R}^2$ s'écrit comme une combinaison linéaire de $\begin{pmatrix}1\\1\end{pmatrix}, \begin{pmatrix}-1\\1\end{pmatrix}$ et $\begin{pmatrix}0\\1\end{pmatrix}$. En effet,

$$\begin{pmatrix}x_1\\x_2\end{pmatrix} = x_2\begin{pmatrix}0\\1\end{pmatrix} + \tfrac{x_1}{2}\begin{pmatrix}1\\1\end{pmatrix} - \tfrac{x_1}{2}\begin{pmatrix}-1\\1\end{pmatrix}$$

ou encore :

$$\begin{pmatrix}x_1\\x_2\end{pmatrix} = \tfrac{x_1+x_2}{2}\begin{pmatrix}1\\1\end{pmatrix} + \tfrac{x_2-x_1}{2}\begin{pmatrix}-1\\1\end{pmatrix}.$$

Il y a donc plusieurs possibilités. ♣

Exemple 1.2.4. Soit $\mathbf{e}_{(k)} \in \mathbb{R}^n$ (où $k \in \{1,2,\ldots,n\}$) le vecteur tel que

$$(\mathbf{e}_{(k)})_i = \begin{cases} 1 & \text{si } k = i, \\ 0 & \text{si } k \neq i. \end{cases}$$

Autrement dit, le vecteur dont seule une coordonnée est non-nulle. L'ensemble de ces vecteurs est parfois appelé **base canonique** de $\mathbb{R}^n$. La raison est que, par les opérations ci-dessus :

$$\mathbf{x} = x_1\mathbf{e}_{(1)} + x_2\mathbf{e}_{(2)} + \ldots + x_k\mathbf{e}_{(k)}.$$

Tout vecteur est une combinaison linéaire d'élément de la base canonique. De plus, cette écriture est unique. ♣

Les définitions ci-dessous qui sont écrites pour des vecteurs colonnes, se transposent sans grande difficulté aux vecteurs lignes.

Définition 1.2.5. La norme [3] d'un vecteur $\mathbf{x}$ de taille n est le réel (positif ou nul) $\|\mathbf{x}\| = \sqrt{\sum_{i=1}^{n} x_i^2}$.

Le produit scalaire entre deux vecteurs $\mathbf{x}$ et $\mathbf{y}$ de même taille (disons n) est défini par $\langle \mathbf{x}, \mathbf{y} \rangle := \sum_{i=1}^{n} x_i y_i$.

★

2. la racine du mot est "échelle", pour dire qu'on fait un changement d'échelle.
3. Dans certains textes, le mot "module" est employé.

La norme au carré est le produit scalaire d'un vecteur avec lui-même : $\langle \mathbf{x}, \mathbf{x} \rangle = \sum_{i=1}^{n} x_i^2 = \|\mathbf{x}\|^2$.

Remarque 1.2.6. Par Pythagore, la norme est la distance entre l'origine et le point décrit par le vecteur dans le repère. ♠

Un vecteur peut toujours s'exprimer comme une amplitude (sa norme) et une direction (un vecteur de norme 1) : $\mathbf{x} = \|\mathbf{x}\|(\frac{1}{\|\mathbf{x}\|}\mathbf{x})$. Autrement dit, un déplacement depuis l'origine peut se décrire comme une suite de déplacement le long de directions fixées (les axes) ou comme une direction et une amplitude.

Proposition 1.2.7

Le produit scalaire a les propriétés suivantes :

- *Il est symétrique :* $\langle \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{y}, \mathbf{x} \rangle$;
- *Il est linéaire :* $\langle r\mathbf{x} + t\mathbf{y}, \mathbf{z} \rangle = r\langle \mathbf{x}, \mathbf{z} \rangle + t\langle \mathbf{y}, \mathbf{z} \rangle$.

DÉMONSTRATION: Il faut juste réécrire les termes de gauche et de droite en utilisant la définition, puis voir, en utilisant les propriétés habituelles des nombres réels (commutativité, associativité, etc...) que les deux côtés sont les mêmes. ■

Même si la seconde propriété ne porte que sur le premier vecteur dans le produit, par symétrie, elle s'applique aussi au deuxième. Par souci de culture générale, nous mentionnons l'inégalité de Cauchy-Schwartz

Proposition 1.2.8 (*Inégalité de Cauchy-Schwartz*)

Si $\mathbf{x}$ *et* $\mathbf{y}$ *sont des vecteurs de* $\mathbb{R}^m$*, alors* $|\langle \mathbf{x}, \mathbf{y} \rangle| \leq \|\mathbf{x}\|\|\mathbf{y}\|$.

DÉMONSTRATION: Si $\mathbf{x}$ ou $\mathbf{y}$ sont nuls, il n'y a rien à montrer (l'inégalité dit $0 \leq 0$ ce qui est vrai). Sinon, on regarde $P(t) = \|\mathbf{x} + t\mathbf{y}\|^2$. Comme la norme au carré d'un vecteur est le produit scalaire du vecteur avec lui-même, on réécrit, en utilisant les propriétés du produit scalaire,

$$\begin{array}{rclcrcl} P(t) & = & \langle \mathbf{x}+t\mathbf{y}, \mathbf{x}+t\mathbf{y} \rangle & = & \langle \mathbf{x}, \mathbf{x}+t\mathbf{y} \rangle + t\langle \mathbf{y}, \mathbf{x}+t\mathbf{y} \rangle & = & \langle \mathbf{x}+t\mathbf{y}, \mathbf{x} \rangle + t\langle \mathbf{x}+t\mathbf{y}, \mathbf{y} \rangle \\ & = & \langle \mathbf{x}, \mathbf{x} \rangle + t\langle \mathbf{y}, \mathbf{x} \rangle + t\langle \mathbf{x}, \mathbf{y} \rangle + t^2\langle \mathbf{y}, \mathbf{y} \rangle & & & = & \|\mathbf{x}\|^2 + 2t\langle \mathbf{x}, \mathbf{y} \rangle + t^2\|\mathbf{y}\|^2. \end{array}$$

Autrement dit, $P(t)$ est un polynôme de degré 2 en t. D'autre part, $P(t) \geq 0$. Ceci veut dire que le polynôme a au plus une solution (comme $\mathbf{y}$ et $\mathbf{x}$ ne sont pas nuls, ce polynôme est bien un polynôme de degré 2 qui ne peut pas toujours être 0). Le discriminant doit être ≤ 0, sinon on aurait deux solutions au polynôme. Mais le discriminant est

$$\Delta = 4\langle \mathbf{x}, \mathbf{y} \rangle^2 - 4\|\mathbf{x}\|^2\|\mathbf{y}\|^2.$$

Or $\Delta \leq 0$ implique

$$4\langle \mathbf{x}, \mathbf{y} \rangle^2 - 4\|\mathbf{x}\|^2\|\mathbf{y}\|^2 \leq 0 \iff \langle \mathbf{x}, \mathbf{y} \rangle^2 \leq \|\mathbf{x}\|^2\|\mathbf{y}\|^2 \iff |\langle \mathbf{x}, \mathbf{y} \rangle| \leq \|\mathbf{x}\|\|\mathbf{y}\|$$

Ce qui termine la démonstration. ■

Une formule géométriquement très importante pour exprimer le produit scalaire (et qui est une conséquence de l'inégalité ci-dessus) est

$$\langle \mathbf{x}, \mathbf{y} \rangle = \|\mathbf{x}\| \cdot \|\mathbf{y}\| \cdot \cos \angle(\mathbf{x}, \mathbf{y}),$$

où $\angle(\mathbf{x}, \mathbf{y})$ est l'angle entre les deux vecteurs. Le côté pratique de ceci est qu'un vecteur de norme 1 est une direction et :

Proposition 1.2.9

Si $\mathbf{y}$ est un vecteur de norme 1, alors $\langle \mathbf{x}, \mathbf{y} \rangle$ est le déplacement dans la direction de $\mathbf{y}$ qui est effectuée en se déplaçant de l'origine vers $\mathbf{x}$.

Exemple 1.2.10. Soit $\mathbf{y} = \binom{0}{1}$ et $\mathbf{x} = \binom{2}{1.5}$. Puisque $\mathbf{y}$ est la direction (vecteur de norme 1) du premier axe, le déplacement de $\mathbf{x}$ dans cette direction est, sans surprise, $\langle \mathbf{y}, \mathbf{x} \rangle = 1 \cdot 2 + 0 \cdot 1.5 = 2$.

Soit $\mathbf{y} = \binom{1/\sqrt{2}}{1/\sqrt{2}}$ et $\mathbf{x} = \binom{2}{1.5}$. Cette fois-ci $\mathbf{y}$ est toujours de norme 1, mais sa direction est la direction exactement entre les deux axes. Le déplacement de $\mathbf{x}$ dans cette direction est $\langle \mathbf{y}, \mathbf{x} \rangle = \frac{1}{\sqrt{2}} \cdot 2 + \frac{1}{\sqrt{2}} \cdot \frac{3}{2} = \frac{7}{2\sqrt{2}} \simeq 2.4748\ldots$ ♣

Un des points crucial de l'ACP, est de changer la manière d'exprimer les données en une autre plus claire. Le produit scalaire sera donc très utile puisqu'il nous permettra de traduire les données dans un repère vers un autre repère.

Définition 1.2.11. Une **base** de $\mathbb{R}^n$ est un ensemble de n vecteurs $\{\mathbf{q}_{(1)}, \mathbf{q}_{(2)}, \ldots, \mathbf{q}_{(n)}\}$ dans $\mathbb{R}^n$ tels que tout vecteur $\mathbf{x} \in \mathbb{R}^n$ possède une écriture *unique* comme une combinaison linéaire des $\mathbf{q}_{(i)}$.

Une base est dite **orthonormée** si

$$\langle \mathbf{q}_{(k)}, \mathbf{q}_{(\ell)} \rangle = \begin{cases} 1 & \text{si } k = \ell, \\ 0 & \text{si } k \neq \ell. \end{cases}$$

Autrement dit, tous les vecteurs sont de norme 1, et le produit scalaire de deux vecteur distincts est nul. ★

La définition ci-dessus est un peu redondante : un théorème nous assure que si n vecteurs de $\mathbb{R}^n$ permettent d'écrire tous les autres, alors l'écriture est unique. Inversement, si, dans $\mathbb{R}^n$, l'écriture est unique, alors il y aura n vecteurs.

Notons que les vecteurs d'une base orthonormée sont tous de norme 1 : puisque $\langle \mathbf{q}_{(k)}, \mathbf{q}_{(k)} \rangle = 1 = \|\mathbf{q}_{(k)}\|^2$, et donc, comme la norme est ≥ 0, $\|\mathbf{q}_{(k)}\| = 1$.

Le mot "coordonnée" sera dorénavant utiliser pour parler des réels qui interviennent devant l'écriture dans une base, *i.e.* si $\mathbf{x} = \sum_{i=1}^{n} a_i \mathbf{q}_{(i)}$ les coordonnées de $\mathbf{x}$ dans la base des $\mathbf{q}_{(i)}$ sont les a_i. Ceci est cohérent avec le faite que les x_i sont les coordonnées de $\mathbf{x}$ dans la base canonique.

Proposition 1.2.12

Si $\{\boldsymbol{q}_{(1)}, \boldsymbol{q}_{(2)}, \ldots, \boldsymbol{q}_{(n)}\}$ est une base orthonormée, l'écriture de $\mathbf{x}$ dans les coordonnées de cette base est :

$$\boldsymbol{x} = \langle \boldsymbol{x}, \boldsymbol{q}_{(1)} \rangle \boldsymbol{q}_{(1)} + \langle \boldsymbol{x}, \boldsymbol{q}_{(2)} \rangle \boldsymbol{q}_{(2)} + \ldots + \langle \boldsymbol{x}, \boldsymbol{q}_{(n)} \rangle \boldsymbol{q}_{(n)}.$$

DÉMONSTRATION: ... ■

Remarque 1.2.13. Ceci implique que l'écriture préserve bien la longueur :

$$\begin{aligned}\|\mathbf{x}\|^2 = \langle \mathbf{x}, \mathbf{x}\rangle &= \Big\langle \langle \mathbf{x}, \mathbf{q}_{(1)}\rangle \mathbf{q}_{(1)} + \langle \mathbf{x}, \mathbf{q}_{(2)}\rangle \mathbf{q}_{(2)} + \ldots + \langle \mathbf{x}, \mathbf{q}_{(n)}\rangle \mathbf{q}_{(n)}, \mathbf{x} \Big\rangle \\ &= \langle \mathbf{x}, \mathbf{q}_{(1)}\rangle^2 + \langle \mathbf{x}, \mathbf{q}_{(2)}\rangle^2 + \ldots + \langle \mathbf{x}, \mathbf{q}_{(n)}\rangle^2.\end{aligned}$$

Autrement dit, la somme au carré des coordonnées dans la nouvelle base donne toujours la longueur de $\mathbf{x}$. ♠

Exemple 1.2.14. La base canonique est une base orthonormée. Dans $\mathbb{R}^2$, une autre base orthonormée possible est $\{\binom{1/\sqrt{2}}{1/\sqrt{2}}, \binom{-1/\sqrt{2}}{1/\sqrt{2}}\}$. En fait, en utilisant les propriétés de cos et sin (*e.g.* $(\cos\theta)^2 + (\sin\theta)^2 = 1$), il est possible de voir que $\{\binom{\cos\theta}{\sin\theta}, \binom{-\sin\theta}{\cos\theta}\}$ est aussi une base orthonormée. ♣

1.2.ii Matrices

Dans la suite du cours les matrices (qui ne sont pas des vecteurs) seront la plupart du temps $n \times p$ (où n =nombre de variables et p =nombre d'individus dans l'échantillon) ou $n \times n$ (n =nombre de variables).

Définition 1.2.15. Quelques opérations usuelles sur les matrices :

1. L'addition : X et Y deux matrices de même taille, alors $(X+Y)_{ij} = (X)_{ij} + (Y)_{ij}$.
2. Multiplication par un scalaire : si X est une matrice et $r \in \mathbb{R}$ un nombre réel (dans ce cours **scalaire** n'est qu'un mot savant pour dire "nombre réel"), alors $(rX)_{ij} = r(X)_i j$.
3. Multiplication matricielle : si X est une matrice $n \times p$ et Y est une matrice $p \times q$, alors

$$(XY)_{ij} = \sum_{k=1}^{p} (X)_{ik}(Y)_{kj}.$$

   Cette règle est souvent appliquée en répétant le mantra "ligne avec colonne" pour dire que l'entrée ij est le produit scalaire de la i^{ligne} (vu comme un vecteur colonne) avec la $j^{\text{ème}}$ colonne (vue comme un vecteur).
4. Transposition : si X est une matrice de taille $n \times p$, la **transposée** de X est la matrice, notée X^T, donnée par $(X^\mathsf{T})_{ij} = (X)_{ji}$. X^T est de taille $p \times n$.

★

Même s'il n'est pas toujours possible de faire le produit matricielle entre deux matrices X et Y arbitraire, il est utile de remarquer qu'il est toujours possible de faire le produit (matriciel) XX^T ou $X^\mathsf{T}X$. À ce propos, il est bon de remarquer que, même si X et Y sont toutes deux de taille $n \times n$ (de sorte que XY et YX seront aussi de taille $n \times n$) le résultat de XY est différent de YX (sauf en de rares exceptions).

Remarque 1.2.16. Le produit scalaire est une composition de transposition et produit matriciel : $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^\mathsf{T}\mathbf{y}$. Notons aussi que, comme la transposée d'une matrice 1×1 est la même matrice 1×1, $\mathbf{x}^\mathsf{T}\mathbf{y} = (\mathbf{x}^\mathsf{T}\mathbf{y})^\mathsf{T} = \mathbf{y}^\mathsf{T}\mathbf{x}$.

En paticulier, le produit matriciel d'un vecteur ligne avec un vecteur colonne est $\underline{x}\mathbf{y} = \langle \underline{x}^\mathsf{T}, \mathbf{y} \rangle$. Par contre, $\mathbf{y}\underline{x} = A$ est une matrice $n \times n$ telle que $(A)_{ij} = y_i x_j$. ♠

L'intérêt de la multiplication matricielle est qu'elle permet d'exprimer une relation linéaire. Ceci mérite un court exemple.

Exemple 1.2.17. Toto ouvre son kiosque à Sandwich, il propose les sandwich suivants

ingrédient \ type	tomate-fromage	œuf	complet	pain-beurre
pain	1	1	1	1
tomate	2	0	1	0
salade	0	1	1	0
fromage	3	0	1	0
œuf	0	2	1	0
beurre	1	1	1	2

Ceci donne une matrice $X = \begin{pmatrix} 1&1&1&1 \\ 2&0&1&0 \\ 0&1&1&0 \\ 3&0&1&0 \\ 0&2&1&0 \\ 1&1&1&2 \end{pmatrix}$. Si quelqu'un commande 2 sandwich tomate-fromage, 1 aux œufs, 1 complet et 2 pain-beurre, les ingrédients utilisés seront :

$$\begin{pmatrix} 1&1&1&1 \\ 2&0&1&0 \\ 0&1&1&0 \\ 3&0&1&0 \\ 0&2&1&0 \\ 1&1&1&2 \end{pmatrix} \begin{pmatrix} 2 \\ 1 \\ 1 \\ 2 \end{pmatrix} = \begin{pmatrix} 6 \\ 4 \\ 2 \\ 7 \\ 3 \\ 8 \end{pmatrix}.$$

Par exemple, il aura besoin de 8 morceaux de beurre, 4 tomates, etc..... Pour ce qui est du prix d'achat des ingrédients, écrivons les dans le vecteur ligne : $\underline{c} = (.5, .3, .1, .4, .3, .1)$. Le coût des ingrédient d'un sandwich est alors :

$$(.5, .3, .1, .4, .3, .1) \begin{pmatrix} 1&1&1&1 \\ 2&0&1&0 \\ 0&1&1&0 \\ 3&0&1&0 \\ 0&2&1&0 \\ 1&1&1&2 \end{pmatrix} = (2.4, 1.3, 1.7, .7)$$

Toto devant se faire un peut de profit pour payer les autres frais de ces activités, le prix de vente (par ingrédient) sera $\underline{v} = \underline{c} + .2\underline{c} = 1.2\underline{c}$. Par exemple la commande ci-haut lui coûtera :

$$(2.4, 1.3, 1.7, .7) \begin{pmatrix} 2 \\ 1 \\ 1 \\ 2 \end{pmatrix} = 9.2$$

et lui rapportera 1.84. ♣

Proposition 1.2.18

Quelques propriétés utiles :

- $(X+Y)^\mathsf{T} = X^\mathsf{T} + Y^\mathsf{T}$;
- $(XY)^\mathsf{T} = Y^\mathsf{T} X^\mathsf{T}$;

— $(X^\mathsf{T})^\mathsf{T} = X$.

La démonstration de ces propriétés est par identification directe des coefficients.

Une matrice qui joue un rôle particulier, est la matrice identité, Id_n (ou Id tout court) de taille $n \times n$. Elle a la propriété que $\forall \mathbf{x} \in \mathbb{R}^n$, $\mathrm{Id}_n \mathbf{x} = \mathbf{x}$. Elle peut aussi être définie par ses coefficients :

$$(\mathrm{Id}_n)_{ij} = \begin{cases} 1 & \text{si } i = j, \\ 0 & \text{si } i \neq j. \end{cases}$$

i.e. ils sont 1 sur la diagonale et 0 ailleurs.

Proposition 1.2.19

Si A est une matrice de taille $q \times n$ et B une matrice de taille $n \times p$, $A\mathrm{Id}_n = A$ et $\mathrm{Id}_n B = B$. Si M et N sont deux matrices de taille $n \times n$ et $MN = \mathrm{Id}_n$ alors[4] *$NM = \mathrm{Id}_n$.*

Toujours dans le cadre de l'ACP, la matrice de changement de base sera celle qui joue le rôle le plus important pour nous. Cependant, pour préparer ce contexte il est préférable de mélanger des vecteurs lignes et des vecteurs colonnes. En effet, les positions seront des vecteurs lignes $\underline{\mathrm{x}}$, et les nouveaux éléments de la base des vecteurs colonnes $\mathbf{q}_{(i)}$. Tout ce qui est a été dit avant continu de s'appliquer, par exemple en posant $\mathbf{x} = \underline{\mathrm{x}}^\mathsf{T}$.

La proposition 1.2.12 se réécrit comme suit :

Proposition 1.2.20

Soit $\{\boldsymbol{q}_{(1)}, \boldsymbol{q}_{(2)}, \ldots, \boldsymbol{q}_{(n)}\}$ une base orthonormée, alors les coordonnées de $\underline{\mathrm{x}}$ dans la base donnée par les $\boldsymbol{q}_{(i)}$ est donné par la matrice Q telle que $(Q)_{ij} = (\boldsymbol{q}_{(j)})_i = q_{i;(j)}$ la $i^{\text{ème}}$ coordonnée du $j^{\text{ème}}$ vecteur :

$$\underline{\mathrm{y}} = \underline{\mathrm{x}} Q = \underline{\mathrm{x}} \begin{pmatrix} | & | & & | \\ \boldsymbol{q}_{(1)} & \boldsymbol{q}_{(2)} & \cdots & \boldsymbol{q}_{(n)} \\ | & | & & | \end{pmatrix}.$$

où $\begin{pmatrix} | \\ \boldsymbol{q}_{(j)} \\ | \end{pmatrix}$ signifie que la $j^{\text{ème}}$ colonne est le vecteur (colonne) $\boldsymbol{q}_{(i)}$.

En effet, par la définition du produit matriciel, $y_i = \underline{\mathrm{x}} \mathbf{q}_{(i)} = \langle \underline{\mathrm{x}}^\mathsf{T}, \mathbf{q}_{(i)} \rangle$, *c'àd.* $\underline{\mathrm{y}}$ est un vecteur (ligne) dont la $i^{\text{ème}}$ coordonnée est le déplacement de $\underline{\mathrm{x}}$ dans la direction de $\mathbf{q}_{(i)}^\mathsf{T}$.

Remarque 1.2.21. Comme les $\mathbf{q}_{(i)}$ ci-haut forment une base orthonormée, pour inverser le changement de coordonnée, il suffit d'utiliser que

$$(Q^\mathsf{T} Q)_{ij} = \mathbf{q}_{(i)}^\mathsf{T} \mathbf{q}_{(j)} = \langle \mathbf{q}_{(i)}, \mathbf{q}_{(j)} \rangle = \begin{cases} 0 & \text{si } i \neq j \\ 1 & \text{si } i = j \end{cases} = (\mathrm{Id})_{ij}.$$

Autrement dit, $Q^\mathsf{T} Q = \mathrm{Id}$. Ainsi, $QQ^\mathsf{T} = \mathrm{Id}$ et donc

$$\underline{\mathrm{x}} = \underline{\mathrm{x}} \mathrm{Id} = \underline{\mathrm{x}} Q Q^\mathsf{T} = \underline{\mathrm{y}} Q^\mathsf{T}$$

Pour s'en convaincre, il faut voir que la base canonique exprimée dans les coordonnées des $\mathbf{q}_{(i)}$ aura pour matrice de changement de coordonnée Q^T. ♠

1.2.iii La trace et le déterminant

La trace et le déterminant d'une matrice interviendront de manière fréquente dans la suite du cours. La définition du déterminant paraîtra certainement absconse au lecteur. En réalité, cette quantité a beaucoup de propriétés qui rend son calcul simple (même s'il peut être long). À la première lecture, la définition du déterminant n'est pas si importante, c'est surtout ses propriétés qui comptent.

LA TRACE

La trace est assez facile à introduire.

Définition 1.2.22. La **trace** d'une matrice de taille $n \times n$ est la somme de ses coefficients diagonaux :

$$\mathrm{Tr}A = \sum_{i=1}^{n} (A)_i i. \qquad \bigstar$$

Les propriétés de cette trace qui nous intéresse sont :

Lemme 1.2.23

Soit A un matrice de taille $p \times k$ alors

$$\mathrm{Tr}(A^\mathsf{T} A) = \mathrm{Tr}(AA^\mathsf{T})$$

Soit A et B des matrices de taille $p \times p$ alors

$$\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$$

DÉMONSTRATION: La première égalité est un calcul :

$$\begin{aligned}
\mathrm{Tr}(A^\mathsf{T} A) &= \sum_{i=1}^{k} (A^\mathsf{T} A)_{ii} &&= \sum_{i=1}^{k}\sum_{j=1}^{p} (A^\mathsf{T})_{ij}(A)_{ji} \\
&= \sum_{i=1}^{k}\sum_{j=1}^{p} (A)_{ji}(A)_{ji} &&= \sum_{i=1}^{k}\sum_{j=1}^{p} (A)_{ji}(A^\mathsf{T})_{ij} \\
&= \sum_{j=1}^{p}\sum_{i=1}^{k} (A)_{ji}(A^\mathsf{T})_{ij} &&= \sum_{j=1}^{p} (AA^\mathsf{T})_{jj} \\
&= \mathrm{Tr}(AA^\mathsf{T}).
\end{aligned}$$

En fait, on aurait pu s'arrêter au milieu du premier calcul : il apparaît que la trace de $A^\mathsf{T}A$ est la somme des coefficients de la matrice A au carré. Cette somme est identique à la somme des coefficients de la matrice A^T au carré... La seconde égalité est aussi un calcul direct :

$$\begin{aligned}
\mathrm{Tr}(AB) &= \sum_{i=1}^{p} (AB)_{ii} &&= \sum_{i=1}^{p}\sum_{j=1}^{p} (A)_{ij}(B)_{ji} \\
&= \sum_{i=1}^{p}\sum_{j=1}^{p} (B)_{ji}(A)_{ij} &&= \sum_{j=1}^{p}\sum_{i=1}^{p} (B)_{ji}(A)_{ij} \\
&= \sum_{j=1}^{p} (BA)_{jj} &&= \mathrm{Tr}(BA).
\end{aligned}$$

■

La définition du déterminant

Le déterminant est moins facile à introduire (ou à calculer). Il y a un ordre habituel sur $\{1,2,\ldots,n\}$ qui consiste à écrire les éléments en ordre croissant : $1,2,3,\ldots,n-1,n$.

Définition 1.2.24. Une **permutation** sur n éléments est une façon de réordonner l'ensemble $\{1,2,\ldots,n\}$. ★

Il est commun de voir une permutation comme une application $\sigma : \{1,2,\ldots,n\} \to \{1,2,\ldots,n\}$. En effet, $\sigma(i)$ est la position du nombre i dans le nouvel ordre.

Exemple 1.2.25. Par exemple, on pourrait renverser l'ordre en listant ces éléments comme suit : $n,n-1,\ldots,1$. Dans ce cas, $\sigma(i) = n-i+1$.

Exemple de la vie de tous les jours : dans une course, on donne un numéro à chaque participant. On mets ensuite ces numéros dans l'ordre d'arrivée des participants, $\sigma(i) =$ la position du coureur numéro i. ♣

Avec un peu de travail, on constante que
· $\sigma(i) = \sigma(j)$ si et seulement si $i = j$;
· et que pour tout k, il existe un j tel que $\sigma(j) = k$. Ce types d'application est aussi appelée "bijection" en langage savant.

Lorsque la permutation ne consiste qu'à échanger la place de deux éléments, on parle de transposition. Même si la démonstration n'est pas si simple, il est assez intuitif de voir qu'une permutation peut-être obtenue en faisant une suite de transposition [5]. Il y a bien sûr plusieurs moyens d'y arriver. Il est plus subtil de remarquer que, peu importe comment on écrit une permutation en tant que suite de transposition, la parité du nombre de transposition est toujours la même :

Lemme 1.2.26

Il y a deux type de permutations : les permutations σ qui s'écrivent comme un nombre pair de transpositions, et celles qui s'écrivent comme un nombre pair de transpositions.

Pour la remarque, ce lemme permet de faire quelques tours de "magie".

Définition 1.2.27. Le **signe** d'une permutation est 1 si elle s'écrit comme un nombre pair de transposition et -1 sinon. ★

Exemple 1.2.28. Sur l'ensemble à trois éléments, il y a 6 permutations possibles :

signe	$+1$	-1
	$1,2,3$	$1,3,2$
	$2,3,1$	$3,2,1$
	$3,1,2$	$2,1,3$

5. Pour reprendre la métaphore de la course, une transposition correspond à un coureur qui double un autre coureur.

La première permutation (en haut à gauche) est l'ordre "habituel". Voici les 24 permutations sur l'ensemble à 4 éléments

signe	$+1$		-1	
	$1,2,3,4$	$1,4,2,3$	$1,3,2,4$	$1,4,3,2$
	$2,3,1,4$	$2,4,3,1$	$3,2,1,4$	$3,4,2,1$
	$3,1,2,4$	$3,4,1,2$	$2,1,3,4$	$2,4,1,3$
	$1,3,4,2$	$4,1,3,2$	$1,2,4,3$	$4,1,2,3$
	$3,2,4,1$	$4,3,2,1$	$2,3,4,1$	$4,2,3,1$
	$2,1,4,3$	$4,2,1,3$	$3,1,4,2$	$4,3,1,2$

Le lecteur pourra s'apercevoir que les permutations sur 4 éléments sont obtenues de celle sur 3 éléments en ajoutant un 4 quelque part dans la liste. ♣

Cette remarque permet de calculer le nombre de permutation :

Lemme 1.2.29

Il y a $n! = n \cdot (n-1) \cdot (n-2) \cdots 3 \cdot 2 \cdot 1$ *permutations sur n éléments.*

DÉMONSTRATION: Il s'agit d'un induction assez simple : supposons qu'il y a k permutations sur l'ensemble à $n-1$ éléments. Pour créer une permutation de l'ensemble à n éléments, il suffit d'ajouter n quelque part dans un ordre sur $n-1$ éléments [6]. Il y a n choix possibles. De plus, si σ et σ' sont deux ordres différents, peu importe où on ajoute n dans σ et dans σ' deux ordres différents sont obtenus [7]. Du coup, il y a $n \cdot k$ permutations sur n éléments.

Soit k_i le nombre de permutation sur i éléments. On trouve d'abord $k_1 = 1$ (ça c'est facile). Puis en utilisant l'argument qu'on vient de donner : $k_2 = 2k_1 = 2 \cdot 1 = 2$, $k_3 = 3k_2 = 3 \cdot 2 \cdot 1 = 6$, etc... ■

Les préliminaires étant faits, la définition peut être donnée.

Définition 1.2.30. Soit $\mathcal{S}_n$ l'ensemble de toutes les permutations sur n éléments, écrites comme fonctions de $\{1,\ldots,n\}$ dans lui-même. Le **déterminant** d'une matrice A de taille $n \times n$ est le nombre réel donné par

$$\mathrm{Det}A = \sum_{\sigma \in \mathcal{S}_n} \mathrm{sgn}\sigma \prod_{i=1}^{n} (A)_{i\sigma(i)}. \qquad \bigstar$$

La somme ci-haut contient donc $n!$ termes, ce qui devient rapidement monstrueux. Sans compter qu'on risque d'oublier une permutation...

Exemple 1.2.31. Pour les matrice 2×2, c'est très facile à calculer :

$$\mathrm{Det}\begin{pmatrix} a & b \\ c & d \end{pmatrix} = ad - bc \qquad \clubsuit$$

6. Dans les termes de la course : on fait courir les $n-1$ premiers participant, et on obtient une permutation sur $n-1$ éléments. Ensuite, on fait courir le $n^{\text{ème}}$ et on le place dans le classement pour avoir une permutation.

7. En effet, si ces deux ordres sur n coureurs étaient les mêmes, on pourrait "disqualifier" n et on tomberait sur un seul et même classement sur $n-1$ coureurs.

Pour les matrices 3×3, ça reste faisable. Mais déjà, pour une matrice 4×4 les calculs sont trop longs.

LES PROPRIÉTÉS DU DÉTERMINANT

Pour l'évaluation du déterminant, beaucoup de méthodes sont disponibles. Avant d'en présenter une, il faut introduire une notation : lorsque A est une matrice $n \times n$, soit $A^{[ij]}$ la même matrice où on aurait retiré la $i^{\text{ème}}$ et la $j^{\text{ème}}$ colonne.

Proposition 1.2.32

Soit A une matrice de taille $n \times n$ et $i \in \{1, 2, \ldots, n\}$. Alors

$$\mathrm{Det}A = \sum_{j=1}^{n} (-1)^{i+j} (A)_{ij} \mathrm{Det}A^{[ij]} \quad \textit{et aussi} \quad \mathrm{Det}A = \sum_{j=1}^{n} (-1)^{i+j} (A)_{ji} \mathrm{Det}A^{[ji]}.$$

IDÉE DE LA DÉMONSTRATION: L'ingrédient principal de la démonstration est présent dans l'exemple 1.2.28 et dans le lemme 1.2.29. Un élément de $\mathcal{S}_n$ peut s'écrire comme un choix pour la position de n puis une permutation sur $n-1$ éléments (*i.e.* quelque chose de $\mathcal{S}_{n-1}$). Ainsi, la somme $\sum_{\sigma \in \mathcal{S}_n}$ peut s'écrire comme une somme de sommes : $\sum_{j=1}^{n} \sum_{\sigma' \in \mathcal{S}_{n-1}}$. La permutation σ est construite de j et de σ' en rajoutant n en $j^{\text{ème}}$ position dans l'ordre de σ'.

Il faut alors faire un peu attention aux signes. L'idée est alors la suivante : si n est ajouté en dernière position, ça ne change pas le signe, $\mathrm{sgn}\sigma = \mathrm{sgn}\sigma$. En effet, comme il est à la même place qu'au début, σ s'obtient exactement par la même suite de transposition [8] que σ'. Ensuite pour mettre n une position plus tôt, on fait une transposition. Donc pour mettre n en position $n-k$ on fait k transpositions de plus. Le signe change donc d'un facteur $(-1)^k$. Ceci permet de calculer le signe (le calcul précis ne sera pas fait ici).

En mettant tout ça correctement ensemble on obtient la formule pour $i = n$. Le cas général demande de porter plus attention aux détails... ■

La propriété géométrique importante du déterminant est que c'est un volume avec un signe. Étant donné deux vecteur de $\mathbb{R}^2$, on peut produire un parallélogramme en regardant les 4 points $\mathbf{0}$, $\mathbf{a}$, $\mathbf{b}$ et $\mathbf{a}+\mathbf{b}$. Plus généralement, étant donné n vecteurs de $\mathbb{R}^n$, on peut former un objet similaire (déterminer par 2^n points) nommé parallélépipède.

Proposition 1.2.33

Étant donné n vecteurs dans $\mathbb{R}^n$. Si A est la matrice qui possède ces vecteurs pour colonnes, alors Det$A$ *est, à un signe près, le volume du parallélépipède engendrer par ces n vecteurs.*

Cette propriété est un réécriture géométrique de la proposition précédente. La démonstration est surtout technique.

En fait, la proposition 1.2.33 dit plus que ce qu'elle annonce. Le cube Q_n dans $\mathbb{R}^n$ est l'ensemble des vecteurs dont les coordonnées sont dans $[0,1]$:

$$Q_n = [0,1]^n = \{\mathbf{v} \in \mathbb{R}^n \mid \text{pour tout } i, v_i \in [0,1]\}.$$

8. Dans la métaphore de la course : comme n arrive en $n^{\text{ème}}$ position (*i.e.* en dernière position) il n'a dépassé personne, donc il y a le même nombre de dépassements dans la course sans n et dans celle avec n.

Une matrice A de taille $n \times n$ envoie un vecteur de longueur n sur un autre vecteur de longueur n. Ainsi elle transforme le cube Q_n en un autre ensemble. Cet ensemble est le parallélépipède mentionné ci-dessus. Le lemme suivant utilise le fait que le déterminant dit à quel point l'application A déforme le volume.

Lemme 1.2.34

Soit A et B deux matrices de taille $n \times n$ alors $\mathrm{Det}(AB) = (\mathrm{Det}A)(\mathrm{Det}B)$.

DÉMONSTRATION: La démonstration utilise la vision géométrique du déterminant, *i.e.* la proposition 1.2.33. D'une part, $\mathrm{Det}(AB)$ est le volume du parallélépipède $(AB)Q_n$. De l'autre la matrice B envoie un cube Q_n de volume 1 sur un parallélépipède P de volume Det$B$. Quitte à faire une petite erreur (le bord du parallélépipède n'est pas très cubique), on redécoupe ce parallélépipède P en une myriade de petits cubes. Comme A est une application linéaire (*i.e.* $A(\mathbf{v}+\mathbf{w}) = A\mathbf{v}+A\mathbf{w}$), elle enverra chacun de ces petits cubes sur un petit parallélépipède. Le volume du petit parallélépipède est Det$A$ fois celui du petit cube (comme $A(r\mathbf{v}) = rA\mathbf{v}$). Donc le volume de $A(BQ_n)$ est Det$A$ fois le volume de BQ_n (qui est Det$B$). ■

Finalement, le résultat suivant sera souvent utilisé :

Lemme 1.2.35

Soit A une matrice de taille $n \times n$ alors, les trois conditions suivantes sont équivalentes :

(1) $\mathrm{Det}A = 0$;

(2) *il existe un vecteur colonne de taille n, disons $\mathbf{v} \in \mathbb{R}^n$, différent de $\mathbf{0}$, tel que $A\mathbf{v} = \mathbf{0}$;*

(3) *A possède NE possède PAS d'inverse,* i.e. *il N'existe PAS de matrice A^{-1} telle que* $A^{-1}A = \mathrm{Id}_n$.

DÉMONSTRATION: Supposons que la condition (1) est vérifiée, *i.e.* Det$A = 0$. Par la proposition 1.2.33, ceci veut dire que le parallélépipède (formé par les vecteur $\mathbf{a}_{(i)}$ qui sont des colonnes de A) est de volume nul. Ceci se produit seulement si un des vecteurs est une combinaison linéaire des autres : disons $\mathbf{a}_{(1)} = c_2\mathbf{a}_2 + \ldots + c_n\mathbf{a}_{(n)}$. Mézalors, si $\mathbf{v} = (-1, c_2, c_3, \ldots, c_n)^{\mathsf{T}}$, on trouve $A\mathbf{v} = \mathbf{0}$. Ceci montre $(1) \Rightarrow (2)$

Supposons que la condition 2 est vérifiée, *i.e.* on connaît un $\mathbf{v}$ avec $A\mathbf{v} = \mathbf{0}$. Pour n'importe quelle matrice B, $B\mathbf{0} = \mathbf{0}$. Si A possède un inverse A^{-1}, alors $\mathbf{v} = \mathrm{Id}_n\mathbf{v} = A^{-1}A\mathbf{v} = A^{-1}\mathbf{0} = \mathbf{0}$. Ceci est une contradiction ; ainsi, A ne possède pas d'inverse. Ainsi $(2) \Rightarrow (3)$

Supposons que la condition 1 N'est PAS vérifiée, *i.e.* que Det$A \neq 0$. Alors par la proposition 1.2.33, les vecteurs colonnes sont linéairement indépendants. Puisqu'il y en a n, ils forment une base. Pour trouver A^{-1}, il faut écrire chacun des éléments de la base canonique comme une combinaison linéaire des $\mathbf{a}_{(i)}$: $\mathbf{e}_{(j)} = \sum_i c_{j;i}\mathbf{a}_{(i)}$. Si C est la matrice avec coefficients $(C)_{ji} = c_{j;i}$, la définition du produit matriciel donne que $CA = \mathrm{Id}_n$ et donc que $C = A^{-1}$. Ainsi on a montré que "(non 1) $\Rightarrow$ (non 3)" ce qui est équivalent à "$(3) \Rightarrow (1)$".

Au final, on a $(1) \Rightarrow (2) \Rightarrow (3) \Rightarrow (1)$. Ce qui veut dire que les 3 conditions sont équivalentes. ■

Quitte à être redondant, voici le même résultat exprimé pour les condition opposées.

Lemme 1.2.36

Soit A une matrice de taille $n \times n$ alors, les trois conditions suivantes sont équivalentes :

(1') $\mathrm{Det}A \neq 0$;

(2') *si $\mathbf{v} \in \mathbb{R}^n$ est un vecteur colonne de taille n, alors $A\mathbf{v} = \mathbf{0}$ implique que $\mathbf{v} = 0$;*

(3') *A possède possède un inverse,* i.e. *il existe une matrice A^{-1} telle que $A^{-1}A = \mathrm{Id}_n = AA^{-1}$.*

1.3 Moyenne et variance, écriture matricielle

Les données recueillies sont mises dans un tableau, qui sera dorénavant une matrice notée X de taille $n \times p$. x_{ij} est la valeur prise par la $j^{\text{ème}}$ variable sur le $i^{\text{ème}}$ individu.

$$X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}$$

La $i^{\text{ème}}$ ligne

$$\underline{\mathrm{x}}_{(i)} = (x_{i1}, ..., x_{ip}), \forall i = 1, ..., n;$$

représente le $i^{\text{ème}}$ individu (ou *sample* ou site). La $j^{\text{ème}}$ colonne

$$\mathbf{x}_{(j)} = \begin{pmatrix} x_{1j} \\ \vdots \\ x_{nj} \end{pmatrix}, \forall j = 1, ..., p;$$

représente l'ensemble de toutes les valeurs (ou mesures) de la $j^{\text{ème}}$ variable (ou descripteur ou *species*).

1.3.i Paramètre de position : moyenne

Il y a trois paramètres de position : la moyenne, la médiane et le mode. Étant donné plusieurs mesures d'une variable (*e.g.* $\mathbf{x}_{(j)}$ ci-haut) la moyenne est

$$\widehat{\mathrm{x}}_{(j)} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_{(j)})_i = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_{i;(j)} = \frac{1}{n}\sum_{i=1}^{n}x_{ij} = \frac{1}{n}\langle \mathbb{1}_n, \mathbf{x}_{(j)} \rangle$$

où $\mathbb{1}_n$ est le vecteur (colonne) de longueur n dont toutes les entrées sont égales à 1. On place souvent la moyenne dans une matrice :

$$\hat{X} = \begin{pmatrix} \widehat{\mathrm{x}}_{(1)} & \dots & \widehat{\mathrm{x}}_{(p)} \\ \widehat{\mathrm{x}}_{(1)} & \dots & \widehat{\mathrm{x}}_{(p)} \\ \vdots & \ddots & \vdots \\ \widehat{\mathrm{x}}_{(1)} & \dots & \widehat{\mathrm{x}}_{(p)} \end{pmatrix} = \mathbb{1}_n(\tfrac{1}{n}\mathbb{1}_n^{\mathsf{T}}X)$$

les moyennes ont été mises dans un vecteur (ligne) : $\underline{\mathrm{m}}_X := \frac{1}{n}\mathbb{1}_n^\mathsf{T} X = (\widehat{\mathrm{x}}_{(1)}, \ldots, \widehat{\mathrm{x}}_{(p)})$. Ce vecteur a un sens géométrique. En effet, si on voit les individus comme des billes (de masses égales) placées au coordonnées $\underline{\mathrm{x}}_{(i)}$ dans l'espace p-dimensionnel défini par les variables, alors $\underline{\mathrm{m}}$ est le centre de masse de toutes ces billes.

Exemple 1.3.1. ... ♣

Proposition 1.3.2

Si $\mathbf{x}$ *et* $\mathbf{y}$ *sont deux variables de même taille (*i.e. *deux vecteurs de* $\mathbb{R}^n$*). Soit* $\mathbf{z} = a\mathbf{x} + b\mathbf{y} + c\mathbb{1}_n$ *où* $a, b, c \in \mathbb{R}$*. Alors* $\widehat{\mathrm{z}} = a\widehat{\mathrm{x}} + b\widehat{\mathrm{y}} + c$.

DÉMONSTRATION: Exercice (utiliser les propriétés du produit scalaire). ■

La médiane d'une variable est la valeur qui sépare en deux l'échantillon : $M_{(j)}$ est médiane pour la $j^{\text{ème}}$ variable si le nombre de fois que cette variable prend une valeur supérieure ou égal à $M_{(j)}$ est égal au nombre de fois que cette variable prend un nombre inférieure ou égal à $M_{(j)}$:

$$\sharp\{i \mid x_{ij} \geq M_{(j)}\} = \sharp\{i \mid x_{ij} \leq M_{(j)}\}.$$

Il y a parfois plusieurs nombres qui peuvent remplir ce rôle. Dans les cas litigieux, on prend la valeur intermédiaire à toutes ces valeurs possible. Par exemple, si les valeurs d'une variable sont $1, 2, 3, 4$, tout nombre (strictement) entre 2 et 3 satisfait la propriété ci-haut. Dans ce cas, on pose que la médiane est 2.5.

Une autre façon de faire est de classer les données en ordre croissant, puis de prendre $k^{\text{ème}}$ valeur (si $n = 2k+1$ est impair) ou alors le nombre au milieu de la $k^{\text{ème}}$ et $k+1^{\text{ème}}$ valeur si $n = 2k$ est pair.

Le mode est la valeur le plus souvent prise par une variable (pour une variable quantitative discrète ou une variable qualitative). Lorsque la variable est quantitative continue, il vaut mieux se fixer un nombre ε, puis chercher la valeur M telle que le nombre de valeurs entre $M-\varepsilon$ et $M+\varepsilon$ est le plus grand possible. Il n'y a pas de choix du nombre ε qui soit plus raisonnable que d'autre. Il dépend de l'expérience, de l'incertitude des instruments de mesure et du sens de la variable.

Remarque 1.3.3. La médiane et le mode ne sont généralement pas pratiques à traiter (même si ils sont néanmoins significatifs, voire plus significatifs). Il est en général une bonne idée de vérifier que la moyenne n'est pas trop éloignées de ces deux valeurs pour s'assurer que notre analyse statistique a un sens. ♠

Exemple 1.3.4. On récolte les données suivantes on mesure les concentrations de deux bactéries dans les intestins d'un mammifère et on note l'âge du spécimen. Il y a 5 individus. La matrice des données est

$$X = \begin{pmatrix} 4.62 & 10.00 & 4 \\ 5.23 & 10.00 & 4 \\ 5.08 & 9.64 & 3 \\ 5.28 & 8.33 & 4 \\ 5.76 & 8.93 & 5 \end{pmatrix}. \text{ Les moyennes sont } \begin{array}{lll} \widehat{\mathrm{x}}_{(1)} & = \frac{4.62+5.23+5.08+5.28+5.76}{5} & = 5.194 \\ \widehat{\mathrm{x}}_{(2)} & = \frac{10+10+9.64+8.33+8.93}{5} & = 9.38 \\ \widehat{\mathrm{x}}_{(3)} & = \frac{4+4+3+4+5}{5} & = 4 \end{array}$$

Les médianes sont 5.23 pour $\mathbf{x}_{(1)}$, 9.64 pour $\mathbf{x}_{(2)}$ et 4 pour $\mathbf{x}_{(3)}$. Le seul mode dont on puisse parler est celui de $\mathbf{x}_{(3)}$ qui est 4. La matrice des moyennes est :

$$\hat{X} = \begin{pmatrix} 5.194 & 9.38 & 4 \\ 5.194 & 9.38 & 4 \\ 5.194 & 9.38 & 4 \\ 5.194 & 9.38 & 4 \\ 5.194 & 9.38 & 4 \end{pmatrix} \qquad \clubsuit$$

1.3.ii Paramètre de dispersion : variance

La moyenne donne donc le "centre de masse" du nuage de points formé par les données. Ceci n'indique cependant pas grand chose. On fera donc appel à une autre quantité : la variance. Celle-ci est une tentative de calculer "la moyenne de l'écart à la moyenne". Encore une fois pour des raisons de calcul, il est plus facile de calculer la moyenne du carré de l'écart à la moyenne.

$$\mathrm{Var}(\mathbf{x}_{(j)}) = \tfrac{1}{n}\sum_{i=1}^{n}(x_{ij} - \widehat{\mathrm{x}}_{(j)})^2 = \frac{1}{n}\|\mathbf{x}_{(j)} - \widehat{\mathrm{x}}_{(j)}\mathbb{1}_n\|^2$$

L'écart-type est alors défini comme $\sigma(\mathbf{x}_{(j)}) = \sqrt{\mathrm{Var}(\mathbf{x}_{(j)})}$. Afin de raccourcir les notations, on écrira souvent $\sigma_{(j)}$ et $\mathrm{Var}_{(j)}$.

La variance possède une interprétation géométrique aussi : il s'agit de l'inertie le long de l'axe donné par la $j^{\text{ème}}$ variable [9].

Théorème 1.3.5

$\mathrm{Var}(\mathbf{x}_{(j)}) = \frac{1}{n}\|\mathbf{x}_{(j)}\|^2 - \widehat{\mathrm{x}}_{(j)}^2 = \left(\frac{1}{n}\sum_{i=1}^{n} x_{ij}^2\right) - \widehat{\mathrm{x}}_{(j)}^2$, c'àd. *la variance est la moyenne des carrés moins le carré de la moyenne.*

DÉMONSTRATION: Il s'agit d'un calcul assez direct, ici en formulation "matricielle".

$$\begin{aligned}
\tfrac{1}{n}\|\mathbf{x}_{(j)} - \widehat{\mathrm{x}}_{(j)}\mathbb{1}_n\|^2 &= \tfrac{1}{n}\langle \mathbf{x}_{(j)} - \widehat{\mathrm{x}}_{(j)}\mathbb{1}_n, \mathbf{x}_{(j)} - \widehat{\mathrm{x}}_{(j)}\mathbb{1}_n\rangle \\
&= \tfrac{1}{n}\Big(\langle \mathbf{x}_{(j)}, \mathbf{x}_{(j)}\rangle - \langle \mathbf{x}_{(j)}, \widehat{\mathrm{x}}_{(j)}\mathbb{1}_n\rangle - \langle \widehat{\mathrm{x}}_{(j)}\mathbb{1}_n, \mathbf{x}_{(j)}\rangle + \langle -\widehat{\mathrm{x}}_{(j)}\mathbb{1}_n, \widehat{\mathrm{x}}_{(j)}\mathbb{1}_n\rangle\Big) \\
&= \tfrac{1}{n}\Big(\|\mathbf{x}_{(j)}\|^2 - 2\widehat{\mathrm{x}}_{(j)}\langle \mathbf{x}_{(j)}, \mathbb{1}_n\rangle + n\widehat{\mathrm{x}}_{(j)}^2\Big) \\
&= \tfrac{1}{n}\Big(\|\mathbf{x}_{(j)}\|^2 - 2\widehat{\mathrm{x}}_{(j)}(n\widehat{\mathrm{x}}_{(j)}) + n\widehat{\mathrm{x}}_{(j)}^2\Big) \\
&= \tfrac{1}{n}\|\mathbf{x}_{(j)}\|^2 - \widehat{\mathrm{x}}_{(j)}^2
\end{aligned}$$

La première utilise le lien entre norme et produit scalaire, la seconde et la troisième la linéarité du produit scalaire, la quatrième utilise la définition de la moyenne. ■

9. l'inertie est ce qui remplace agréablement la masse lors de la transcription des formules de la cinétique "simple" (des translations) vers la cinétique des corps rigides en rotation. L'inertie le long d'un axe est l'inertie si la rotation considérée fixe cet axe.

Exemple 1.3.6. Suite de l'exemple 1.3.4. La matrice des données moins celle des moyennes est

$$X-\hat{X}=\begin{pmatrix}4.62 & 10.00 & 4\\ 5.23 & 10.00 & 4\\ 5.08 & 9.64 & 3\\ 5.28 & 8.33 & 4\\ 5.76 & 8.93 & 5\end{pmatrix}-\begin{pmatrix}5.194 & 9.38 & 4\\ 5.194 & 9.38 & 4\\ 5.194 & 9.38 & 4\\ 5.194 & 9.38 & 4\\ 5.194 & 9.38 & 4\end{pmatrix}=\begin{pmatrix}-0.574 & 0.62 & 0\\ 0.036 & 0.62 & 0\\ -0.114 & 0.26 & -1\\ 0.086 & -1.05 & 0\\ 0.566 & -0.45 & 1\end{pmatrix}$$

On peut alors calculer la matrice des variances (les nombres sont arrondis à la $3^{\text{ème}}$ décimale) :

$$V_X=\frac{1}{5}(X-\hat{X})^{\mathsf{T}}(X-\hat{X})=\begin{pmatrix}0.134 & -0.141 & 0.136\\ -0.141 & 0.428 & -0.142\\ 0.136 & -0.142 & 0.4\end{pmatrix}$$

Il est toujours bon de vérifier que les entrées sur la diagonales de la matrice des variances sont positives. Elles sont aussi extrêmement rarement nulles : la variance est nulle seulement lorsque toutes les mesures d'une variable sont identiques ! ♣

Il est souvent utile de considérer les variables centrées réduites :

$$Z=(X-\hat{X})S^{-1}$$

où S^{-1} est une matrice diagonale (*i.e.* $(S^{-1})_{ij}=0$ si $i\neq j$) et $(S^{-1})_{jj}=\sigma_{(j)}^{-1}$. Une autre façon de réécrire la formule ci-dessus est :

$$\mathbf{z}_{(j)}=\frac{\mathbf{x}_{(j)}-\widehat{\mathrm{x}}_{(j)}\mathbb{1}_n}{\sigma_{(j)}}\qquad\text{ou}\qquad z_{ij}=\frac{x_{ij}-\widehat{\mathrm{x}}_{(j)}}{\sigma_{(j)}}.$$

Le mot centré vient du fait que le centre de masse est maintenant à l'origine (le vecteur 0), ou autrement dit que les moyennes des $\mathbf{z}_{(j)}$ sont toutes nulles. Le mot réduite vient du fait que leur variance est toujours 1. Une variable est centrée réduite si elle satisfait ces deux conditions (moyenne nulle et variance égale à 1).

Remarque 1.3.7. Il y a deux avantages à utiliser des variables centrées réduites :

1- Une variable $\mathbf{x}$ est obtenue par des mesures. Si on avait changer le 0 sur notre échelle (on aurait alors obtenu comme mesures $\mathbf{x}'=\mathbf{x}+r\mathbb{1}_n$), les variable centrées réduites associées à $\mathbf{x}$ et $\mathbf{x}'$ sont les mêmes. Par exemple, si on mesure la température sur un site, on pourrait choisir les degrés Celsius ou les degrés Kelvin. Ce choix a une influence sur les valeurs de la variable, mais pas sur celles de la variable centrée réduite.

2- Si on avait changer l'échelle de la mesure par un facteur multiplicatif, cela n'affecte pas non plus la variable centrée réduite. Pour reprendre l'exemple de la température, on aurait aussi pu choisir entre degrés Celsius et degré Farenheit. Ce choix n'a pas d'influence sur la variable centrée réduite.

En bref, la variable centrée réduite gomme les choix reliés aux unités de mesures. Ceci est un avantage dans certaines situations, mais peut aussi être indésirable dans certains cas. ♠

Prendre les variables centrées réduites est aussi une bonne idée si on veut éviter que le fait qu'une variable prenne des valeurs très grandes et très petites (*i.e.* varie énormément) ne lui donne trop d'importance. Inversement, une variable qui varie peu prendra plus d'importance lors du passage au variables centrées réduites.

Exemple 1.3.8. ... ♣

Il existe d'autres paramètres de dispersion habituels. Le premier est l'étendue : la différence entre la plus grande est la plus petite valeur prise par une variable. Il n'est cependant pas commode à manipuler et le lien attendu qu'il entretien avec l'écart-type dépend de la taille de l'échantillon. Le second est le "véritable écart-type", *c'àd.* la moyenne des écarts à la moyenne en valeur absolue : $\frac{1}{n}\sum_{i=1}^{n}|x_{ij}-\widehat{\mathrm{x}}_{(j)}|$. Il est tout aussi difficile à tenir en compte dans les calculs.

1.3.iii Covariance

La covariance indique dans quelle mesure deux variables varient ensembles. Vu la nature du produit scalaire, une définition assez naturelle est :

$$\mathrm{Cov}(\mathbf{x}_{(j)},\mathbf{x}_{(k)}) = \frac{1}{n}\sum_{i=1}^{n}(x_{ij}-\widehat{\mathrm{x}}_{(j)})(x_{ik}-\widehat{\mathrm{x}}_{(k)}) = \frac{1}{n}\langle \mathbf{x}_{(j)}-\widehat{\mathrm{x}}_{(j)}\mathbb{1}_n, \mathbf{x}_{(k)}-\widehat{\mathrm{x}}_{(k)}\mathbb{1}_n\rangle$$

En effet, si (en moyenne !) les personnes avec un écart à la moyenne positif pour la variable "longueur des pieds" ont aussi un écart positif à la moyenne pour "longueur des doigts", on présente cela comme le fait que les variables ont tendance à varier ensemble (covariance positive). Si cette un écart est positif et l'autre négatif, la covariance est négative.

Pour raccourcir les notations, on écrit parfois : $\mathrm{Cov}(\mathbf{x}_{(j)},\mathbf{x}_{(k)}) = \mathrm{Cov}_{(j),(k)}$. La covariance d'une variable avec elle-même est sa variance : $\mathrm{Var}_{(j)} = \mathrm{Cov}_{(j),(j)}$.

Proposition 1.3.9

Quelques propriétés de la covariance : si λ et μ sont des nombres réels,

1. $\mathrm{Var}_{(j)} = \mathrm{Cov}_{(j),(j)}$;
2. $\mathrm{Cov}_{(i),(j)} = \mathrm{Cov}_{(j),(i)}$;
3. $\mathrm{Cov}(\lambda\boldsymbol{x}+\mu\boldsymbol{x}',\boldsymbol{x}'') = \lambda\mathrm{Cov}(\boldsymbol{x},\boldsymbol{x}'')+\mu\mathrm{Cov}(\boldsymbol{x}',\boldsymbol{x}'')$;
4. *En particulier,* $\mathrm{Var}(\lambda\boldsymbol{x}) = \lambda^2\mathrm{Var}(\boldsymbol{x})$;
5. $\mathrm{Cov}(\boldsymbol{x}+\lambda\mathbb{1}_n,\boldsymbol{x}') = \mathrm{Cov}(\boldsymbol{x},\boldsymbol{x}')$.

DÉMONSTRATION: Exercice (utiliser les définitions et les propriétés du produit scalaire). ■

Il y a aussi une formule plus simple pour calculer la covariance :

Théorème 1.3.10

$$\mathrm{Cov}(\boldsymbol{x}_{(j)},\boldsymbol{x}_{(k)}) = \frac{1}{n}\langle \boldsymbol{x}_{(j)},\boldsymbol{x}_{(k)}\rangle - \widehat{\mathrm{x}}_{(j)}\widehat{\mathrm{x}}_{(k)} = \left(\frac{1}{n}\sum_{i=1}^{n}x_{ij}x_{ik}\right) - \widehat{\mathrm{x}}_{(j)}\widehat{\mathrm{x}}_{(k)}.$$

DÉMONSTRATION: Exercice (il s'agit des mêmes arguments que pour la variance). ■

Exemple 1.3.11. Mettons-nous dans une situation artificielle où deux variables sont complètement reliées par une relation linéaire. Disons

$$\mathbf{x}_{(1)} = \begin{pmatrix} 5 \\ 6 \\ 7 \\ 8 \end{pmatrix} \text{ et } \mathbf{x}_{(2)} = 3\mathbf{x}_{(1)} + 2\mathbb{1}_4 = 3\begin{pmatrix} 5 \\ 6 \\ 7 \\ 8 \end{pmatrix} + 2\begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 16 \\ 19 \\ 22 \\ 25 \end{pmatrix}$$

Les moyennes sont $\widehat{\mathrm{x}}_{(1)} = \frac{1}{4}(5+6+7+8) = 6.5$ et [10] $\widehat{\mathrm{x}}_{(2)} = \frac{1}{4}(16+19+22+25) = 20.5$.

On peut ensuite calculer la covariance de ces deux variables à la main :

$$\begin{aligned} \mathrm{Cov}(\mathbf{x}_{(1)}, \mathbf{x}_{(2)}) &= \tfrac{1}{4}\big[(5-6.5)(16-20.5) + (6-6.5)(19-20.5) \\ &\qquad + (7-6.5)(22-20.5) + (8-6.5)(25-20.5)\big] \\ &= \tfrac{1}{4}\big[\tfrac{-3}{2}\cdot\tfrac{-9}{2} + \tfrac{-1}{2}\cdot\tfrac{-3}{2} + \tfrac{1}{2}\cdot\tfrac{3}{2} + \tfrac{3}{2}\cdot\tfrac{9}{2}\big] \\ &= \tfrac{30}{4} \qquad\qquad = 7.5 \end{aligned}$$

D'autre part, on peut vérifier que $\mathrm{Var}(\mathbf{x}_{(1)}) = \frac{1}{4}\big[\frac{9}{4} + \frac{1}{4} + \frac{1}{4} + \frac{9}{4}\big] = 2.5$.

En fait, il suffit de calculer $\mathrm{Var}(\mathbf{x}_{(1)})$ pour trouver $\mathrm{Cov}(\mathbf{x}_{(1)}, \mathbf{x}_{(2)})$ et $\mathrm{Var}(\mathbf{x}_{(2)})$. En utilisant les propriétés de la covariance, pour n'importe quels deux réels a et b, on a

$$\begin{aligned} & \mathrm{Cov}(\mathbf{x}, a\mathbf{x} + b\mathbb{1}_n) = \mathrm{Cov}(\mathbf{x}, a\mathbf{x}) = a\mathrm{Cov}(\mathbf{x}, \mathbf{x}) = a\mathrm{Var}(\mathbf{x}), \\ \text{et} \quad & \mathrm{Var}(a\mathbf{x} + b\mathbb{1}_n) = \mathrm{Var}(a\mathbf{x}) = a^2\mathrm{Var}(\mathbf{x}). \end{aligned}$$ ♣

La matrice des variances, de taille $p \times p$ est la matrice :

$$V := \frac{1}{n}(X - \hat{X})^{\mathsf{T}} \cdot (X - \hat{X}),$$

ou encore $(V)_{ij} := \mathrm{Cov}(\mathbf{x}_{(i)}, \mathbf{x}_{(j)})$. Pour insister sur la dépendance sur les données X, on écrit parfois aussi V_X ou $\mathrm{Cov}(X)$. Géométriquement, cette matrice permet de trouver l'inertie du nuage de point (par rapport à n'importe quel axe de rotation).

La matrice des variances est symétrique :$V^{\mathsf{T}} = V$, par la proposition 1.3.9.

Une quantité utile et important pour une matrice est sa trace (voir la définition 1.2.22). En général, si M est une matrice $k \times k$, $\mathrm{Tr}M = \sum_{i=1}^{k} (M)_{kk}$, *i.e.* c'est la somme des coefficients qui sont sur la diagonale.

Définition 1.3.12. La **variabilité totale** des données X est

$$\mathrm{Tr}V_X = \sum_{i=1}^{n} \mathrm{Var}(\mathbf{x}_{(i)}).$$ ★

Si on regarde les variables centrées réduites, la variabilité totale est égale au nombre de variables.

10. Par la propriété de la moyenne, voir proposition 1.3.2 : si $\mathbf{y} = a\mathbf{x} + c\mathbb{1}_n$, alors $\widehat{\mathrm{y}} = a\widehat{\mathrm{x}} + c$. Ainsi, il n'est pas nécessaire de faire le calcul ici car $\mathbf{x}_{(2)} = 3\mathbf{x}_{(1)} + \mathbb{1}_4$. Par conséquent, $\widehat{\mathrm{x}}_{(2)} = 3\widehat{\mathrm{x}}_{(1)} + 1 = 3 \cdot 6.5 + 1 = 20.5$.

1.3.iv Corrélation

La corrélation (de Pearson) est un indicateur d'une relation *linéaire* entre deux variables. La matrice des corrélations est définie soit à partir des variables centrées réduites Z

$$R := R_X := \tfrac{1}{n} Z^{\mathsf{T}} Z, \qquad \text{soit par} \qquad (R)_{ij} = \frac{\mathrm{Cov}_{(i),(j)}}{\sigma_{(i)}\sigma_{(j)}} = \frac{(V)_{ij}}{\sqrt{(V)_{ii}(V)_{jj}}}$$

Il est bon de souligner que $(R)_{ii} = 1$, puisque la formule donne alors $\frac{(V)_{ii}}{\sqrt{(V)_{ii}(V)_{ii}}} = 1$. Quelques avertissements habituels mais très importants sur la corrélation.

Remarque 1.3.13. La corrélation est un indicateur d'une relation, et n'a rien à voir avec la causalité. Par exemple, les variables "produit net des ventes de crème solaire" et "produit net de ventes de lunettes solaire" sont souvent très corrélées, particulièrement dans les stations balnéaires. Ce n'implique pourtant pas que l'achat de l'un provoque l'achat de l'autre. Dans ce cas particulier, il est assez facile de voir la relation : quand il y a du soleil, les gens ont tendance à avoir besoin de ces deux objets. ♠

1.3.v Relation linéaire et non-linéaire

Remarque 1.3.14. Le fait que la corrélation soit nulle n'indique absolument pas l'absence de relation. Par exemple, voici un tableau de donnée :

$$\begin{pmatrix} -3 & 9 \\ -2 & 4 \\ -1 & 1 \\ 0 & 0 \\ 1 & 1 \\ 2 & 4 \\ 3 & 9 \end{pmatrix}$$

Un calcul rapide donne que $\widehat{\mathrm{x}}_{(1)} = 0$ et $\widehat{\mathrm{x}}_{(2)} = 4$. Ce qui donne

$$V = \frac{1}{2}(X-\hat{X})^{\mathsf{T}}(X-\hat{X}) = \frac{1}{7}\begin{pmatrix} -3 & -2 & -1 & 0 & 1 & 2 & 3 \\ 5 & 0 & -3 & -4 & -3 & 0 & 5 \end{pmatrix}\begin{pmatrix} -3 & 5 \\ -2 & 0 \\ -1 & -3 \\ 0 & -4 \\ 1 & -3 \\ 2 & 0 \\ 3 & 5 \end{pmatrix} = \begin{pmatrix} 2 & 0 \\ 0 & 12 \end{pmatrix}$$

Ainsi, il est assez facile de voir que la corrélation est nulle. Cependant, il y a une relation assez flagrante : la seconde variable est la première variable mise au carré. ♠

Ce qui rend cette corrélation néanmoins utile est que si les variations sont faibles, une relation, même non-linéaire sera quand même détectée. Par exemple, supposons deux variables sont reliées

par $y = f(x)$. Si seules des petites variations de x sont faites, le principe même de la dérivée est de dire que $f(x+\varepsilon) \simeq f(x)+\varepsilon f'(x)$. Ainsi, sauf si l'on est près d'un point où la dérivée de la fonction est nulle (ce qui serait un coup de malchance assez improbable), on s'attend à ce qu'une corrélation linéaire apparaisse.

Voici un peu plus de détails sur ce qui se passe.

Exemple 1.3.15. Soit f une fonction (à valeur réelle) qui est dérivable. Être dérivable signifie qu'en chaque réel x, si on se fixe un erreur numérique δ pour tout nombre réel ε assez petit (le "assez petit" dépend de l'erreur numérique permise et de x), l'égalité $f(x+\varepsilon) = f(x)+\varepsilon f'(x)+err(\varepsilon)$ a lieu avec *err* une "erreur" qui satisfait $|err(\varepsilon)| < \varepsilon\delta$. Autrement dit, l'erreur est "négligeable" par rapport à la variation ε.

Soit $\mathbf{z} \in \mathbb{R}^n$ un vecteur tel que $\widehat{\mathrm{z}} = 0$ et soit $\mathbf{x} = c\mathbb{1}_n + \mathbf{z}$. On a $\widehat{\mathrm{x}} = c$. Soit $\mathbf{y}$ le vecteur définit par $y_i = f(x_i) = f(c+z_i)$.

Dans un premier temps, on suppose vraiment l'erreur comme étant complètement négligeable, *i.e.* on suppose que $f(x+\varepsilon) = f(x)+\varepsilon f'(x)$. Alors $\mathbf{y} = f(c)\mathbb{1}_n + f'(c)\mathbf{z}$. Ainsi, $\widehat{\mathrm{y}} = f(c)$ et

$$\mathrm{Cov}(\mathbf{x},\mathbf{y}) = \mathrm{Cov}(c\mathbb{1}_n+\mathbf{z}, f(c)\mathbb{1}_n + f'(c)\mathbf{z}) = \mathrm{Cov}(\mathbf{z}, f'(c)\mathbf{z}) = f'(c)\mathrm{Cov}(\mathbf{z},\mathbf{z}) = f'(c)\mathrm{Var}(\mathbf{z}).$$

Ainsi, même si *a priori* $\mathbf{x}$ et $\mathbf{y}$ sont reliés par une relation non-linéaire, celle-ci se traduit par une covariance (*i.e.* elle se détecte de manière linéaire), pourvu que $f'(c) \neq 0$ (ce qui est en pratique assez rare). ♣

Pour ce qui s'y intéresse vraiment, voici des détails supplémentaires en tenant les erreurs en compte.

Exemple 1.3.16. Continuons l'exemple précédent mais sans supposer que les erreurs sont complètement négligeables. Soit $\mathbf{r}$ le vecteur des erreurs, *i.e.* $r_i = err(z_i)$. Alors $\mathbf{y} = f(c)\mathbb{1}_n + f'(c)\mathbf{z}+\mathbf{r}$. Alors, par les propriétés de la moyenne (et comme $\widehat{\mathrm{z}} = 0$),

$$\widehat{\mathrm{y}} = f(c) + f'(c)0 + \widehat{\mathrm{r}} = f(c) + \tfrac{1}{n}\sum_{i=1}^{n} err(z_i)$$

Par l'inégalité du triangle, $|\sum_{i=1}^n err(z_i)| \leq \sum_{i=1}^n |err(z_i)| \leq \delta\sum_{i=1}^n |z_i|$. Puis on utilise une inégalité connue [11] : pour n'importe quels nombres $a_i \in \mathbb{R}$, $\sum_{i=1}^n |a_i| \leq \sqrt{n}(\sum a_i^2)^{1/2}$. De là, on obtient $|\sum_{i=1}^n err(z_i)| \leq \delta\sqrt{n}\sigma(\mathbf{z})$. Ainsi $\widehat{\mathrm{y}} \simeq f(c)$ avec une erreur [12] d'au plus $\delta\sqrt{\frac{\mathrm{Var}(\mathbf{z})}{n}}$.

On peut aussi calculer

$$\mathrm{Cov}(\mathbf{x},\mathbf{y}) = \mathrm{Cov}(\mathbf{z}+c\mathbb{1}_n, f(c)\mathbb{1}_n + f'(c)\mathbf{z}+\mathbf{r}) = \mathrm{Cov}(\mathbf{z}, f'(c)\mathbf{z}+\mathbf{r}) = f'(c)\mathrm{Cov}(\mathbf{z},\mathbf{z}) + \mathrm{Cov}(\mathbf{z},\mathbf{r})$$

D'un côté, $f'(c)\mathrm{Cov}(\mathbf{z},\mathbf{z}) = f'(c)\mathrm{Var}(\mathbf{z})$. Tandis que de l'autre

$$\mathrm{Cov}(\mathbf{z},\mathbf{r}) = \tfrac{1}{n}\sum_{i=1}^{n}\langle \mathbf{z}, \mathbf{r} - \widehat{\mathrm{r}}\mathbb{1}_n\rangle = \tfrac{1}{n}\sum_{i=1}^{n}\langle \mathbf{z},\mathbf{r}\rangle\tfrac{1}{n} - \sum_{i=1}^{n}\langle \mathbf{z}, \widehat{\mathrm{r}}\mathbb{1}_n\rangle = \tfrac{1}{n}\sum_{i=1}^{n} z_i err(z_i).$$

11. Cette inégalité est un cas particulier de l'inégalité de Cauchy-Schwartz, voir proposition 1.2.8.

12. Quelques principes généraux de probabilité (que le lecteur aurait dû voir dans un premier cours de statistiques) indiquent que si n est grand et que les mesures qui donnent la variable $\mathbf{z}$ satisfont des conditions d'indépendance faibles, $\mathrm{Var}(\mathbf{z})/n$ est de plus en plus petit lorsque n est grand.

En utilisant de nouveau que $|err(z_i)| \leq \delta z_i$ et l'inégalité du triangle, on a

$$|\mathrm{Cov}(\mathbf{z},\mathbf{r})| = \tfrac{1}{n}\big|\sum_{i=1}^{n} z_i err(z_i)\big| \leq \tfrac{1}{n}\sum_{i=1}^{n} |z_i||err(z_i)| \leq \tfrac{1}{n}\delta\sum_{i=1}^{n}|z_i|^2 = \delta\mathrm{Var}(\mathbf{z})$$

Ainsi, $\mathrm{Cov}(\mathbf{x},\mathbf{y})$ est égal à $f'(c)\mathrm{Var}(\mathbf{z})$ avec une erreur d'au plus $\delta\mathrm{Var}(\mathbf{z})$. En mettant toutes ces informations ensemble, on a

$$\widehat{\mathrm{y}} = f(c) \pm \delta\sqrt{\frac{\mathrm{Var}(\mathbf{z})}{n}} \qquad \text{et} \qquad \mathrm{Cov}(\mathbf{x},\mathbf{y}) = f'(c)\mathrm{Var}(\mathbf{z}) \pm \delta\mathrm{Var}(\mathbf{z})$$

où $\pm a$ signifie avec une erreur d'au plus a. ♣

Les arguments des exemples précédents ne sont valables que lorsqu'on se préoccupe de "petites variations". C'est assez naturel : dans un phénomène physique, il y a souvent plusieurs "régimes" différent.

L'exemple le plus simple serait d'étudier les comportements de l'eau selon la température (et la pression). Il vaut mieux rester dans des intervalles de température comme $200°C$ à $300°C$ ou $20°C$ à $60°C$ et mieux éviter les intervalles comme $-10°C$ à $110°C$ (ou des choses radicalement différentes se produisent)...

Lorsque le nombre de paramètres devient trop grand (*e.g.* en écologie ou en climatologie), cet avertissement est d'autant plus vrai que les systèmes qu'on observe peuvent très bien avoir des facteurs qui en amplifient d'autres. Cela est d'autant plus vrai, qu'on ne peut en général pas soupçonner à l'avance où se produiront des changements importants.

Finalement, c'est le bon moment pour rappeler que nous faisons de la statistique *descriptive*. Autrement dit, il s'agit surtout d'un manière d'explorer ces données, et non pas de faire tirer des conséquences avec un certain intervalle de confiance (c'est le propre de la statistique inférentielle).

1.4 La régression linéaire

Le but de cette section est de présenter rapidement la régression linéaire dans le cas de plusieurs variable (et en écriture matricielle).

Soit $\mathbf{y}^{\mathsf{T}} = (y_1, ..., y_n)$ un échantillon de la mesure d'une variable expliquée ou variable réponse [13] (en anglais, *response variable*) sur n individus (y_i est donc la mesure sur le $i^{\text{ème}}$ individu). Soit X la matrice donnée des variables explicatives [14] (en anglais, *explanatory variables*), *i.e.* une matrice $n \times p$ qui contient p autres mesures pour les mêmes n individus.

Faire une régression linéaire c'est tenter d'écrire $\mathbf{y}$ comme une combinaison linéaire des $\mathbf{x}$. Cependant, il est probable que ce ne soit pas possible de le faire : $\mathbf{y}$ est un vecteur de taille n, il y a autant de vecteur $\mathbf{x}_{(i)}$ que de variables et pour que l'analyse soit représentative il y a d'habitude beaucoup plus d'individus que de variables. Ainsi l'espace engendré par $\mathbf{x}_{(i)}$ (appelons le E) est de dimension beaucoup plus petite que l'espace dans lequel est $\mathbf{y}$, il y a donc toutes les chances pour que $\mathbf{y}$ ne soit pas dans E.

13. en écologie, on dit aussi variables espèces
14. toujours en écologie : variables environnementales.

Par conséquent, l'important est d'exprimer cette relation linéaire en minimisant l'erreur ou, autrement dit, de trouver le point $\tilde{\mathbf{y}}$ dans E qui est le plus proche de $\mathbf{y}$. On pose

$$\begin{array}{rl} \tilde{y}_i & = b_0 + \sum_{j=1}^p x_{ij} b_j, \qquad \forall i = 1, ..., n, \\ \text{ou, en notation matricielle, } \tilde{\mathbf{y}} & = b_0 \mathbb{1}_n + X\mathbf{b} \end{array} \tag{1.4.1}$$

où les b_i (ici $i \in \{0, 1, \ldots, p\}$) sont à déterminer de sorte que, si l'erreur sur chaque individu est $u_i = y_i - \tilde{y}_i = y_i - b_0 + \sum_{j=1}^p b_j x_{ij}$, alors $\sum_{i=1}^n u_i^2$ soit minimale. Cette somme de carré n'est rien d'autre que la norme de $\mathbf{u} = (u_1, \ldots, u_n)^\mathsf{T}$.

Pour garder une notation complètement matricielle de (1.4.1), soit $X' = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \ldots & x_{np} \end{pmatrix}$ et $\mathbf{b}' = \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_p \end{pmatrix}$. Alors l'équation (1.4.1) se réécrit

$$\tilde{\mathbf{y}} = X'\mathbf{b}'.$$

Proposition 1.4.2

Supposons que X est centrée [15] *et que* $\mathrm{Det}(X'^\mathsf{T} X') \neq 0$ [16], *alors* $\sum_{i=1}^n u_i^2 = \|\mathbf{y} - X'\boldsymbol{b}'\|^2$ *est minimal lorsque*

$$\boldsymbol{b}' = (X'^\mathsf{T} X')^{-1} X'^\mathsf{T} \mathbf{y},$$

ou, de manière équivalente : $\boldsymbol{b} = (X^\mathsf{T} X)^{-1} X^\mathsf{T} \mathbf{y}$ *et* $b_0 = \widehat{\mathrm{y}}$. *Ainsi,* $\tilde{\mathbf{y}} = (X'^\mathsf{T} X')^{-1} X'^\mathsf{T} \mathbf{y}$.

En particulier, lorsque $\mathbf{y}$ est centrée [17] alors $b_0 = 0$ et

$$\mathbf{b} = (X^\mathsf{T} X)^{-1} X^\mathsf{T} \mathbf{y} \qquad \text{et} \qquad \tilde{\mathbf{y}} = X(X^\mathsf{T} X)^{-1} X^\mathsf{T} \mathbf{y}.$$

DÉMONSTRATION: L'idée est de voir $\mathbf{u}$ comme une fonction des coefficients b_i. Si cette fonction est minimale, alors sa dérivée par rapport à b_i (*c'àd.* en considérant tous les autres termes constants, y compris les b_j pour $j \neq i$) doit être nulle.

En dérivant $\|\mathbf{y} - X'\mathbf{b}'\|^2 = \langle \mathbf{y} - X'\mathbf{b}', \mathbf{y} - X'\mathbf{b}' \rangle$ par rapport à b_i, il apparaît

$$\begin{array}{ll} \text{si } i \in \{1, \ldots, p\} : & 2\langle \mathbf{y} - X'\mathbf{b}', X'\underline{\mathrm{e}}_i \rangle = 2\langle \mathbf{y} - X'\mathbf{b}', \mathbf{x}_{(i)} \rangle \\ \text{si } i = 0 & 2\langle \mathbf{y} - X'\mathbf{b}', \mathbb{1}_n \rangle \end{array}$$

Les dérivées doivent toutes s'annuler en un minimum de cette fonction. Or, si ces dérivées sont nulles, on a :

$$\langle \mathbf{y}, \mathbf{x}_{(i)} \rangle = \langle X'\mathbf{b}', \mathbf{x}_{(i)} \rangle \quad (\forall i \in \{1, \ldots, p\}) \qquad \text{et} \qquad \langle \mathbf{y}, \mathbb{1}_n \rangle = \langle X'\mathbf{b}', \mathbb{1}_n \rangle$$

14. *i.e.* $\hat{X} = 0$, *i.e.* pour tout $i \in \{1, \ldots, p\}$ on a $\widehat{\mathrm{x}}_{(i)} = 0$, *i.e.* $\mathbb{1}_n^\mathsf{T} X = 0$

16. de sorte que la matrice $(X'^\mathsf{T} X)^{-1}$ existe. Cette condition est quasiment toujours vérifiée en pratique. Elle est mise en défaut seulement si les vecteurs $\{\mathbf{x}_{(i)}\}_{i=1}^p$ ne sont pas linéairement indépendants. En mots, cela voudrait dire que certaines des variables $\mathbf{x}_{(j)}$, disons pour certains indices $j \in J$, expliquent parfaitement une des variable $\mathbf{x}_{(i)}$, où $i \notin J$.

17. *i.e.* de moyenne nulle, *i.e.* $\widehat{\mathrm{y}} = 0$, *i.e.* $\mathbb{1}_n^\mathsf{T} \mathbf{y} = 0$.

Puis en passant à l'écriture matricielle du produit scalaire :

$$\mathbf{x}_{(i)}^{\mathsf{T}}\mathbf{y} = \mathbf{x}_{(i)}^{\mathsf{T}}X'\mathbf{b}' \quad (\forall i \in \{1,\dots,p\}) \qquad \text{et} \qquad \mathbb{1}_n^{\mathsf{T}}\mathbf{y} = \mathbb{1}_n^{\mathsf{T}}X'\mathbf{b}'$$

[Au passage, on peut remarque que la dernière équation implique que $n\widehat{\mathrm{y}} = nb_0$, *c'àd.* que $b_0 = \widehat{\mathrm{y}}$.] Ces équations s'écrivent ensemble comme une équation matricielle :

$$X'^{\mathsf{T}}\mathbf{y} = X'^{\mathsf{T}}X'\mathbf{b}' \qquad \Rightarrow \qquad \mathbf{b}' = (X'^{\mathsf{T}}X')^{-1}X'^{\mathsf{T}}\mathbf{y}.$$

ou $\mathbf{b} = (X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}\mathbf{y}$ et $b_0 = \widehat{\mathrm{y}}$. ■

Géométriquement, $X^{\mathsf{T}}\mathbf{y}$ est le vecteur (de taille p) qui donne les différentes projections de $\mathbf{y}$ sur les $\mathbf{x}_{(i)}$. La matrice $(X^{\mathsf{T}}X)^{-1}$ (qui est en fait nV_X, puisque $\hat{X}$ est supposée nulle) est présente pour compenser le fait que les variables $\mathbf{x}_{(i)}$ peuvent être corrélées ; le résultat $\mathbf{b}$ de tout ce produit donne les coordonnées de $\mathbf{y}$ dans la base *non-orthonormée* des $\mathbf{x}_{(i)}$.

Une fois la régression linéaire accomplie, la variance de "l'erreur" $\mathbf{u}$ provenant de cette tentative d'approximation est appelée variance résiduelle.

Lemme 1.4.3

(Voir [8, p.388]) La variance résiduelle de la régression linéaire $\boldsymbol{y} = b_0\mathbb{1}_n + X\boldsymbol{b} + \boldsymbol{u}$ est donnée par

$$\mathrm{Var}(\boldsymbol{u}) = \mathrm{Var}(\boldsymbol{y}) - \frac{\mathrm{Cov}(\boldsymbol{y}, X\boldsymbol{b})^2}{\mathrm{Var}(X\boldsymbol{b})}.$$

DÉMONSTRATION: C'est un calcul direct (mais un peu long) depuis la formule $\mathrm{Var}(\mathbf{u}) = \frac{1}{n}\|\mathbf{u} - \widehat{\mathrm{u}}\|^2$. Pour simplifier la situation, supposons que $\mathbf{y}$ soit aussi centré. Ainsi, par la proposition 1.4.2, $\tilde{\mathbf{y}} = X(X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}\mathbf{y}$. Comme $\mathbb{1}_n^{\mathsf{T}}X = 0$ et $\mathbb{1}_n^{\mathsf{T}}\mathbf{y} = 0$, $\mathbb{1}_n^{\mathsf{T}}(\mathbf{y} - \tilde{\mathbf{y}}) = 0$, ou, autrement dit, les erreurs sont de moyenne 0. Ainsi, la variance est

$$\begin{aligned}\|\mathbf{u}\|^2 &= \langle \mathbf{y} - \tilde{\mathbf{y}}, \mathbf{y} - \tilde{\mathbf{y}}\rangle &&= \langle \mathbf{y}, \mathbf{y}\rangle - \langle \tilde{\mathbf{y}}, \mathbf{y}\rangle - \langle \mathbf{y}, \tilde{\mathbf{y}}\rangle + \langle \tilde{\mathbf{y}}, \tilde{\mathbf{y}}\rangle \\ &&&= \mathrm{Var}(\mathbf{y}) - 2\langle \tilde{\mathbf{y}}, \mathbf{y}\rangle + \langle \tilde{\mathbf{y}}, \tilde{\mathbf{y}}\rangle\end{aligned}$$

Ensuite, il suffit de substituer $\tilde{\mathbf{y}} = X\mathbf{b} = X(X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}\mathbf{y}$. Une étape intermédiaire consiste à remarquer que $\mathrm{Cov}(\mathbf{y}, \tilde{\mathbf{y}}) = \langle \tilde{\mathbf{y}}, \mathbf{y}\rangle = \langle \tilde{\mathbf{y}}, \tilde{\mathbf{y}}\rangle = \mathrm{Var}(\tilde{\mathbf{y}})$: (on utilisera que $X^{\mathsf{T}}X$ est symétrique [18] et donc que $(X^{\mathsf{T}}X)^{-1}$ est symétrique)

$$\begin{aligned}\langle \tilde{\mathbf{y}}, \tilde{\mathbf{y}}\rangle &= \left(X(X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}\mathbf{y}\right)^{\mathsf{T}} X(X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}\mathbf{y} &&= \mathbf{y}^{\mathsf{T}}X(X^{\mathsf{T}}X)^{-1\mathsf{T}}X^{\mathsf{T}}X(X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}\mathbf{y} \\ &= \mathbf{y}^{\mathsf{T}}X(X^{\mathsf{T}}X)^{-1}(X^{\mathsf{T}}X)(X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}\mathbf{y} &&= \mathbf{y}^{\mathsf{T}}X(X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}\mathbf{y} \\ &= \langle \mathbf{y}, \tilde{\mathbf{y}}\rangle\end{aligned}$$

Le premier calcul se poursuit alors comme suit :

$$\begin{aligned}\|\mathbf{u}\|^2 &= \mathrm{Var}(\mathbf{y}) - 2\langle \tilde{\mathbf{y}}, \mathbf{y}\rangle + \langle \tilde{\mathbf{y}}, \tilde{\mathbf{y}}\rangle &&= \mathrm{Var}(\mathbf{y}) - \langle \tilde{\mathbf{y}}, \mathbf{y}\rangle \\ &= \mathrm{Var}(\mathbf{y}) - \langle \tilde{\mathbf{y}}, \mathbf{y}\rangle \frac{\langle \tilde{\mathbf{y}}, \mathbf{y}\rangle}{\langle \tilde{\mathbf{y}}, \tilde{\mathbf{y}}\rangle} &&= \mathrm{Var}(\mathbf{y}) - \frac{\langle \tilde{\mathbf{y}}, \mathbf{y}\rangle^2}{\langle \tilde{\mathbf{y}}, \tilde{\mathbf{y}}\rangle} \quad = \mathrm{Var}(\mathbf{y}) - \frac{\mathrm{Cov}(\tilde{\mathbf{y}}, \mathbf{y})^2}{\mathrm{Var}(\tilde{\mathbf{y}})}\end{aligned}$$

En se souvenant que $\tilde{\mathbf{y}} = X\mathbf{b}$, le résultat est obtenu. ■

18. une matrice est symétrique si $M^{\mathsf{T}} = M$

Au passage, si l'expression $\text{Var}(\mathbf{u}) = \text{Var}(\mathbf{y}) - \text{Cov}(\mathbf{y}, X\mathbf{b})$ n'est pas donnée dans l'énoncé c'est qu'elle n'est valable que si les données sont centrées tandis que l'expression écrite dans le lemme est valable en général.

Chapitre 2

L'analyse en composantes principales et ses dérivés

L'objectif des méthodes de réduction est de réduire le nombre de variable de manière à y voir plus clair dans la masse de données qui se présentent à nous. L'objectif est de le faire en perdant le moins d'information possible, *c'àd.* le moins de variabilité totale possible. Cela peut *a priori* paraître impossible : si par exemple on regarde des variables centrées réduites, enlever une variable réduit la variabilité de 1.

2.1 L'analyse en composantes principales (ACP)

ou, plus communément, *PCA* en anglais (pour *Principal Component Analysis*).

2.1.i Les vecteur propres

La méthode de l'ACP est de faire un changement de base de sorte à y voir plus clair. On passera donc de la matrice des données X à une nouvelle matrice des données $Y = XQ$, où Q est une matrice $p \times p$ de changement de base "bien choisie". Pour être plus précis sur comment choisir cette matrice Q, on voudrait que :

- la variance des variables aille en ordre décroisant $\mathrm{Var}(\mathbf{y}_{(1)}) \geq \mathrm{Var}(\mathbf{y}_{(2)}) \geq \ldots \geq \mathrm{Var}(\mathbf{y}_{(p)})$;
- les variables $\mathbf{y}_{(i)}$ ne sont plus corrélées, *i.e.* la matrice des variance pour les données transformées Y est diagonale.

De cette façon, on voit que c'est la première "nouvelle variable" $\mathbf{y}_{(1)}$ qui "contient le plus d'information" (la variabilité est souvent interprétée comme la partie des données qui comporte de l'information). La seconde condition est très importante : il serait très facile de faire de nouvelles variables de variance très élevée en répétant simplement $\mathbf{y}_{(1)}$ plusieurs fois. La condition de non-corrélation est là pour s'assure que toute la variabilité de la seconde $\mathbf{y}_{(2)}$ est indépendante de celle de la première.

Ainsi, on a des nouvelles variables $\mathbf{y}_{(i)}$ qui portent chacune de l'information (les premières en portent le plus) et, si on considère deux variables distinctes, l'information qu'elles portent est

complémentaire.

Une fois ceci fait (ce n'est, *a priori*, pas du tout clair que c'est possible !), on gardera seulement les 2 ou 3 premières "nouvelles" variables, de sorte que tout se représente naturellement dans le plan (ou l'espace trois dimensionnel). Ces variable étant celles qui contiennent le plus de variance possible, on s'assure qu'en ne regardant que ce qui se passe dans ces variables, la perte d'information est aussi petite que possible.

Rappel : (voir les propositions 1.2.12 et 1.2.20) Soit $\underline{x}$ un vecteur ligne de longueur p (représentant la position d'un des individus dans l'espace des variables), et $\{\mathbf{q}_{(1)}, \mathbf{q}_{(2)}, \ldots, \mathbf{q}_{(p)}\}$ une base orthonormée de $\mathbb{R}^p$ (écrite comme des colonnes). Soit $\underline{y}$ le vecteur qui donne les coordonnées de $\underline{x}$ dans la base des $\{\mathbf{q}_{(j)}^\mathsf{T}\}$: y_i dit à quel point $\underline{x}$ se déplace dans la direction $\mathbf{q}_{(i)}^\mathsf{T}$. Une autre manière d'écrire ceci est $\underline{x} = \sum_{i=1}^{n} y_i \mathbf{q}_{(i)}^\mathsf{T}$. Alors

$$\underline{y} = \underline{x}Q \qquad \text{où} \qquad Q = \begin{pmatrix} | & | & & | \\ \mathbf{q}_{(1)} & \mathbf{q}_{(2)} & \cdots & \mathbf{q}_{(p)} \\ | & | & & | \end{pmatrix}.$$

(Autrement dit Q est la matrice dont la $i^{\text{ème}}$ colonne est $\mathbf{q}_{(i)}$.) Ainsi, dans ces nouvelles coordonnées, les données s'écrivent $Y = XQ$. Ainsi,

$$\hat{Y} = \mathbb{1}_n(\tfrac{1}{n}\mathbb{1}_n^\mathsf{T} Y) = \tfrac{1}{n}\mathbb{1}_n\mathbb{1}_n^\mathsf{T} XQ = \hat{X}Q.$$

En fait, on choisit toujours les nouvelles données Y de sorte à ce qu'elles soient des données centrées (*i.e.* de moyenne nulle). Pour ce faire il suffit de poser :

$$Y = (X - \hat{X})Q.$$

De toute façon, ce qui nous intéresse c'est la variance et,

$$V_Y = \tfrac{1}{n}(Y-\hat{Y})^\mathsf{T}(Y-\hat{Y}) = \tfrac{1}{n}(XQ-\hat{X}Q)^\mathsf{T}(XQ-\hat{X}Q) = \tfrac{1}{n}Q^\mathsf{T}(X-\hat{X})^\mathsf{T}(X-\hat{X})Q = Q^\mathsf{T}V_X Q.$$

Ainsi, poser $Y = XQ$ ou $Y = (X-\hat{X})Q$ donne la même matrice des variances V_Y. Enfin, les conditions de l'ACP, se réécrivent en terme des propriétés des vecteurs $\mathbf{q}_{(j)}$ grâce à :

$$\begin{aligned} \operatorname{Var}(\mathbf{y}_{(j)}) &= \mathbf{q}_{(j)}^t V_X \mathbf{q}_{(j)} \\ \operatorname{Cov}(\mathbf{y}_{(j)}, \mathbf{y}_{(i)}) &= \mathbf{q}_{(j)}^t V_X \mathbf{q}_{(i)} \end{aligned}$$

Encore une autre façon de présenter ces deux conditions est de dire que la matrice V_Y est une matrice diagonale (et les entrées de la diagonale sont en ordre décroissant). Les conditions de l'ACP (vu sur les vecteurs $\mathbf{q}_{(j)}$) sont que

$$\begin{aligned} \mathbf{q}_{(j)}^t V_X \mathbf{q}_{(j)} &= \lambda_j \text{ sont des nombres en ordre décroissant} \\ \text{et, si } i \neq j, \quad \mathbf{q}_{(j)}^t V_X \mathbf{q}_{(i)} &= 0 \end{aligned}$$

Le problème de l'ACP est ainsi de trouver la matrice Q qui satisfait aux conditions ci-haut. Un argument théorique (présenté en section 2.A) permet de dire que les nouvelles variables s'expriment comme des vecteurs propres de la matrice. Ici nous tenterons seulement d'introduire ces vecteurs propres, de montrer pourquoi ils résolvent le problème, puis d'expliquer comment ils sont calculés.

Définition 2.1.1. Soit V une matrice symétrique (*i.e.* $V = V^\mathsf{T}$) de taille $p \times p$. Un **vecteur propre** (en anglais, *eigenvector*) de V (de longueur p) est un vecteur (colonne) $\mathbf{q}$ tel que pour un certain nombre réel λ, $V\mathbf{q} = \lambda\mathbf{q}$ et $\mathbf{q}$ n'est pas le vecteur trivial $\begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}$. Le nombre λ est appelé la **valeur propre** (en anglais, *eigenvalue*) de $\mathbf{q}$. ★

Ces vecteurs propres sont exactement les vecteurs qu'il nous faut pour le changement de variable. Un rappel de vocabulaire n'est pas de trop : l'ensemble de vecteurs $\{\mathbf{w}_{(1)}, \ldots, \mathbf{w}_{(k)}\}$ est dit **linéairement indépendant**, si aucun d'entre eux ne s'exprime comme une combinaison linéaire des autres.

Supposons que V a n vecteurs propres $\mathbf{w}_{(1)}$, $\mathbf{w}_{(2)}$, ... , $\mathbf{w}_{(n)}$. Alors :

— si $\mathbf{w}_{(i)}$ et $\mathbf{w}_{(j)}$ ont une valeur propre distincte $\lambda_i \neq \lambda_j$ alors

$$\mathbf{w}_{(i)}^\mathsf{T} V \mathbf{w}_{(j)} = \lambda_j \langle \mathbf{w}_{(i)}, \mathbf{w}_{(j)} \rangle,$$

puisque $\mathbf{w}_{(j)}$ est un vecteur propre de valeur propre λ_j. D'autre part, $V = V^\mathsf{T}$ ainsi

$$\mathbf{w}_{(i)}^\mathsf{T} V \mathbf{w}_{(j)} = \mathbf{w}_{(i)}^\mathsf{T} V^\mathsf{T} \mathbf{w}_{(j)} = (V\mathbf{w}_{(i)})^\mathsf{T} \mathbf{w}_{(j)} = \lambda_i \langle \mathbf{w}_{(i)}, \mathbf{w}_{(j)} \rangle.$$

Donc $\lambda_j \langle \mathbf{w}_{(i)}, \mathbf{w}_{(j)} \rangle = \lambda_i \langle \mathbf{w}_{(i)}, \mathbf{w}_{(j)} \rangle$ où $\lambda_i \neq \lambda_j$. Par conséquent, $\langle \mathbf{w}_{(i)}, \mathbf{w}_{(j)} \rangle = 0$.

— si $\mathbf{w}_{(i)}$ et $\mathbf{w}_{(j)}$ ont la même valeur propre [1] (et sont linéairement indépendants), il est possible de les recombiner de sorte qu'ils soient orthogonaux : prendre $\mathbf{w}_{(i)}$ et comme second vecteur

$$\mathbf{w}' = \mathbf{w}_{(j)} - \frac{\langle \mathbf{w}_{(j)}, \mathbf{w}_{(i)} \rangle}{\|\mathbf{w}_{(i)}\|^2} \mathbf{w}_{(i)}.$$

En effet ce dernier satisfait $\langle \mathbf{w}', \mathbf{w}_{(i)} \rangle = 0$. Cette idée peut être aussi appliquée si le nombre de vecteurs est plus grand (et toujours linéairement indépendants) ; cela s'appelle le processus d'orthonormalisation de Gram-Schmidt.

— Multiplier par un réel un vecteur propre, donne toujours un vecteur propre (de même valeur propre) :

$$V(r\mathbf{w}) = rV\mathbf{w} = r\lambda\mathbf{w} = \lambda(r\mathbf{w}).$$

Ainsi, si ces vecteurs propres existent, il est possible de supposer qu'ils forment une base orthonormée. D'où

Proposition 2.1.2

Si V possèdent n vecteurs propres et qu'ils sont choisis de sorte à faire une base orthonormée, alors ces vecteurs sont la solution du problème de l'ACP.

DÉMONSTRATION: Supposons que $\{\mathbf{q}_{(1)}, \ldots, \mathbf{q}_{(p)}\}$ est une base orthonormée formée de vecteurs propres de V_X, soit Q la matrice de changement de base associée, alors $V_Y = Q^\mathsf{T} V_X Q$ est une matrice diagonale. En effet :

$$(V_Y)_{ij} = \mathbf{e}_{(i)}^\mathsf{T} V_Y \mathbf{e}_{(j)} = \mathbf{e}_{(i)}^\mathsf{T} Q^\mathsf{T} V_X Q \mathbf{e}_{(j)} = (Q\mathbf{e}_{(i)})^\mathsf{T} V_X \mathbf{q}_{(j)} = \mathbf{q}_{(i)}^\mathsf{T} V_X \mathbf{q}_{(j)} = \lambda_j \langle \mathbf{q}_{(i)}, \mathbf{q}_{(j)} \rangle = \begin{cases} 0 & \text{si } i \neq j \\ \lambda_j & \text{si } i = j \end{cases}$$

1. Dans la pratique, il est extrêmement rare que ceci se produise. Le lecteur peut sauter ce paragraphe en première lecture.

Autrement dit, V_Y est une matrice diagonale. ■

Quitte à réordonner les vecteurs propres de la valeur propre la plus grande λ_1 à la plus petite λ_p, on a ramené le problème à celui de trouver les vecteurs propres.

Remarque 2.1.3. Les matrice des variance ont toujours des valeurs propres ≥ 0 (en terme savant, elles sont semi-définies positives). Ceci est assez facile à voir : soit $A = \frac{1}{\sqrt{n}}(X - \hat{X})$, alors $V_X = A^\mathsf{T} A$. Si $\mathbf{w}$ est un vecteur (colonne de taille p) quelconque, alors

$$\mathbf{w}^\mathsf{T} V_X \mathbf{w} = \mathbf{w}^\mathsf{T} A^\mathsf{T} A \mathbf{w} = (A\mathbf{w})^\mathsf{T} (A\mathbf{w}) = \|A\mathbf{w}\|^2 \geq 0.$$

Si de plus, c'est un vecteur propre de valeur propre λ : $\mathbf{w}^\mathsf{T} V_X \mathbf{w} = \lambda \mathbf{w}^\mathsf{T} \mathbf{w} = \lambda \|\mathbf{w}\|^2$. Puisque $\|\mathbf{w}\|^2 \neq 0$ (*i.e.* $\mathbf{w}$ n'est pas le vecteur dont toutes les coordonnées sont 0), $\lambda = \|A\mathbf{w}\|^2 / \|\mathbf{w}\|^2 \geq 0$.

(Il est possible que $A\mathbf{w} = \mathbf{0}$ même si $\mathbf{w} \neq \mathbf{0}$.) ♠

Le calcul des vecteurs propres est malgré tout simple. En effet, si

$$V\mathbf{q} = \lambda \mathbf{q} \Leftrightarrow V\mathbf{q} - \lambda \mathrm{Id}_p \mathbf{q} = \mathbf{0} \Leftrightarrow (V - \lambda \mathrm{Id}_p)\mathbf{q} = \mathbf{0}.$$

Or,

Proposition 2.1.4

Une matrice M de taille $p \times p$ possède un vecteur $\mathbf{q}$ tel que $M\mathbf{q} = \mathbf{0}$ si et seulement si $\mathrm{Det} M = 0$ *où* Det *est le déterminant de M.*

Le déterminant est facile à calculer pour un humain lorsque la taille de la matrice est petite, plus important, elle est facile à calculer pour un ordinateur même lorsque la taille de la matrice est petite. On donnera plus de détail sur ce déterminant dans les calculs.

Corollaire 2.1.5

Les valeurs propres sont les solutions du polynôme $P(\lambda) = \mathrm{Det}(V - \lambda \mathrm{Id}_p)$.

Or un polynôme n'est pas une équation trop vilaine à résoudre. De plus, une fois la valeur propre connue, le vecteur propre $\mathbf{q}$ est une des solutions (non-triviale) de l'équation

$$V\mathbf{q} = \lambda \mathbf{q}$$

Remarque 2.1.6. Si cette équation n'a pas de solution autre que $\mathbf{0}$, c'est qu'on a fait une erreur dans le calcul des valeurs propres ! ♠

Exemple 2.1.7. Cet exemple est pour montrer que le calcul des vecteurs et valeurs propres n'est pas aussi compliqué qu'on pourrait le croire. Il n'est pas du tout réaliste (puisqu'il correspondrai au cas où on aurait seulement deux variables, et ainsi, aucun besoin d'en réduire le nombre). Supposons que la matrice est $V = \left(\begin{smallmatrix} 9 & 12 \\ 12 & 19 \end{smallmatrix}\right)$, alors

$$\mathrm{Det}(V - \lambda \mathrm{Id}) = \mathrm{Det} \begin{pmatrix} 9-\lambda & 12 \\ 12 & 19-\lambda \end{pmatrix} = (9-\lambda)(19-\lambda) - 12^2 = \lambda^2 - 28\lambda + 27.$$

D'où

$$\lambda = 14 \pm \tfrac{1}{2}\sqrt{28^2 - 4 \cdot 27} = 14 \pm \tfrac{1}{2}\sqrt{784 - 108} = 14 \pm 13.$$

Autrement dit, il y a deux valeurs propres : $\lambda_1 = 27$et $\lambda_2 = 1$. Maintenant, les vecteurs propres... celui de $\lambda_1 = 27$ est une solution de

$$\begin{pmatrix} 0 \\ 0 \end{pmatrix} = (V - 27\mathrm{Id}) \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} = \begin{pmatrix} -18 & 12 \\ 12 & -8 \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} = \begin{pmatrix} -18w_1 + 12w_2 \\ 12w_1 - 8w_2 \end{pmatrix}.$$

Il y a (superficiellement) deux équations : $-18w_1 + 12w_2 = 0$ et $12w_1 - 8w_2 = 0$. Après division de la première par -6 et de la seconde par 4, les deux équations ne sont en faite qu'un seule[2] : $3w_1 - 2w_2 = 0$. Ainsi $\binom{2}{3}$ est un vecteur propre. Afin qu'il soit de norme 1, il vaut mieux le diviser par sa norme, d'où $\mathbf{q}_{(1)} = \binom{2/\sqrt{13}}{3/\sqrt{13}}$.

Pour le second vecteur propre (de valeur propre 1), c'est une solution de

$$\begin{pmatrix} 0 \\ 0 \end{pmatrix} = (V - \mathrm{Id}) \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} = \begin{pmatrix} 8 & 12 \\ 12 & 18 \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} = \begin{pmatrix} 8w_1 + 12w_2 \\ 12w_1 + 18w_2 \end{pmatrix}.$$

Les deux équations se ramènent cette fois-ci à $2w_1 + 3w_2 = 0$. Ainsi $\binom{3}{-2}$ est un vecteur propre. La multiplication par $\frac{1}{\sqrt{13}}$ donnera un vecteur de norme 1 : $\mathbf{q}_{(2)} = \binom{3/\sqrt{13}}{-2/\sqrt{13}}$. ♣

2.1.ii Interprétation et représentation

L'interprétation géométrique du procédé est simple. Pour se faire, plaçons-nous dans le cas où il n'y a que trois variables. Nos données peuvent être représentées par un ensemble de points dans l'espace (trois dimensionnel), chaque individu a la position $\underline{x}_i$. Le vecteur moyenne $\underline{m}$ donne le centre de masse de ce nuage de points. La matrice des variance V donne son inertie : soit E l'image par V de la boule unité (l'ensemble des vecteurs de norme ≤ 1), *i.e.*

$$E = \{\underline{v} \in \mathbb{R}^p \mid \exists \underline{w} \in \mathbb{R}^p \text{ tel que } \|\underline{w}\| \leq 1 \text{ et } \underline{w}V = \underline{v}\}.$$

Alors E (translaté par $\underline{m}$) est un ellipsoïde qui imite le mieux possible la forme du nuage de points.

Cet ellipsoïde est aussi déterminé par ses semi-axes. Or les trois semi-axes sont exactement les trois directions des vecteurs propres et leur longueur est la valeur propre associée.

Faire une ACP, c'est donc choisir les directions (d'habitude deux, parfois un peu plus) dans lesquelles notre nuage de point présente la plus grande variabilité. Ces deux vecteurs engendrent un plan, et ensuite on projette toutes les données sur ce plan de forme à obtenir quelque chose de lisible, sans avoir perdu trop d'information. Le concept clef est :

L'ACP est une projection !
(sur un espace de dimension inférieure, engendré par les $\mathbf{q}_{(i)}$)

La perte d'information est donc automatique, fatale et inévitable. Pour rendre cette "perte d'information" plus précise, rappelons un résultat important :

2. Il est crucial d'avoir moins d'équations que d'inconnues : sinon la seule réponse serait le vecteur trivial. En fait, si, à cette étape une des équations ne disparaît pas, c'est qu'on a fait une erreur.

Proposition 2.1.8

Soit V et U des matrices de taille $p \times p$. S'il existe une matrice U^{-1} telle que $U^{-1}U = \mathrm{Id}$, alors $\mathrm{Tr}(U^{-1}VU) = \mathrm{Tr}V$.

DÉMONSTRATION: Par le lemme 1.2.23, $\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$ si A et B sont deux matrice de taille $p \times p$. On applique ceci avec $A = U^{-1}$ et $B = VU$:

$$\mathrm{Tr}(U^{-1}VU) = \mathrm{Tr}(AB) = \mathrm{Tr}(BA) = \mathrm{Tr}(VUU^{-1}) = \mathrm{Tr}(V\mathrm{Id}_p) = \mathrm{Tr}V,$$

car $UU^{-1} = \mathrm{Id}_p$ et, pour n'importe quelle matrice de taille $p \times p$, $M\mathrm{Id}_p = M$. ■

Corollaire 2.1.9

Soit V_X la matrice des variances et Q la matrice des vecteurs propres. Soit $V_Y = Q^{\mathsf{T}}V_XQ$, i.e. *la matrice des variances des données exprimées dans la nouvelle base $Y = XQ$. Alors la variabilité totale de Y est égale à celle de X : $\mathrm{Tr}V_Y = \mathrm{Tr}V_X$. Autrement dit,*

$$\mathrm{Tr}V_X = \sum_{i=1}^{p} \lambda_i$$

où $\lambda_i = (V_Y)_{ii}$ est la $i^{\text{ème}}$ valeur propre de V_X.

En particulier, si on ne garde que les k premières [3] variables des données Y, la variabilité totale restante est $\lambda_1 + \ldots + \lambda_k$. La fraction de la variabilité que ceci représente est

$$\frac{\lambda_1 + \ldots + \lambda_k}{\mathrm{Tr}V_X}.$$

Lors d'une ACP, il est important de vérifier que ce nombre est assez proche de 1. Voir la sous-section 2.1.iv pour plus de détails.

L'interprétation d'une combinaison linéaire peut être difficile. En effet, le sens des variables $\mathbf{y}_{(i)}$ n'est pas forcément clair à interpréter. La meilleure façon de décrire ceci en termes communs est de faire allusions aux odeurs [4].

Supposons que l'on cherche à mesurer l'odeur d'un plat/un animal/une personne. On utilisera alors des appareils pour mesure la concentration de divers composés chimiques que dégage l'individu dans l'air (des "olfactomètres"). Ces appareils nous donne un résultat généralement peu parlant : la $i^{\text{ème}}$ variables originale sera la concentration du composé numéro i. Or une odeur "naturelle" n'est jamais un pur concentré d'une seule composante [5]. Ainsi, si on étudie quels odeurs sont attirante (*e.g.* pour les chimpanzés mâles, ou encore pour un charognard affamé), on s'attend à ce qu'une "combinaison" des divers composés apparaisse comme attirante, *i.e.* une combinaison linéaire des variables initiales.

Un peu de terminologie.

3. Dans la vaste majorité des cas $k = 2$ ou 3.

4. La même analogie pourrait être faite avec la musique (ou les sons).

5. Par exemple, le H_2S est la composante principale de l'odeur des "oeufs pourris" mais ce n'est probablement pas la seule. Idem pour "ça sent l'alcool" ou "ça sent la fondue", il y a certainement des composés dominant, mais il s'agit néanmoins d'un mélange.

— Les **composantes principales** (anglais : *principal component*) parfois notées $PC_k := \mathbf{y}_{(k)}$. Rappelons que ces composantes principales sont non-corrélées et $\mathrm{Var}(\mathbf{y}_{(k)}) = \lambda_k$, où $k \in \{1,\dots,p\}$.

— **score/coordonnée des individus/site/échantillons** (anglais : *component/sample/site scores*) pour le $i^{\text{ème}}$ individu, ce sont ses coordonnées centrées [6] dans la base des vecteurs propres :

$$\underline{\mathrm{y}}_i = (\underline{\mathrm{x}}_i - \widehat{\mathrm{x}}_{(i)})Q$$

— **scores/coordonnées des composantes** (anglais : *component loadings* ou *species scores*) sont les coordonnées (dans la base canonique) des vecteurs propres, *i.e.* $(Q)_{jk} = \mathbf{q}_{j;(k)}$.

— La **proportion de la variance expliquée par la composante** PC_k (anglais : *proportion of variance explained by* PC_k) est le rapport $\dfrac{\lambda_k}{\sum_j \lambda_j} = \dfrac{\lambda_k}{\mathrm{Tr}V_X}$.

— Les **corrélations** entre les variables (initiales) et les PC :

$$r(\mathbf{x}_{(j)}, \mathbf{y}_{(k)}) = \frac{\sqrt{\lambda_k} q_{jk}}{\sqrt{V_{jj}}}.$$

Pour voir ce dernier point, si $\underline{\mathrm{v}}_{(i)}$ est la $i^{\text{ème}}$ ligne de la matrice V_X :

$$\mathrm{Cov}(\mathbf{x}_{(j)}, \mathbf{y}_{(k)}) = \tfrac{1}{n}(\mathbf{x}_{(j)} - \widehat{\mathrm{x}}_{(j)})^{\mathsf{T}}(\mathbf{y}_{(k)} - \widehat{\mathrm{y}}_{(k)} = \tfrac{1}{n}(\mathbf{x}_{(j)} - \widehat{\mathrm{x}}_{(j)})^{\mathsf{T}}(X - \hat{X})\mathbf{q}_{(k)} = \underline{\mathrm{v}}_{(i)}\mathbf{q}_{(k)} = \lambda_k \mathbf{q}_{i;(k)}$$

Puis, il faut diviser par la racine carré de la variance de $\mathbf{x}_{(j)}$, soit V_{jj}, et la racine carrée de la variance de $\mathbf{y}_{(k)}$, soit la $k^{\text{ème}}$ valeur propre.

(!) Il arrive que les coordonnées des individus et des composantes soient définies comme des éléments de U et W où $U\Sigma W^{\mathsf{T}}$ est la *décomposition en valeur singulière* (anglais : *singular value decomposition*, ou SVD) de $X - \hat{X}$. La lien avec notre convention est dicté par :

$$Y = U\Sigma, \quad W = Q, \quad \text{et} \quad (\Sigma)_{ij} = \begin{cases} \tilde{\lambda}_i & \text{si } i = j; \\ 0 & \text{si } i \neq j; \end{cases}$$

où $\tilde{\lambda}_i$ est la $i^{\text{ème}}$ valeur propre de la matrice $(n-1)V_X$.

La représentation graphique des données est le but principal de l'ACP. Les noms ci-dessous sont tirés des logiciels "Canoco" ou "R" (package *vegan*). Il y a deux méthodes standard :

— Le **bi-graphe des distances** (anglais : *distance biplot (scaling 1)* est une représentation graphique où les vecteurs propres sont de longueur 1. La distance entre les points (représentant des individus/sites/échantillons) est aussi proche que possible de la distance originelle.
Si on a décidé de garder les k premières valeurs propres, ce graphique est obtenu tout simplement en plaçant un point pour le $i^{\text{ème}}$ individu dont les coordonnées sont les k première coordonnée de $\underline{\mathrm{y}}_i$, *i.e.*

$$\big(\langle \underline{\mathrm{x}}_i, \mathbf{q}_{(1)} \rangle, \dots, \langle \underline{\mathrm{x}}_i, \mathbf{q}_{(k)} \rangle\big).$$

(La plupart du temps il n'y aura que 2 coordonnées !) Les "axes" du graphique sont les deux vecteurs propres. On ajoute aussi parfois les "anciens axes" (qui correspondent aux vecteurs

6. centrées : de sorte que la moyenne de chaque variable est 0.

canoniques, *i.e.* aux variables initiales). La $j^{\text{ème}}$ variable initiale sera représentée par une flèche qui commence à l'origine et se termine en

$$\left(\mathbf{q}_{j;(1)}, \dots, \mathbf{q}_{j;(k)}\right)$$

— Le **bi-graphe des corrélations** (anglais : *correlation biplot (scaling 2)* est une représentation où les vecteurs (représentant les variables initiales) ont leur $i^{\text{ème}}$ coordonnée multipliée par la racine de la $i^{\text{ème}}$ valeur propre, tandis que les $i^{\text{èmes}}$ coordonnées des individus sont multipliées par l'inverse de la racine carrée de la $i^{\text{ème}}$ valeur propre. Dans cette représentation, l'angle entre les variables est approximativement leur corrélation. La distance entre les individus ne réflète plus rien. La projection des individus le long de l'axe engendré par un des vecteurs des variables initiales est approximativement la présence de la dite variable dans les données de l'individu.
Pour expliquer brièvement pourquoi cette modification permet de retrouver les corrélations, écrivons

$$\mathbf{e}_{(i)} = \sum_{k=1}^{p} c_{i,k}\mathbf{q}_{(k)} \qquad \text{où } c_{i,k} = \langle \mathbf{e}_{(i)}, \mathbf{q}_{(k)} \rangle = \mathbf{q}_{i;(k)}$$

Alors

$$\begin{aligned}(V)_{ij} &= \mathbf{e}_{(i)}^{\mathsf{T}} V \mathbf{e}_{(j)} \\ &= \left(\sum_{k=1}^{p} c_{i,k}\mathbf{q}_{(k)}\right)^{\mathsf{T}} V \left(\sum_{k=1}^{p} c_{j,k}\mathbf{q}_{(k)}\right) \\ &= \sum_{k=1}^{p} c_{i,k}c_{j,k}\lambda_k \qquad\qquad = \sum_{k=1}^{p} (\sqrt{\lambda_k}c_{i,k})(\sqrt{\lambda_k}c_{j,k})\end{aligned}$$

Ainsi, si on donne au vecteur $\mathbf{e}_{(i)}$ les coordonnées $\mathbf{v}_{(i)} = \begin{pmatrix} \lambda_k c_{1,k} \\ \vdots \\ \lambda_k c_{p,k} \end{pmatrix}$, il apparaît :

$$(R)_{ij} = \frac{(V)_{ij}}{\sqrt{(V)_{ii}(V)_{jj}}} = \frac{\langle \mathbf{v}_{(i)}, \mathbf{v}_{(j)} \rangle}{\|\mathbf{v}_{(i)}\| \cdot \|\mathbf{v}_{(j)}\|}.$$

Évidemment, comme on fait une projection, l'angle obtenu dans le graphique bi-dimensionnel (ou tri-) n'est qu'une approximation. Plus la proportion de la variabilité contenue dans les composantes principales est grande, plus cette approximation sera bonne.
Un calcul similaire montre que la multiplication de la $i^{\text{ème}}$ coordonnée des individus par $1/\sqrt{\lambda_i}$ préserve (exactement avant projection, approximativement après projection) l'amplitude des individus le long des variables initiales.

2.1.iii Le point de vue de la régression linéaire

Une autre perspective sur l'ACP provient du problème de la régression linéaire (voir la section 1.4). Rappelons que géométriquement, les valeurs ajustées $\tilde{\mathbf{s}} = \mathbf{s} - \mathbf{u} = b_0\mathbb{1}_n + \sum_{i=1}^{k} b_i\mathbf{r}_{(i)}$ représente tout simplement la projection (dans $\mathbb{R}^n$) de $\mathbf{s}$ sur l'espace (normalement, de dimension $k+1$) engendré par $\mathbb{1}_n, \mathbf{r}_{(1)}, \dots, \mathbf{r}_{(k)}$. Voir la section 1.4 pour un peu plus de détails.

La proposition suivante essaye de mettre l'emphase sur l'interprétation suivante : si on devait tenter d'expliquer toutes les données recueillies X, par une petite parties de celles-ci (obtenues par

des combinaisons linéaires des données initiales), quelle serait le meilleur choix possible ? Il se trouve que l'ACP est aussi une solution à ce problème.

Proposition 2.1.10

Soient $\mathbf{r}_{(\ell)} = X\,\mathbf{c}_{(\ell)}$ (où $\ell = 1,...,k$) des combinaisons linéaires des variables $\mathbf{x}_{(j)}$. Soit $\mathrm{Var}(\mathbf{u}_{(i)})$ la variance résiduelle dans la régression $\mathbf{x}_{(i)}$ par les $\mathbf{r}_{(\ell)}$ (où $\ell = 1,...,k$). Alors le minimum de la somme des variances résiduelles

$$\min_{\mathbf{c}_{(1)},\dots,\mathbf{c}_{(k)}} \sum_{i=1}^{p} \mathrm{Var}(\mathbf{u}_{(i)})$$

est obtenu lorsque les $\mathbf{c}_{(\ell)}$ sont les k vecteurs propres de V_X de valeur propres maximale.

DÉMONSTRATION SI $k = 1$: Comme cela simplifie les choses significativement, il vaut mieux s'attarder d'abord au cas $k = 1$. Aussi, on supposera X centrée. Soit $\mathbf{r} = X\mathbf{c}$ l'unique combinaison linéaire. La régression linéaire de $\mathbf{x}_{(i)}$ ne correspond alors qu'à trouver β_i dans $\mathbf{x}_{(i)} = \beta_i\mathbf{r} + \mathbf{u}$ de sorte que la variance résiduelle soit minimale. Or $\beta_i = (\mathbf{r}^\mathsf{T}\mathbf{r})^{-1}\mathbf{r}^\mathsf{T}\mathbf{x}_{(i)} = \dfrac{\mathbf{r}^\mathsf{T}\mathbf{x}_{(i)}}{\mathbf{r}^\mathsf{T}\mathbf{r}}$ et la variance résiduelle est

$$\begin{array}{lll} \mathrm{Var}(\mathbf{u}_{(i)}) & = \mathrm{Var}(\mathbf{x}_{(i)}) - \mathrm{Cov}(\mathbf{x}_{(i)}, \beta_i\mathbf{r}) & = \mathrm{Var}(\mathbf{x}_{(i)}) - \beta_i\mathrm{Cov}(\mathbf{x}_{(i)}, \mathbf{r}) \\ & = \mathrm{Var}(\mathbf{x}_{(i)}) - \beta_i\mathbf{x}_{(i)}^\mathsf{T}\mathbf{r} & = \mathrm{Var}(\mathbf{x}_{(i)}) - \dfrac{\mathbf{r}^\mathsf{T}\mathbf{x}_{(i)}\mathbf{x}_{(i)}^\mathsf{T}\mathbf{r}}{\mathbf{r}^\mathsf{T}\mathbf{r}} \end{array}$$

En prenant la somme sur i, il apparaît

$$\sum_i \mathrm{Var}(\mathbf{u}_{(i)}) = \mathrm{Tr}(V_X) - \frac{\mathbf{c}^\mathsf{T}X^\mathsf{T}XX^\mathsf{T}X\mathbf{c}}{\mathbf{c}^\mathsf{T}X^\mathsf{T}X\mathbf{c}} = \mathrm{Tr}(V_X) - \frac{\mathbf{c}^\mathsf{T}V_X^2\mathbf{c}}{\mathbf{c}^\mathsf{T}V_X\mathbf{c}}.$$

Une application du lemme 2.A.4, montre que le vecteur $\mathbf{b}$ qui minimise $\sum_i \mathrm{Var}(\mathbf{u}_{(i)})$ est solution de

$$(V_X^2 - \lambda V_X)\mathbf{c} = 0 \qquad (\text{qu'on réécrit} \quad V_X(V_X - \lambda\mathrm{Id})\mathbf{c} = 0 \quad)$$

pour la plus grande valeur de λ possible. La réécriture montre (presque [7]) que $\mathbf{c}$ est bien un vecteur propre de V_X. ■

DÉMONSTRATION GÉNÉRALE : Il est possible de supposer que les $\mathbf{r}_{(\ell)} = X\,\mathbf{c}_{(\ell)}$ sont non-corrélés (*i.e.* de covariances nulles) et de variance 1. Soit R la matrice dont les colonnes sont les $\mathbf{r}_{(\ell)}$ et C la matrice dont les colonnes sont les $\mathbf{c}_{(\ell)}$: alors $R = XC$. Si les X sont supposées centrées, alors les R seront aussi centrées et $V_R = R^\mathsf{T}R = \mathrm{Id}$ (comme les variances sont 1 et les covariances 0).

La régression de $\mathbf{x}_{(i)}$ expliquée par la matrice R donne la valeur ajustée

$$\tilde{\mathbf{x}}_{(i)} = R\mathbf{b}_{(i)} = R(R^\mathsf{T}R)^{-1}R^\mathsf{T}\mathbf{x}_{(i)} = RR^\mathsf{T}\mathbf{x}_{(i)}$$

par la proposition 1.4.2 et puisque $(R^\mathsf{T}R)^{-1} = R^\mathsf{T}R = \mathrm{Id}$. En utilisant le lemme 1.4.3, il apparaît que la somme des variances résiduelles est

$$\begin{array}{lll} \sum_{j=1}^p \mathrm{Var}(u_{(j)}) & = \sum_{j=1}^p \left(\mathrm{Var}(\mathbf{x}_{(j)}) - \mathrm{Cov}(\mathbf{x}_{(j)}, R\mathbf{b}_{(j)})\right) & = \mathrm{Tr}V_X - \sum_{j=1}^p \left(\mathbf{x}_{(j)}^\mathsf{T}RR^\mathsf{T}\mathbf{x}_{(j)}\right) \\ & = \mathrm{Tr}V_X - \sum_{j=1}^p \mathbf{e}_{(j)}^\mathsf{T}X^\mathsf{T}RR^\mathsf{T}X\mathbf{e}_{(j)} & = \mathrm{Tr}V_X - \mathrm{Tr}(X^\mathsf{T}RR^\mathsf{T}X) \\ & = \mathrm{Tr}V_X - \mathrm{Tr}(R^\mathsf{T}XX^\mathsf{T}R) & = \mathrm{Tr}V_X - \mathrm{Tr}(C^\mathsf{T}X^\mathsf{T}XX^\mathsf{T}XC) \\ & = \mathrm{Tr}V_X - \mathrm{Tr}(C^\mathsf{T}V_X^2C) & \end{array}$$

7. voire la démonstration générale : il faut en fait supposer que $\mathrm{Det}V_X \neq 0$.

où V_X est la matrice des variances de X. Ainsi, il faut chercher le maximum de $\sum_{j=1}^{k} \mathbf{c}_{(j)}^{\mathsf{T}} V_X^2 \mathbf{c}_{(j)}$ sous la contrainte

$$R^{\mathsf{T}}R = \mathrm{Id} \qquad \Leftrightarrow C^{\mathsf{T}}X^{\mathsf{T}}XC = \mathrm{Id} \qquad \Leftrightarrow \mathbf{c}_{(j)}^{\mathsf{T}} V_X \mathbf{c}_{(i)} = \begin{cases} 1 & \text{si } i = j; \\ 0 & \text{si } i \neq j. \end{cases}$$

À ce moment on peut déjà voir que si les $\mathbf{c}_{(i)} = \lambda_i^{-1/2}\mathbf{v}_{(i)}$ où $\mathbf{v}_{(i)}$ est un vecteur propre, alors

$$\mathbf{c}_{(j)}^{\mathsf{T}} V_X \mathbf{c}_{(i)} = \lambda_i^{-1/2}\lambda_j^{-1/2}\mathbf{v}_{(j)}^{\mathsf{T}} V_X \mathbf{v}_{(i)} = \lambda_i^{1/2}\lambda_j^{-1/2}\mathbf{v}_{(j)}^{\mathsf{T}}\mathbf{v}_{(i)} = \begin{cases} 1 & \text{si } i = j; \\ 0 & \text{si } i \neq j; \end{cases}$$

et $\sum_{j=1}^{p} \mathrm{Var}(u_{(j)}) = \sum_{i=1}^{k} \lambda_k$.

Pour se convaincre que c'est bien la solution, il vaut mieux supposer que V_X n'est pas de déterminant nul. Dans ce cas, par le lemme 2.A.3, la matrice $V_X^{-1/2}$ existe et il est possible de poser $\mathbf{c}_{(j)} = V_X^{-1/2}\mathbf{v}_{(j)}$. Alors, le problème est de maximiser $\sum_{j=1}^{k} \mathbf{v}_{(j)}^{\mathsf{T}} V_X \mathbf{v}_{(j)}$ sous la contrainte $\mathbf{v}_{(j)}^{\mathsf{T}}\mathbf{v}_{(j)} = 1$ (*i.e.* $\|\mathbf{v}_{(j)}\| = 1$).

Voir la section 2.A pour l'explication que la solution de ce dernier problème est un vecteur propre de V_X. ■

2.1.iv Quelques remarques supplémentaires

Le nombre de dimensions (*i.e.* le nombre de composantes principales conservées) à la fin de l'analyse ne doit jamais être élevé : 2, parfois 3, très rarement 4. Cela tient cependant seulement à des raisons de lisibilité. Ainsi, ce nombre doit être petit (lisibilité) mais avoir une grande variabilité (pour ne pas perdre trop d'information).

Il n'y a pas de règles strictes. Si la proportion de la variabilité expliquée par les valeurs propres (correspondantes aux variables conservées) est inférieure à 60%, c'est cependant *très* mauvais signe.

Voici deux exemples basés sur le fait que certains programmes ne donnent que les 4 premières valeurs propres de l'analyse (les autres restent inconnues).

Exemple 2.1.11. Supposons le cas suivant : notons p_i la $i^{\text{ème}}$ valeur propre divisée par la variabilité totale (*i.e.* la proportion de la variabilité qu'elle explique). Supposons que $p_1 = 0.281$, $p_2 = 0.072$, $p_3 = 0.063$ et $p_4 = 0.008$. Comme elles sont données en ordre décroissant, tous les p_i pour $i \geq 5$ auront une contribution de moins de 0.008 (*i.e.* moins de 1%). Conséquemment, il faudrait vraiment cumulé beaucoup de variables supplémentaires (*i.e.* au moins 50, car $50 \times 0.008 = 0.4$) pour obtenir une contribution plus grande que celle des trois premières ($p_1 + p_2 + p_3 \simeq 0.416 = 41.6\%$). Il n'est alors pas complètement déraisonnable de tirer des conclusions, même si elles doivent être très prudentes. Il est surtout plus raisonnable soit de refaire des expériences (en laboratoire, où les autres variables peuvent être contrôlées) soit d'augmenter sa base de donnée. ♣

Exemple 2.1.12. Si on avait $p_1 = 0.230$, $p_2 = 0.105$, $p_3 = 0.104$ et $p_4 = 0.101$, les conclusions sur les trois premiers axes seraient fatalement douteuses : en effet, il ne serait pas impossible d'avoir que p_5, p_6 et p_7 soit d'environs 0.1 dans quel cas considérer les axes correspondant à p_4, p_5, p_6 et p_7 aurait à peu près autant de poids que p_1, p_2 et p_3 (environs 40% contre 43.9%). ♣

Lorsque les variables qu'on garde expliquent une variation de moins de 70%, ce qui est surtout utile c'est de voir combien de variables supplémentaires sont nécessaires pour expliquer autant de variations que les variables qu'on garde. Il est assez facile de voir que ceci donne :

$$\frac{p_1 + \ldots + p_j}{p_{j+1}}.$$

En appliquant bêtement cette formule dans 2.1.11, en partant du fait qu'on voudrait garder deux variables, on obtiendrai :

$$\frac{0.281 + 0.072}{0.063} = 5.603$$

Mais ce nombre est beaucoup trop petit. Dans ces situations, il est préférable de se servir de toute l'information connue. Dans l'exemple 2.1.11, comme on connaît p_3, le nombre minimal de variables supplémentaires pour retrouver autant de variabilité que dans $p_1 + p_2$ est

$$1 + \frac{0.281 + 0.072 - 0.063}{0.008} = 37.25$$

Ce qui est beaucoup plus confortable.

Les transformations préalables sur les variables. En pratique, il est préférable d'utiliser les variables centrées réduites. En effet, il est impossible au moment de faire l'expérience de savoir quelle est la bonne échelle à utiliser. Parfois, les variables sont aussi de sens très différent, il est ainsi futile de les comparer sans une renormalisation.

Si, dans une expérience, on mesure en centimètres, la taille des jambes et la taille du petit orteil, les variables centrées réduites sont plus appropriées : une variation de 1cm sur la taille du petit orteil est une grande variation, tandis qu'une variation de 1cm sur la taille des jambes est une faible variation. En bref, si les variables ne sont pas comparables, il vaut mieux utiliser les variables centrées réduites.

Par contre, si l'on mesure les différentes odeurs dégagées par un animal, on pourrait considérer rester avec les données initiales. Cependant, la sensibilité à différentes odeurs peut-être extrêmement variable d'un animal à l'autre, et il est impensable de connaître la sensibilité des cellules à une odeur donnée. D'autre part, pour ces données sensorielles, une transformation logarithmique peut s'imposer : les échelles logarithmique correspondent souvent mieux aux perceptions sensorielles par des animaux (*e.g.* les décibels pour l'intensité du son). Une transformation logarithmique a aussi tendance à gommer les changements d'échelle : $\ln(ab) = \ln a + \ln b$ ainsi un changement d'échelle ne contribuera qu'un terme additif (et disparaîtra dans la moyenne) [8].

Un certain nombre d'autres transformations sont usuelles. Par exemple, quand le nombre de données ayant la valeur 0 sont présentes en grand nombres.

2.2 L'analyse de redondance (ou ARD)

... ou *redundancy analysis* (RDA).

8. Faites attention : la plupart du temps les transformations logarithmiques sont de la forme $a \mapsto \ln(1+a)$ et n'ont pas cette propriété,

Dans l'analyse de redondance on a X, la matrice [des données] explicative de taille $n \times p$ (*e.g.* des données environnementales : acidité du sol, pluviométrie, ...) et Y, la matrice [des données] expliquées [9] de taille $n \times q$ (*e.g.* des données espèces : abondance de tel ou tel autre type de plante/champignon/...).

Le principe de l'ARD est de chercher la combinaison de variables "environnementales" qui explique le mieux la variation (ou la dispersion) de la matrice "espèces".

La méthode de l'ARD consiste d'abord à régresser tour à tour chaque variable expliquée sur les variables explicatives. De ces régressions multiples sont extraites les valeurs ajustées. Cette nouvelle matrice des valeurs ajustées est ensuite soumise à une ACP.

Supposons que matrices X et Y sont centrées et réduites [10] (*i.e.* toutes les variable sont de moyenne 0 et de variance 1). Les valeurs ajustées sont données par la formule de régression suivante (voir la section 1.4) :

$$\widetilde{Y} = X(X^\mathsf{T}X)^{-1}X^\mathsf{T}Y$$

La matrice de covariance de valeurs ajustées est donnée par

$$V_{\widetilde{Y}\widetilde{Y}} = \frac{1}{n}\widetilde{Y}^\mathsf{T}\widetilde{Y} = \frac{1}{n}Y^\mathsf{T}X(X^\mathsf{T}X)^{-1\mathsf{T}}X^\mathsf{T}X(X^\mathsf{T}X)^{-1}X^\mathsf{T}Y = \frac{1}{n}Y^\mathsf{T}X(X^\mathsf{T}X)^{-1\mathsf{T}}X^\mathsf{T}Y = V_{YX}V_{XX}^{-1}V_{XY}$$

où $V_{YX} = \frac{1}{n}(Y-\hat{Y})^\mathsf{T}\cdot(X-\hat{X}) = \frac{1}{n}Y^\mathsf{T}X$ (et on a utiliser que V_{XX}^{-1} est [tout comme V_{XX}] aussi une matrice symétrique).

Théorème 2.2.1

Soit X, Y les matrices de données, et supposons qu'elles sont centrées réduites. Supposons aussi que V_{XX} est de déterminant non-nul [11]*. Trouver les vecteurs $\mathbf{v}$ qui maximisent la somme des carres de corrélations entre les colonnes de Y et $X\mathbf{v}$, sous la contrainte* $\mathrm{Var}(X\mathbf{v}) = 1$*, revient a résoudre l'équation de l'ARD*

$$(V_{XY}V_{YX} - \lambda V_{XX})\mathbf{v} = 0 \tag{2.2.2}$$

pour les λ les plus grands. De plus, les vecteurs $\mathbf{s} = V_{YX}\mathbf{v}$ sont des vecteurs propres pour la matrice $V_{\widetilde{Y}\widetilde{Y}}$, i.e. *ils donnent la solution au problème de l'ACP pour la matrice des données ajustées $\widetilde{Y}$ (pour la même valeur de λ) :*

$$(V_{YX}V_{XX}^{-1}V_{XY} - \lambda \mathrm{Id}_q)\mathbf{s} = 0, \tag{2.2.3}$$

En particulier, si $X = Y$ le problème est identique à celui de l'ACP.

DÉMONSTRATION: Comme les variables sont centrées et réduites, et si $\mathbf{z}$ est une autre variable centrée réduite, alors la corrélation entre la variable $\mathbf{y}_{(i)}$ la variable $\mathbf{z}$ est donnée par la formule

$$\mathrm{Cov}(\mathbf{y}_{(i)}, \mathbf{z}) = \frac{1}{n}\langle \mathbf{y}_{(i)}, \mathbf{z}\rangle = \frac{1}{n}\mathbf{y}_{(i)}^\mathsf{T}\mathbf{z}.$$

9. ou plutôt, données "à expliquer" (expliquandes), parfois aussi appelée données/matrice "réponse[s]"
10. En particulier, dans toute cette section il n'y aura pas d'apparition des $\hat{X}$ ou $\hat{Y}$ car ils seront supposés nuls.

Ainsi, les divers corrélations entre les $\mathbf{y}_{(i)}$ et $X\mathbf{v}$ est donnée par le vecteur $Y^\mathsf{T}X\mathbf{v}$. La somme des corrélations au carré est la norme de ce vecteur soit (à une constante $\frac{1}{n^2}$ près) :

$$\|Y^\mathsf{T}X\mathbf{v}\|^2 = \langle Y^\mathsf{T}X\mathbf{v}, Y^\mathsf{T}X\mathbf{v}\rangle = \mathbf{v}^\mathsf{T}X^\mathsf{T}YY^\mathsf{T}X\mathbf{v}.$$

La fonction à maximiser est donc $\mathbf{v}^\mathsf{T}X^\mathsf{T}YY^\mathsf{T}X\mathbf{v}$ avec la contrainte $\mathbf{v}^\mathsf{T}X^\mathsf{T}X\mathbf{v} = 1$. Par le lemme 2.A.3, comme Det$V_{XX} \neq 0$, la matrice $V_{XX}^{-1/2}$ existe et il est possible de faire un "changement de variable" en posant : $V_{XX}^{1/2}\mathbf{v} = \mathbf{w}$. Puisque $X^\mathsf{T}X = V_{XX}$, le problème revient a maximiser $\mathbf{w}^\mathsf{T}(V_{XX}^{-1/2}V_{XY}V_{YX}V_{XX}^{-1/2})\mathbf{w}$ avec la contrainte $\mathbf{w}^\mathsf{T}\mathbf{w} = 1$.

Comme on a vu dans le cadre de l'ACP, ce problème est résolu lorsque $\mathbf{w}$ est un vecteur propre de $V_{XX}^{-1/2}V_{XY}V_{YX}V_{XX}^{-1/2}$, *i.e.* il existe λ tel que

$$(V_{XX}^{-1/2}V_{XY}V_{YX}V_{XX}^{-1/2} - \lambda\mathrm{Id}_p)\mathbf{w} = 0$$

Pour retrouver l'équation de $\mathbf{v} = V_{XX}^{-1/2}\mathbf{w}$, il suffit d'écrire :

$$0 = (V_{XX}^{-1/2}V_{XY}V_{YX}V_{XX}^{-1/2} - \lambda\mathrm{Id}_p)\mathbf{w} = (V_{XX}^{-1/2}V_{XY}V_{YX} - \lambda V_{XX}^{1/2})\mathbf{v},$$

puis en multipliant à gauche par $V_{XX}^{1/2}$, ceci est équivalent (comme V_{XX} est inversible) à :

$$(V_{XY}V_{YX} - \lambda V_{XX})\mathbf{v} = 0$$

De plus, si $\mathbf{s}$ est un vecteur propre de $V_{\widetilde{Y}\widetilde{Y}} = V_{YX}V_{XX}^{-1}V_{XY}$ (*i.e.* une solution à l'ACP sur $V_{\widetilde{Y}\widetilde{Y}}$), alors

$$(V_{YX}V_{XX}^{-1}V_{XY} - \lambda\mathrm{Id}_q)\mathbf{s} = 0$$

En posant, $\mathbf{s} = V_{YX}\mathbf{v} = V_{YX}V_{XX}^{-1/2}\mathbf{w}$, il apparaît :

$$\begin{aligned}(V_{YX}V_{XX}^{-1}V_{XY} - \lambda\mathrm{Id}_q)V_{YX}\mathbf{v} &= (V_{YX}V_{XX}^{-1}V_{XY}V_{YX} - \lambda V_{YX})\mathbf{v} \\ &= V_{YX}(V_{XX}^{-1}V_{XY}V_{YX} - \lambda\mathrm{Id}_p)\mathbf{v} \\ &= V_{YX}V_{XX}^{-1}(V_{XY}V_{YX} - \lambda V_{XX})\mathbf{v} \quad = 0\end{aligned}$$

Autrement dit, pour trouver $\mathbf{s}$ (le vecteur "solution" au problème de l'ACP pour $\widetilde{Y}$), il suffit de trouver $\mathbf{v}$ (le vecteur pour résoudre le problème de l'ARD). ■

Corollaire 2.2.4

*(Voir [6]) Les r combinaisons linéaires des variables $\mathbf{x}_{(i)}$ (disons $X\mathbf{v}_{(k)}$ où $k = 1, ..., \ell$) qui font en sorte que la variance résiduelle (*cf. *lemme 1.4.3) dans l'explication des variables Y est minimale, sont les r vecteurs (propres) solutions de l'équation* (2.2.2) *:*

$$(V_{XY}V_{YX} - \lambda V_{XX})\mathbf{v} = 0$$

correspondant aux valeurs de λ les plus grandes.

DÉMONSTRATION: Supposons que X et Y sont centrées. Et, pour simplifier l'exposition, supposons que $\ell = 1$ (*i.e.* qu'on ne cherche qu'une combinaison linéaire des $\mathbf{x}_{(i)}$ qui satisfait). Soit $\mathbf{z} = X\mathbf{b}$

cette combinaison linéaire. La régression linéaire de $\mathbf{y}_{(i)}$ ne correspond alors qu'à trouver β_i dans $\mathbf{y}_{(i)} = \beta_i \mathbf{z} + \mathbf{u}$ de sorte que la variance résiduelle soit minimale. Or cette variance résiduelle est

$$\begin{aligned} \mathrm{Var}(\mathbf{u}_{(i)}) &= \mathrm{Var}(\mathbf{y}_{(i)}) - \frac{\mathrm{Cov}(\mathbf{y}_{(i)}, \beta_i \mathbf{z})^2}{\mathrm{Var}(\beta_i \mathbf{z})} &= \mathrm{Var}(\mathbf{y}_{(i)}) - \frac{\mathrm{Cov}(\mathbf{y}_{(i)}, \mathbf{z})^2}{\mathrm{Var}(\mathbf{z})} \\ &= \mathrm{Var}(\mathbf{y}_{(i)}) - \frac{(\mathbf{y}_{(i)}^\mathsf{T} X\mathbf{b})^2}{\mathbf{b}^\mathsf{T} X^\mathsf{T} X\mathbf{b}} &= \mathrm{Var}(\mathbf{y}_{(i)}) - \frac{\mathbf{b}^\mathsf{T} X^\mathsf{T} \mathbf{y}_{(i)} \mathbf{y}_{(i)}^\mathsf{T} X\mathbf{b}}{\mathbf{b}^\mathsf{T} X^\mathsf{T} X\mathbf{b}}. \end{aligned}$$

En prenant la somme sur i, il apparaît

$$\sum_i \mathrm{Var}(\mathbf{u}_{(i)}) = \mathrm{Tr}(V_{YY}) - \frac{\mathbf{b}^\mathsf{T} X^\mathsf{T} Y Y^\mathsf{T} X\mathbf{b}}{\mathbf{b}^\mathsf{T} X^\mathsf{T} X\mathbf{b}}.$$

Une application du lemme 2.A.4, montre que le vecteur $\mathbf{b}$ qui minimise $\sum_i \mathrm{Var}(\mathbf{u}_{(i)})$ est le vecteur solution de l'équation (2.2.2) pour la plus grande valeur de λ possible.

Pour les combinaisons linéaires suivantes (*i.e.* si $\ell > 1$), il faut argumenter comme à la proposition 2.1.10 et cela complique significativement les équations. ■

L'**indice de redondance** (en anglais *redundancy index*) est

$$\rho I(Y,X) = \frac{\mathrm{Tr} V_{\widetilde{Y}\widetilde{Y}}}{\mathrm{Tr} V_{YY}} = \frac{\mathrm{Tr}\left(V_{YX} V_{XX}^{-1} V_{XY}\right)}{\mathrm{Tr} V_{YY}}.$$

Il exprime le rapport entre la variabilité des données ajustées $\widetilde{Y}$ et celle des données (à expliquer) initiales Y.

Corollaire 2.2.5

(Voir [4]) L'ARD produit des combinaisons linéaires des variables X ($\mathbf{r}_{(i)} = X\mathbf{c}_{(i)}$ où $i = 1, \ldots, k$) non-correlées et de variance 1*, de sorte que la somme des indices de redondance pour Y (*i.e. $\sum_{i=1}^{k} \rho I(Y, \mathbf{r}_{(i)})$*) soit maximale.*

2.3 Quelques autres variantes

... en construction ...

2.3.i Analyse factorielle de correspondance ou AFC

... ou encore *correspondance analysis* (CA). Le "factorielle" est parfois aussi abandonné en français.

Analyse en composantes principales généralisée Métrique dans l'espace des individus :

$$\langle \mathbf{x}_{i_1}^t ; \mathbf{x}_{i_2}^t \rangle = \mathbf{x}_{i_1} M \mathbf{x}_{i_2}^t$$

Métrique dans l'espace des variables (la métrique des poids) :

$$\langle \mathbf{x}_{(j_1)} ; \mathbf{x}_{(j_2)} \rangle = \mathbf{x}_{(j_1)}^t D \mathbf{x}_{(j_2)}$$

Soit $\mathcal{P} : \mathbb{R}^p \to W^k$ l'opérateur de projection M-orthogonale sur un sous-espace de dimension k :

$$\mathcal{P}^2 = \mathcal{P}, \quad \mathcal{P}^t M = M\mathcal{P}.$$

On cherche W^k tel que l'inertie du nuage projeté soit maximale, ce qui revient a trouver $\mathcal{P}$ de rang k maximisant $\mathrm{Tr}(VMP)$, où $V = X^t DX$ est la matrice de covariance de X. Les composantes principales seront les vecteurs propres de $XMX^t D$.

L'analyse factorielle des correspondances Si $N = (n_{ij}) \in \mathcal{M}_{m_1 \times m_2}(\mathbb{N})$ un tableau contingence de 2 variables qualitatives. Notons

$$D_1 := \mathrm{Diag}(n_{1\cdot}, \cdots, n_{m_1\cdot}), \quad D_2 := \mathrm{Diag}(n_{\cdot 1}, \cdots, n_{\cdot m_2})$$

les matrices diagonales des effectifs marginaux. On veux rendre compte de la structure des écarts a l'indépendance en projetant sur un espace réduit les nuages de points profil-lignes $D_1^{-1}N$ (et colonnes) en gardant le maximum d'inertie. L'inertie doit alors représenter l'écart a l'indépendance [12] ce que nous conduit à choisir la distance χ^2 entre deux profils-lignes :

$$d_{\chi^2}(i,i') = \sqrt{\sum_{j=1}^{m_2} \frac{n}{n_{\cdot j}} \left(\frac{n_{ij}}{n_{i\cdot}} - \frac{n_{i'j}}{n_{i'\cdot}} \right)^2}.$$

Notons que d_{χ^2} est obtenue du produit scalaire défini par $M = nD_2^{-1}$.

Alors AFC reviens a une ACP avec les données $X = D_1^{-1}N$, utilisant les métriques $M = nD_2^{-1}$ et $D = \frac{1}{n}D_1$. La symétrie avec la construction analogue pour les profiles colonnes nous permet de superposer les plans principaux des deux ACP et obtenir une représentation simultanée des catégories de deux variables croisées dans le tableau de contingence N.

2.3.ii Analyse factorielle multiple

... ou encore *multiple factor analysis* (MFA).

... à venir ...

2.A Théorème spectral et Multiplicateurs de Lagrange

12. Représentée par la relation $n_{ij} = \frac{n_{i\cdot} n_{\cdot j}}{n}$

Théorème 2.A.1

Si M est une matrice symétrique $M = M^{\mathsf{T}}$ de taille $n \times n$, alors il existe n vecteur propres orthogonaux (et de norme 1) : $\boldsymbol{q}_{(1)}, \ldots, \boldsymbol{q}_{(n)}$ tels que si Q est la matrice dont la $i^{\text{ème}}$ colonne est $\boldsymbol{q}_{(i)}$

$$Q^{\mathsf{T}} M Q = \begin{pmatrix} \lambda_1 & 0 & \ldots & 0 & 0 \\ 0 & \lambda_2 & \ldots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & \lambda_{n-1} & 0 \\ 0 & 0 & \ldots & 0 & \lambda_n \end{pmatrix} =: \Lambda$$

où $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$ sont les valeurs propres de M.

DÉMONSTRATION: On emploie ici la méthode des multiplicateurs de Lagrange (ici avec les termes de la statistique).

Étape I. Trouver $\mathbf{q}_{(1)}$ de sorte que $\mathbf{y}_{(1)}$ ait la plus grande variance mais sous la contrainte $\mathbf{q}_{(1)}^t \mathbf{q}_{(1)} = 1$. Soit le lagrangien

$$L(\mathbf{q}_{(1)}) = \mathbf{q}_{(1)}^t V \mathbf{q}_{(1)} - \lambda(\mathbf{q}_{(1)}^t \mathbf{q}_{(1)} - 1)$$

La condition nécessaire pour être un maximum est $\frac{\partial L}{\partial \mathbf{q}_{(1)}} = 0$. Puisque $\frac{\partial L}{\partial \mathbf{q}_{(1)}} = 2V\mathbf{q}_{(1)} - 2\lambda\mathbf{q}_{(1)}$ il apparaît

$$(V - \lambda I_p)\mathbf{q}_{(1)} = 0$$

Ainsi λ doit être une valeur propre et $\mathrm{Var}(\mathbf{y}_{(1)}) = \mathbf{q}_{(1)}^t V \mathbf{q}_{(1)} = \mathbf{q}_{(1)}^t \lambda I_p \mathbf{q}_{(1)} = \lambda$. On prends alors $\lambda = \lambda_1$, la plus grande valeur propre.

Étape II. Trouver $\mathbf{q}_{(2)}$ de sorte que $\mathbf{y}_{(2)}$ est non-correlé à $\mathbf{y}_{(1)}$ et a la plus grande variance possiblesous la contrainte $\mathbf{q}_{(2)}^t \mathbf{q}_{(2)} = 1$. Comme $\mathrm{Cov}(\mathbf{y}_{(2)}, \mathbf{y}_{(2)}) = \lambda_1 \mathbf{q}_{(2)}^t \mathbf{q}_{(1)}$, le lagrangien est cette fois-ci

$$L(\mathbf{q}_{(2)}) = \mathbf{q}_{(2)}^t V \mathbf{q}_{(2)} - \lambda(\mathbf{q}_{(2)}^t \mathbf{q}_{(2)} - 1) - \delta \mathbf{q}_{(2)}^t \mathbf{q}_{(1)}$$

Une condition nécessaire est $\frac{\partial L}{\partial \mathbf{q}_{(2)}} = 0$. Ainsi

$$\frac{\partial L}{\partial \mathbf{q}_{(2)}} = 2(V - \lambda I_p)\mathbf{q}_{(2)}.$$

En prenant le produit scalaire de cette expression avec $\mathbf{q}_{(1)}$, et en utilisant que $\mathbf{q}_{(1)}^{\mathsf{T}} V^{\mathsf{T}} = \lambda_1 \mathbf{q}_{(1)}$, $\mathbf{q}_{(1)}^{\mathsf{T}} \mathbf{q}_{(2)} = 0$ et $\mathbf{q}_{(1)}^{\mathsf{T}} \mathbf{q}_{(1)} = 0$, la condition nécessaire implique que $\delta = 0$. Ceci connu, il apparaît ensuite que λ doit être la deuxième plus grande valeur propre, $\mathbf{q}_{(2)}$ le vecteur propre correspondant. ■

Le sens (important) du théorème spectral est que, dans une bonne base, la matrice est essentiellement la dilatation des axes par un facteur : le $i^{\text{ème}}$ axe $\mathbf{q}_{(i)}$ est dilaté par un facteur $|\lambda_i|$, et, si $\lambda_i < 0$ il est inversé [13].

Un corollaire important de la démonstration :

Corollaire 2.A.2

La matrice Q qui résous le problème de l'ACP est la matrice de vecteurs propres de V.

Lemme 2.A.3

*Si une matrice M s'écrit $M = A^{\mathsf{T}}A$, alors il existe une matrice $M^{1/2}$ qui satisfait $M^{1/2}M^{1/2} = M$. De plus, $\mathrm{Det}(M^{1/2}) = (\mathrm{Det}M)^{1/2}$, ainsi $M^{1/2}$ est inversible (*i.e. *son déterminant est non-nul) exactement lorsque M l'est. L'inverse $M^{1/2}$ est noté $M^{-1/2}$, et la règle de calcul $M^a M^b = M^{a+b}$ est valable pour n'importe quel demi-entier.*

13. Dans les situations présentes dans ce cours, les valeurs propres sont toujours positives, ainsi il ne s'agira toujours que de dilatations.

DÉMONSTRATION: C'est une conséquence du théorème spectral sur les matrices symétriques : si $M = A^\mathsf{T} A$ alors les λ_i sont toutes ≥ 0. Ensuite il faut remarquer que le produit de matrices diagonales D_1 et D_2 est une matrice diagonale D dont l'entrée sur la $i^{\text{ème}}$ ligne et $i^{\text{ème}}$ colonne est le produit des entrées diagonales correspondantes : $(D)_{ii} = (D_1)_{ii}(D_2)_{ii}$.

Ainsi, la matrice Λ du théorème 2.A.1 possède une matrice $\Lambda^{1/2}$ "évidente" :

$$\Lambda^{1/2} = \begin{pmatrix} \lambda_1^{1/2} & 0 & \dots & 0 & 0 \\ 0 & \lambda_2^{1/2} & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & \lambda_{n-1}^{1/2} & 0 \\ 0 & 0 & \dots & 0 & \lambda_n^{1/2} \end{pmatrix}$$

En effet, comme tous les $\lambda \geq 0$, il est possible de prendre la racine carrée. Ensuite, on pose

$$M^{1/2} = Q\Lambda^{1/2}Q^\mathsf{T}.$$

Un petit calcul justifie ce choix :

$$M^{1/2}M^{1/2} = Q\Lambda^{1/2}Q^\mathsf{T}Q\Lambda^{1/2}Q^\mathsf{T} = Q\Lambda^{1/2}\mathrm{Id}\Lambda^{1/2}Q^\mathsf{T} = Q\Lambda^{1/2}\Lambda^{1/2}Q^\mathsf{T} = Q\Lambda Q^\mathsf{T} = M.$$

La relation sur le déterminant peut être vue de plusieurs manières. Par exemple, en utilisant la multiplicativité du déterminant (*cf.* le lemme 1.2.34), on a

$$\mathrm{Det}M = \mathrm{Det}(M^{1/2}M^{1/2}) = \mathrm{Det}(M^{1/2})(\mathrm{Det}M^{1/2}) = \left(\mathrm{Det}M^{1/2}\right)^2$$

Ainsi, Det$M$ est ≥ 0 et Det$M^{1/2}$ est nul exactement lorsque Det$M$ l'est. En particulier, $M^{1/2}$ est inversible exactement lorsque M l'est (*cf.* le lemme 1.2.35). ■

Lemme 2.A.4

(Voir [6, p.74]) Soit A une matrice symmétrique de taille $m \times m$ et C une matrice définie positive (aussi de taille $m \times m$). Soit $\lambda_1 \geq \cdots \geq \lambda_m$ les racines du polynôme $\mathrm{Det}(A - \lambda C) = 0$. *Alors*

$$\sup_{\boldsymbol{x}\in\mathbb{R}^m} \frac{\boldsymbol{x}^\mathsf{T} A\boldsymbol{x}}{\boldsymbol{x}^\mathsf{T} C\boldsymbol{x}} = \sup_{\boldsymbol{x}^\mathsf{T} C\boldsymbol{x}=1} \boldsymbol{x}^\mathsf{T} A\boldsymbol{x} = \lambda_1. \quad \textit{et} \quad \inf_B \sup_{B^\mathsf{T} C\boldsymbol{x}=0} \frac{\boldsymbol{x}^\mathsf{T} A\boldsymbol{x}}{\boldsymbol{x}^\mathsf{T} C\boldsymbol{x}} = \lambda_{k+1}.$$

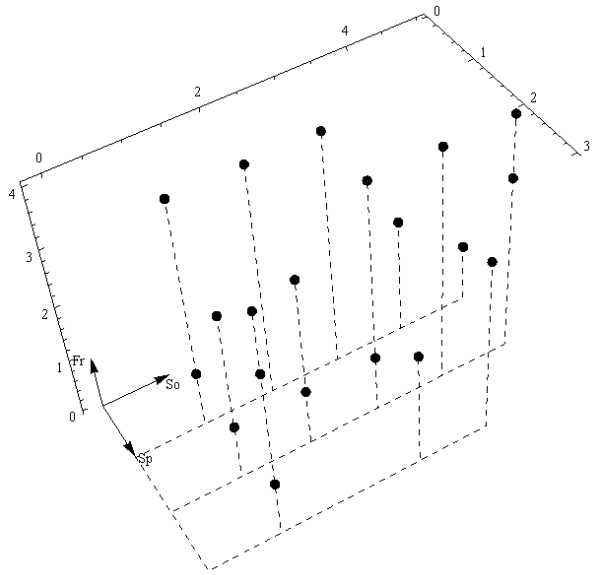
où B est une matrice $m \times k$.

2.B Exemples d'ACP

2.B.i Exemple 1 : Un exemple d'ACP en trois dimension

no de l'individu	sport	sortie	fromage
1	2	3	1
2	2	5	4
3	2	2	1
4	2	2	3
5	2	1	3
6	3	1	1
7	1	5	1
8	2	1	1
9	2	5	3
10	2	1	3
11	3	3	2
12	1	4	2
13	1	1	1
14	3	1	4
15	2	3	4
16	1	2	4
17	2	2	3
18	1	1	4
19	1	4	2
20	3	1	3
21	2	1	3
22	1	3	4
23	2	4	4
24	2	1	1
25	3	4	3

On catégorifie 25 individus selon la fréquence de trois activité : une sportive, une sortie et manger du fromage.



Le dessin à droite est la représentation des individus en fonction des trois mesures dans l'espace à trois dimension. De chaque variable (sport/sortie/fromage), on tire les données habituelles :

\	sport	sortie	fromage
moyenne	1.92	2.44	2.60
médiane	2	2	3
mode	2	1	3
étendue	2	4	3
variance($\simeq$)	0.474	2.086	1.360
écart-type($\simeq$)	0.688	1.444	1.166
étend./éc.-type($\simeq$)	2.906	2.769	2.572

Ici, la seconde variable semble peu fiable : sa moyenne est éloignée du mode. De fait, les deux valeurs les plus fréquentes sont des 1 et des 4, alors qu'en général il est préférable (pour des raisons de modèle statistique) d'avoir des mesures qui sont relativement centrées.

Idéalement, l'étendue divisée par l'écart-type reste inférieur à 6. Cela dépend néanmoins de la taille de l'échantillon : si il est très grand (*e.g.* 100 000) il n'est pas anormal que ce rapport excède 8.

Ceci permet d'écrire la matrice centrée réduite, si X est le contenu du tableau ci-haut et $\bar{X}$ la matrice dont la première colonne est remplie de 1.92, la seconde de 2.44 et la troisième de 2.6, alors les variables centrées sont :

$$X - \bar{X} = \begin{pmatrix} 0,08 & 0,56 & -1,60 \\ 0,08 & 2,56 & 1,40 \\ 0,08 & -0,44 & -1,60 \\ 0,08 & -0,44 & 0,40 \\ 0,08 & -1,44 & 0,40 \\ 1,08 & -1,44 & -1,60 \\ -0,92 & 2,56 & -1,60 \\ 0,08 & -1,44 & -1,60 \\ 0,08 & 2,56 & 0,40 \\ 0,08 & -1,44 & 0,40 \\ 1,08 & 0,56 & -0,60 \\ -0,92 & 1,56 & -0,60 \\ -0,92 & -1,44 & -1,60 \\ 1,08 & -1,44 & 1,40 \\ 0,08 & 0,56 & 1,40 \\ -0,92 & -0,44 & 1,40 \\ 0,08 & -0,44 & 0,40 \\ -0,92 & -1,44 & 1,40 \\ -0,92 & 1,56 & -0,60 \\ 1,08 & -1,44 & 0,40 \\ 0,08 & -1,44 & 0,40 \\ -0,92 & 0,56 & 1,40 \\ 0,08 & 1,56 & 1,40 \\ 0,08 & -1,44 & -1,60 \\ 1,08 & 1,56 & 0,40 \end{pmatrix}$$

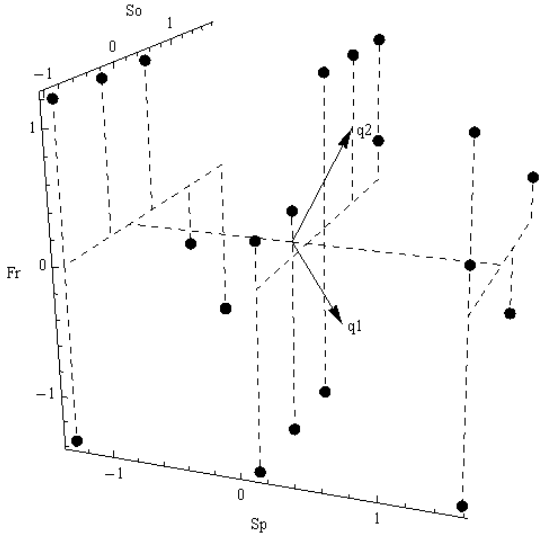
La matrice des variances V est alors donnée par

$$\begin{aligned} V &= \tfrac{1}{25}(X-\bar{X})^{\mathrm{t}}(X-\bar{X}) \\ &\simeq \begin{pmatrix} 0.474 & -0.205 & 0.008 \\ -0.205 & 2.086 & 0.176 \\ 0.008 & 0.176 & 1.360 \end{pmatrix} \end{aligned}$$

Il serait possible de poursuivre avec cette matrice, cependant il est dans ce cas-ci préférable d'utiliser les variables centrées réduites (*i.e.* la matrice des corrélations). En effet, comme les variables ne sont ni comparable en unité et ni en sens, il vaut mieux passer par les variables centrées réduites Z.

Ici, il est difficile de justifier une quelconque homogénéité dans les différentes mesure ; d'où l'emploi de variables centrées réduites. On rappelle que la $i^{\text{ème}}$ colonne de Z, notée $\mathbf{z}_{(j)}$ est donnée par $\mathbf{z}_{(j)} = \frac{\mathbf{x}_{(j)} - \widehat{\mathrm{x}}}{\sigma_j}$ (où σ_j est l'écart-type associé à $\mathbf{x}_{(j)}$). Ceci donne :

$$Z \simeq \begin{pmatrix} 0,116 & 0,388 & -1,372 \\ 0,116 & 1,772 & 1,200 \\ 0,116 & -0,305 & -1,372 \\ 0,116 & -0,305 & 0,343 \\ 0,116 & -0,997 & 0,343 \\ 1,569 & -0,997 & -1,372 \\ -1,337 & 1,772 & -1,372 \\ 0,116 & -0,997 & -1,372 \\ 0,116 & 1,772 & 0,343 \\ 0,116 & -0,997 & 0,343 \\ 1,569 & 0,388 & -0,514 \\ -1,337 & 1,080 & -0,514 \\ -1,337 & -0,997 & -1,372 \\ 1,569 & -0,997 & 1,200 \\ 0,116 & 0,388 & 1,200 \\ -1,337 & -0,305 & 1,200 \\ 0,116 & -0,305 & 0,343 \\ -1,337 & -0,997 & 1,200 \\ -1,337 & 1,080 & -0,514 \\ 1,569 & -0,997 & 0,343 \\ 0,116 & -0,997 & 0,343 \\ -1,337 & 0,388 & 1,200 \\ 0,116 & 1,080 & 1,200 \\ 0,116 & -0,997 & -1,372 \\ 1,569 & 1,080 & 0,343 \end{pmatrix}$$



Le dessin de droite représente de nouveau le "nuages" des mesures, mais cette fois-ci exprimé en variables centrées réduites (en particulier, l'origine des axes est au centre du nuage).

Ensuite on regarde la matrice

$$R = \frac{1}{25} Z^{\mathrm{t}} Z \simeq \begin{pmatrix} 1.000 & -0.206 & 0.010 \\ -0.206 & 1.000 & 0.105 \\ 0.010 & 0.105 & 1.000 \end{pmatrix}.$$

La trace d'une matrice R (qui est essentiellement la matrice V obtenu en commençant avec Z pour données initiales) est toujours égale à sa taille (ses coefficients diagonaux sont tous 1). Ainsi, la variabilité totale pour Z est toujours le nombre de variable, soit dans cet exemple 3.

Résoudre le polynôme d'ordre 3, $\mathrm{Det}(R - \lambda \mathrm{Id})$ en fonction de λ donne les trois valeurs propres des composantes principales. Ici, on trouve (en arrondissant)

$$1.23, 1.01, \text{ et } 0.77$$

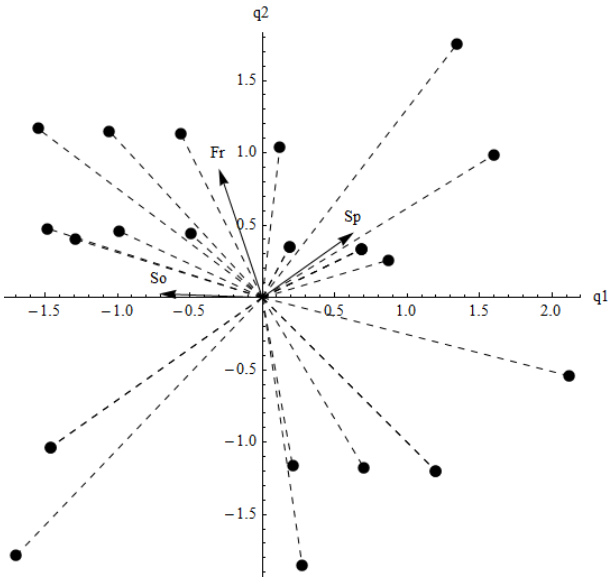
Ainsi, les deux premières valeurs propres donnent $\frac{2.24}{3} \simeq .75 = 75\%$ de la variation. Résoudre

$$(R - 1.23\mathrm{Id}_{3\times 3}) \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

Donne que le premier vecteur propre (la nouvelle variable) est $\mathbf{q}^{\mathrm{t}}_{(1)} \simeq (0.63, -0.71, -0.30)$ (ou, très approximativement, une proportion de 6 sport moins 7 sortie moins 3 fromage).

De la même façon, $\mathbf{q}^{t}_{(2)} \simeq (0.45, 0.03, 0.89)$ (ou, toujours très approximativement, une proportion de 1 sport plus 2 fromage) est la seconde nouvelle variable.

Dans ces nouvelles variables, les données se réduisent au portrait suivant :



2.B.ii Exemple 2 : Un exemple d'ACP à ne pas suivre

Cet exemple illustre le genre de problème qui peut arriver si les composantes principales d'une ACP représentent une trop faible proportion de la variabilité

Dans un cours 25 étudiants ont rempli un formulaire, représentant la fréquence avec laquelle il pratiquait une certaine activité, mangeait un fromage, etc.... Il y avait un total de 17 variables (pour 25 individu) : comme vous vous en doutez c'est très mauvais d'un point de vue méthodologique. Cet exemple est évidemment voué à l'échec ; les données sont trop faibles [14] : trop de variables pour trop peu d'individus.

Un calcul des valeurs propres donne :

$\frac{\lambda_1}{17}$=0.180921 $\frac{\lambda_2}{17}$=0.13301 $\frac{\lambda_3}{17}$=0.124779 $\frac{\lambda_4}{17}$=0.109466
$\frac{\lambda_5}{17}$=0.100226 $\frac{\lambda_6}{17}$=0.0702869 $\frac{\lambda_7}{17}$=0.066662 $\frac{\lambda_8}{17}$=0.0636928
$\frac{\lambda_9}{17}$=0.0450723 $\frac{\lambda_{10}}{17}$=0.0356809 $\frac{\lambda_{11}}{17}$=0.0310489 $\frac{\lambda_{12}}{17}$=0.0179046
$\frac{\lambda_{13}}{17}$=0.0115822 $\frac{\lambda_{14}}{17}$=0.00773748 $\frac{\lambda_{15}}{17}$=0.00186055 $\frac{\lambda_{16}}{17}$=0.0000750771
$\frac{\lambda_{17}}{17}$=0.00000697046.

[La division par 17, vient du fait que, ayant 17 variables, la variabilité totale avec des variables centrées réduite est le nombre de variables, soit 17.] Ainsi, les deux premières valeurs propres expliquent environs 31.4% de la variabilité, les 3 premières 43.8%, les 4 premières 54.8%, et les 5 premières 64.8%. Il est évident difficile de dessiner 5 variables...

14. Ici, non seulement les données sont faibles, mais certaines sont traitées de manière incorrecte : lorsqu'une variable ne peut prendre que deux valeurs possibles, il ne faut pas la traiter avec une ACP !

La dernière page donne les résultats de l'ACP avec les trois premières valeurs propres, la représentation est une simple projection sur l'espace engendré par les valeurs propres (*i.e.* un bi-graphe des distances). Le nom des variables (et donc des vecteurs) sont

age	l'âge	fratrie	nombre frère/soeur	racle	raclette	conc	concert
patin	patin	FH	Femme ou Homme	skialp	ski alpin	tomme	tomme
biere	bière	skifon	ski de fond	cine	cinéma	gruye	gruyère
theat	théâtre	ae	dernière lettre prénom	appenz	appenzeller	danse	danse
		hasard	nombre au hasard				

Les trois première vues sont faites essentiellement pour regarder le dessin sans trop considérer un des axes (qui est alors une profondeur). Les axes sont numérotés q_1, q_2 et q_3 (correspondant à la première, seconde et troisième valeur propre). Les points plus profonds sont plus petits et plus sombre. Dans le dernier dessin, le point de vue n'est pas aussi perpendiculaire aux autres axes. En particulier, pour mieux mettre en valeur la variation selon q_1 et q_3, les points sont plus lumineux quand q_1 est grand (sombre quand q_1 petit, *i.e.* négatif) et change de couleur quand q_3 varie (rouge pour q_3 grand et vert pour q_3 petit).

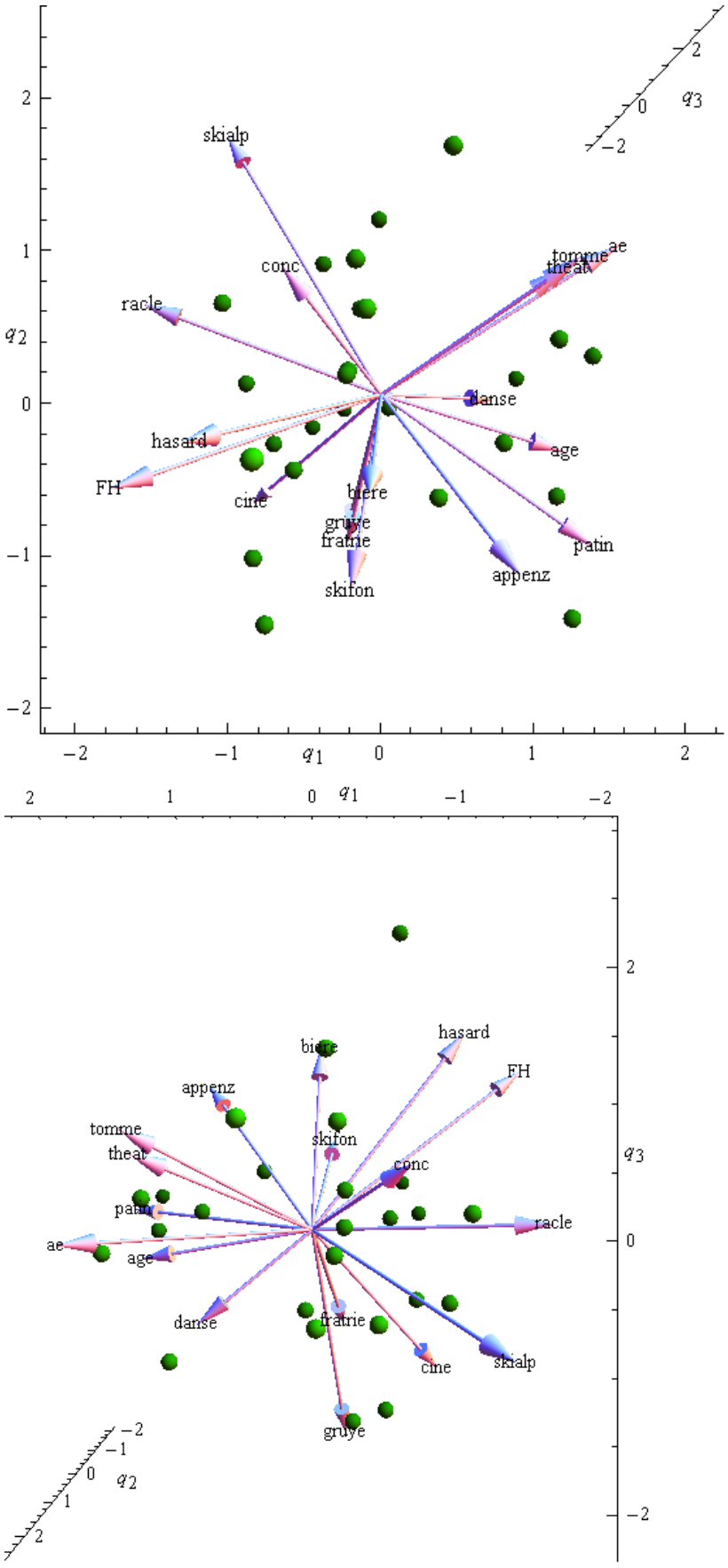
Remarquer la particularité très importante suivante : dans le premier dessin (axes 1 et 2) la bière, le ski de fond, la fratrie et le gruyère pointent dans la même direction. Dans le second, la bière et le ski de fond pointent dans la même direction mais (! ! !) leur direction est opposée à celle du gruyère et de la fratrie. Finalement, dans le troisième la bière et le ski de fond pointent légèrement dans la même direction, la bière est orthogonale à la fratrie, le ski de fond est orthogonal au gruyère, etc...

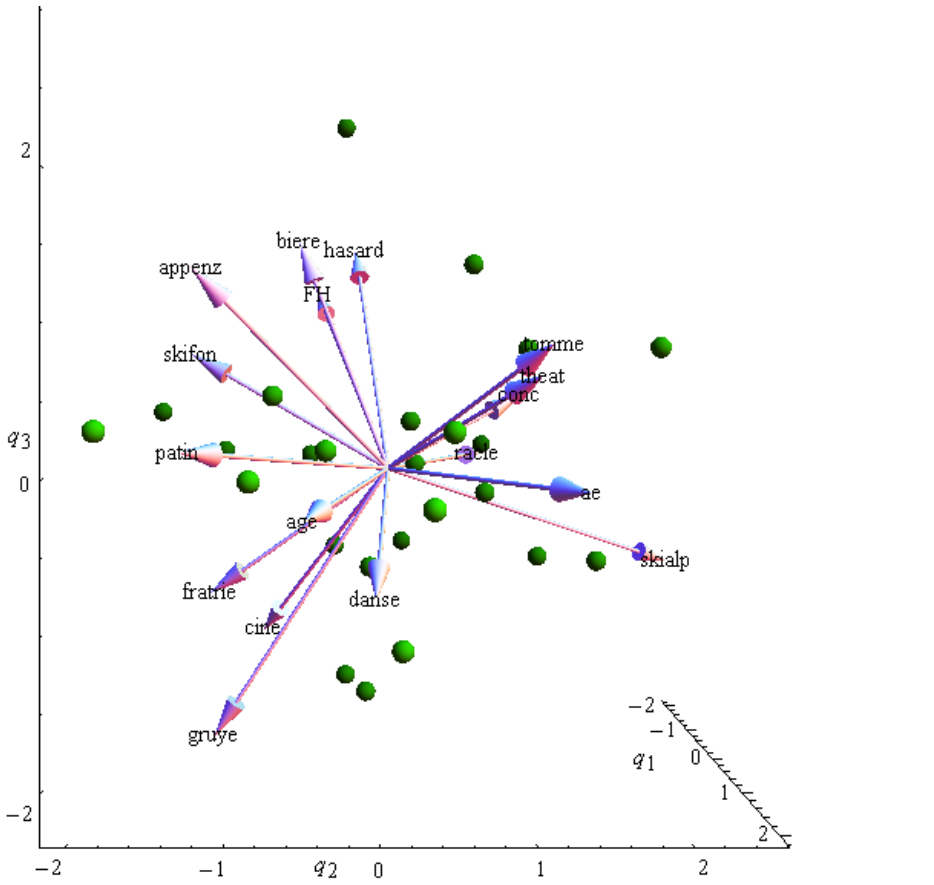
Pourquoi tant de changements ? C'est que deux des trois premières composantes principales expliquent en fait très peu de la variabilité et l'ajout de la composante manquante est suffisant pour tout perturbé. De fait, comme même les trois premières valeurs ne donne que 43.8%, il n'est pas dit que considérer les autres composantes pourrait changer la donne. Dans ce cas-ci, les variations correspondant aux valeurs propres λ_4, λ_5, λ_6, λ_7 et λ_8 représentent 45.5% de la variabilité totale : c'est plus grand que les trois premières valeurs propres ! Ceci tient beaucoup au fait que le nombre de variables est très grand en comparaison au nombre d'individus (et que les variables sont assez aléatoires).

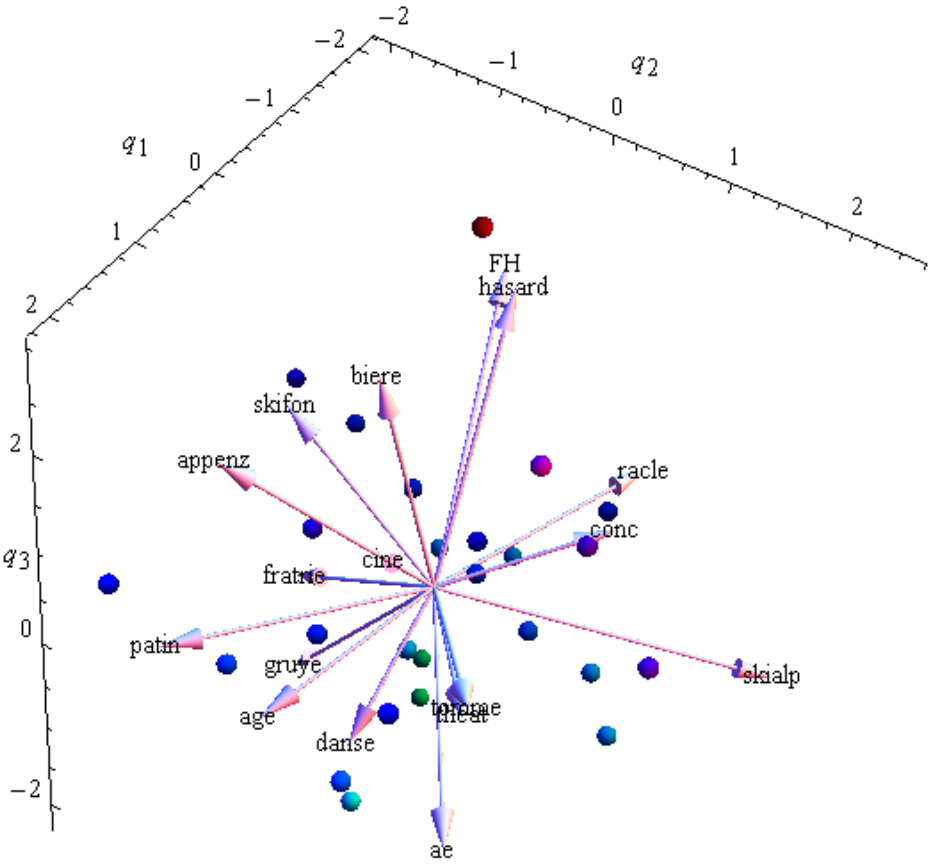
Finalement, ceci est aussi une des raisons pour lesquelles il faut éviter de tirer des conclusions trop fortes lorsque les axes choisis dans l'ACP expliquent trop peu (*e.g.* moins de 75%) de la variation [15]. Dans les cas où la variation est inférieure, il est toujours possible de continuer (avec néanmoins des précautions sur les conclusions) si les prochaines valeurs propres sont très basses.

Comme exercice, vous pouvez vous amuser à tirer des conclusions (à ne pas prendre trop au sérieux) du résultat de cette ACP.

15. Ce qui est admis dépend beaucoup de la communauté scientifique dans laquelle on se trouve.


skialp
conc
racle
tomme
theatv
ae
danse
age
hasard
FH
cine
biere
gruye
fratrie
patin
appenz
skifon
q_1
q_2
q_3
hasard
biere
FH
appenz
skifon
tomme
theat
conc
racle
patin
ae
age
danse
fratrie
cine
skialp
gruye


biere
hasard
appenz
FH
skifon
conc
patin
racle
age
skialp
fratrie
danse
cine
gruye
q_1
q_2
q_3


FH
hasard
biere
skifon
appenz
racle
conc
cine
fratrie
patin
gruye
skialp
age
danse
ae
q_1
q_2
q_3

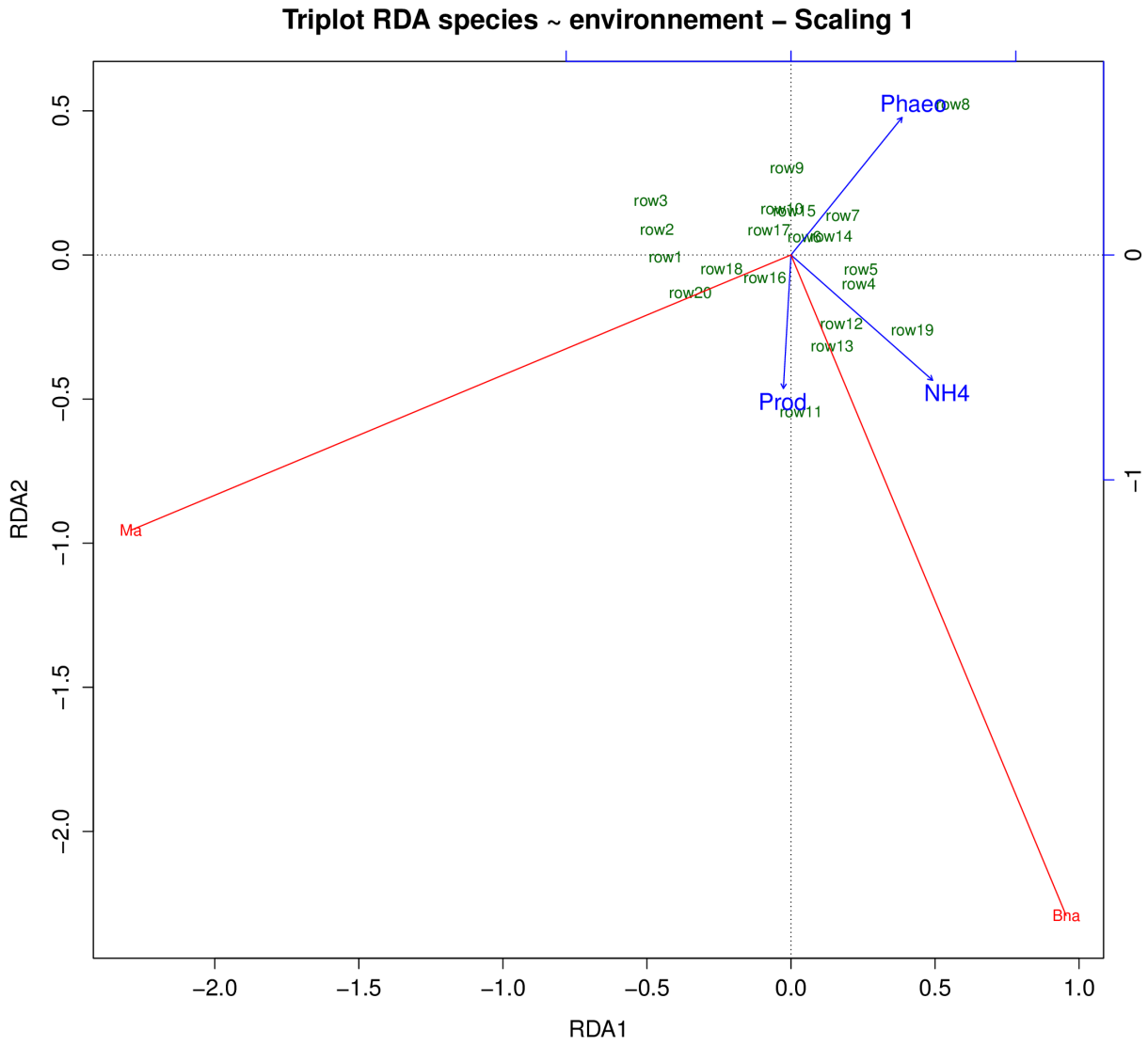
2.B.iii Exemple 3 : une RDA de petite dimension

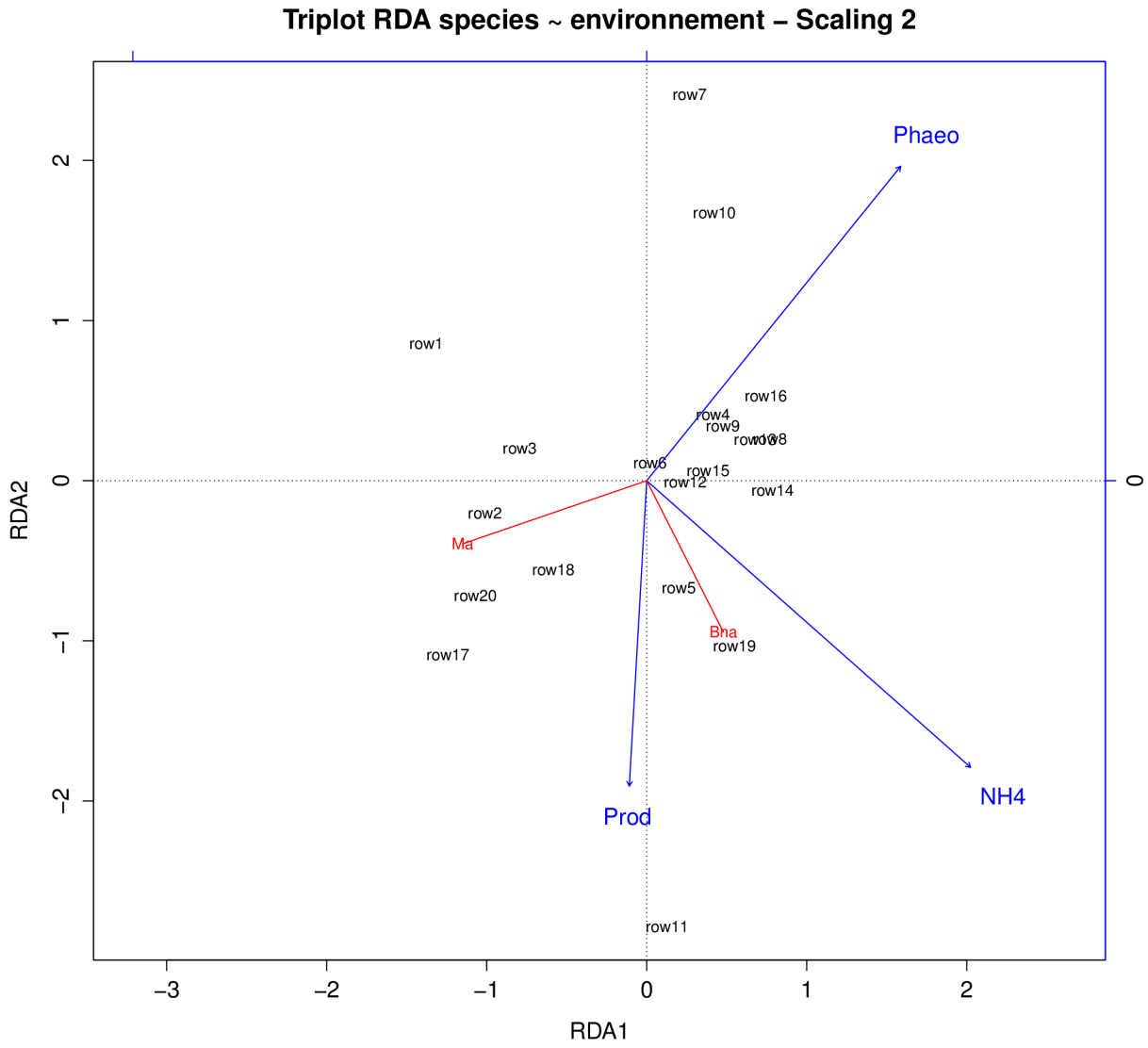
Data collected at 20 sites in the Thau lagoon on 25 October 1988, cf. [5] :

• Two bacterial variables (*response*) : *Bna*, the concentration of colony-forming units of aerobic heterotrophs growing on bioMérieux nutrient agar, with low NaCl concentration ; and *Ma*, the concentration of aerobic heterotrophs growing on marine agar at 34 gL^{-1} salinity ;

• three environmental variables (NH_4 in the water column, in $\mu molL^{-1}$; phaeopigments from degraded chlorophyll *a*, in μgL^{-1} ; and bacterial production, determined by incorporation of tritiated thymidine in bacterial DNA, in $nmolL^{-1}d^{-1}$).

R|E	Bna	Ma	NH4	Phaeo	Prod
1	4.62	10.00	0.31	0.18	0.27
2	5.23	10.00	0.21	0.21	0.21
3	5.08	9.64	0.14	0.23	0.13
4	5.28	8.33	1.37	0.29	0.18
5	5.76	8.93	1.45	0.24	0.09
6	5.33	8.84	0.67	0.53	0.27
7	4.26	7.78	0.30	0.95	0.46
8	5.44	8.02	0.33	1.39	0.25
9	5.33	8.29	0.21	0.76	0.23
10	4.66	7.88	0.22	0.74	0.36
11	6.78	9.74	0.79	0.45	0.82
12	5.44	8.66	1.11	0.40	0.42
13	5.42	8.12	1.27	0.25	0.40
14	5.60	8.12	0.96	0.45	0.17
15	5.44	8.49	0.71	0.46	0.14
16	5.30	7.96	0.64	0.39	0.36
17	5.60	10.54	0.52	0.48	0.26
18	5.50	9.69	0.25	0.47	0.45
19	6.02	8.70	1.66	0.32	0.29
20	5.46	10.24	0.18	0.38	0.51


Triplot RDA species ~ environnement – Scaling 1
RDA2
RDA1
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
0
−1
Phaeo
NH4
Prod
Ma
Bna
row1
row2
row3
row4
row5
row7
row8
row9
row11
row12
row13
row14
row15
row16
row17
row18
row19
row20



2.B.iv Exemple 4 : AFC

Exemple d'une compagnie qui veux déterminer l'état de santé de son personnel [3] :

	0-6 visites	7-12 visites	>12 visites
cadre>40	5	7	3
cadre<40	5	5	2
employé>40	2	12	6
employé<40	24	18	12
RH	12	8	4

Colonnes :

	0-6 visites	7-12 visites	>12 visites	Profil moyen
cadre>40	0.10	0.14	0.11	0.12
cadre<40	0.10	0.10	0.07	0.10
employe>40	0.04	0.24	0.22	0.16
employe<40	0.50	0.36	0.44	0.43
RH	0.25	0.16	0.15	0.19

d'où la proximité entre "7-12 visites" et ">12 visites"

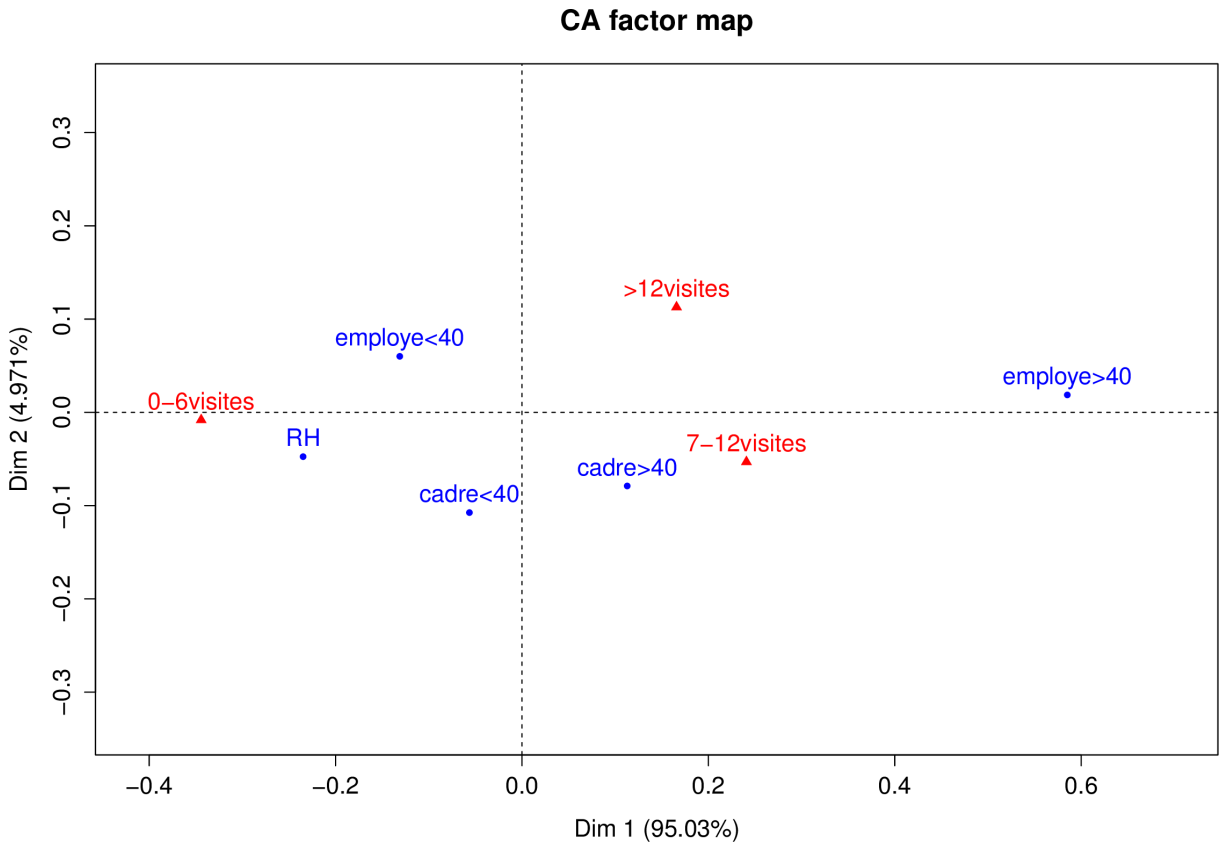
Lignes :

	0-6 visites	7-12 visites	>12 visites
cadre>40	0.33	0.47	0.20
cadre<40	0.42	0.42	0.17
employe>40	0.10	0.60	0.30
employe<40	0.44	0.33	0.22
RH	0.50	0.33	0.17
Profil moyen	0.38	0.40	0.22

d'ou la proximité entre "employé<40" et "RH" ; et pas du tout entre "cadre>40" et "employé>40". Du point de vue croisé, les "employé<40" et "RH" sont proches de "0-6 visites" (car 44% et 50% d'entre eux favorisent peu de visites médicales).

Contributions des cases à l'écart de l'indépendance (pourcentages) :

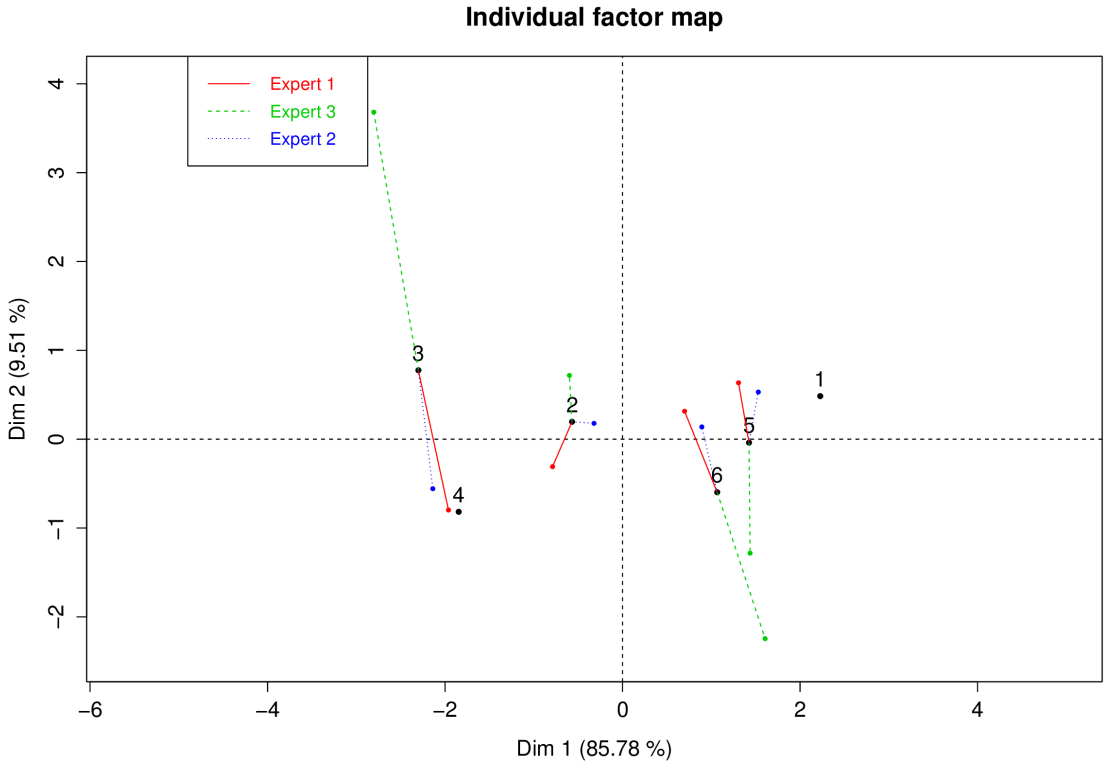
	0-6 visites	7-12 visites	> 12 visites
cadre>40	1.02	1.70	0.18
cadre<40	0.34	0.08	1.38
employé>40	42.79	20.37	6.65
employé<40	5.23	6.11	0.10
RH	8.57	2.72	2.75

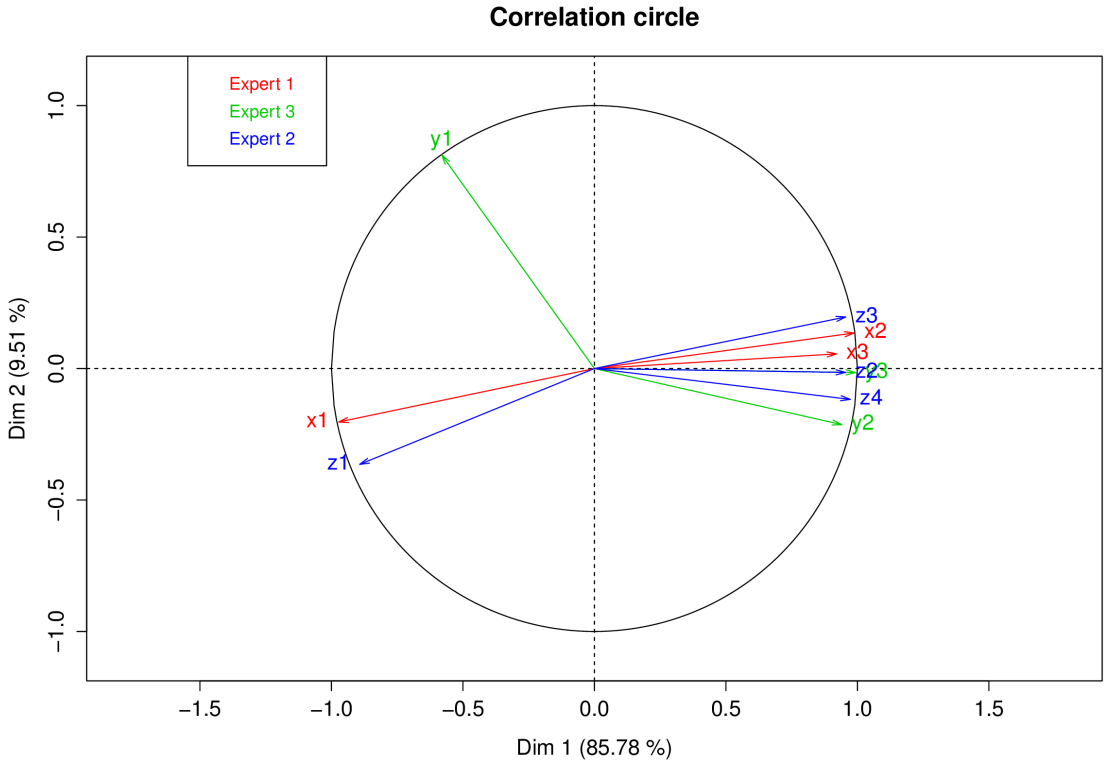


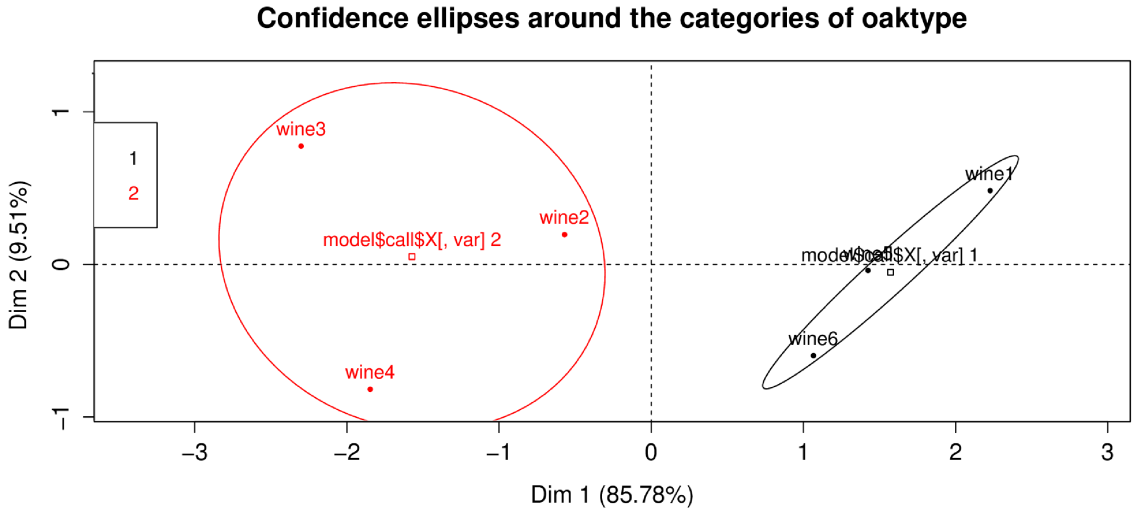
2.B.v Exemple 5 : MFA

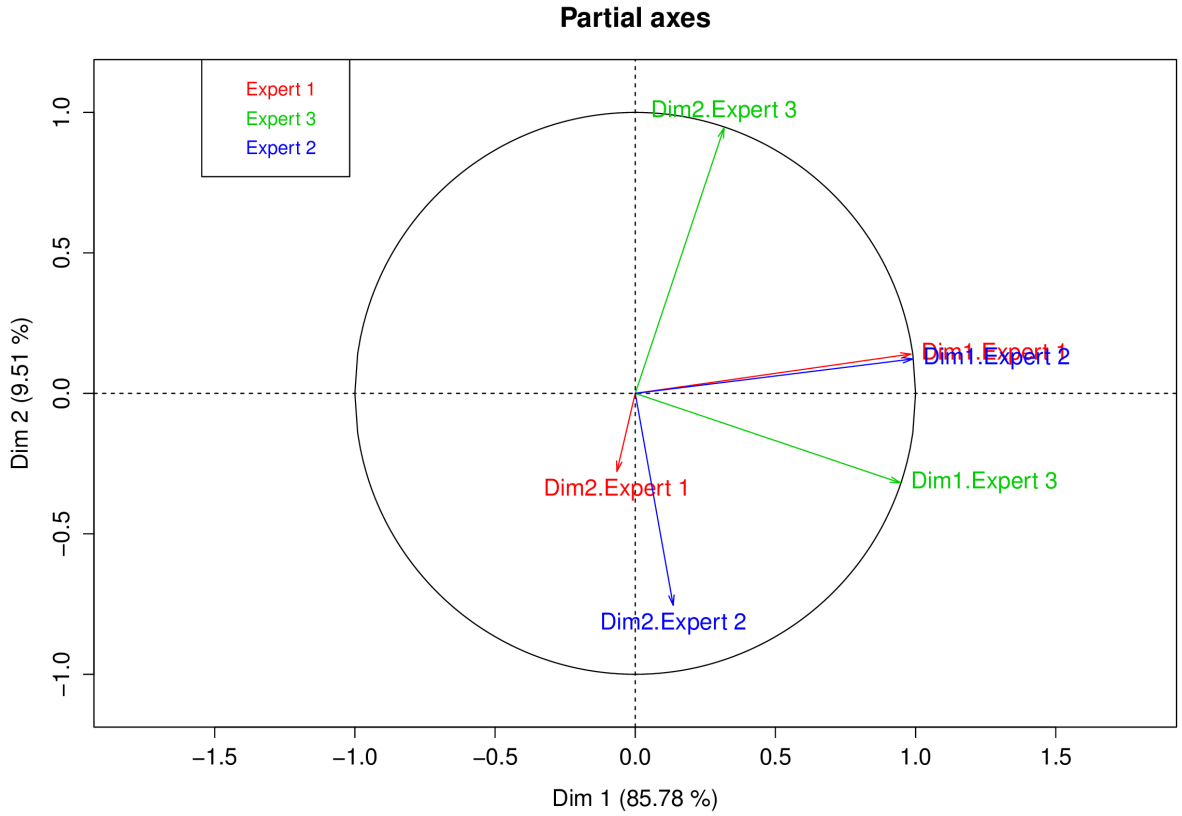
Les notes (sur une échelle de 1 a 9) données par 3 experts sur 6 vins. On cherche une typologie de vins et s'il y a un accord entre les experts [1].

	fruity	woody	coffee	fruity	butter	woody	red-fruit	roast	vanilla	woody	Oak
$wine_1$	1	6	7	3	6	7	2	5	7	6	1
$wine_2$	5	3	2	4	4	3	4	4	4	2	2
$wine_3$	6	1	1	7	1	1	5	2	1	1	2
$wine_4$	7	1	2	2	2	2	7	2	1	2	2
$wine_5$	2	5	4	2	6	6	3	5	6	5	1
$wine_6$	3	4	4	1	7	5	3	5	4	5	1


Individual factor map
Expert 1
Expert 3
Expert 2
Dim 2 (9.51 %)
Dim 1 (85.78 %)
1
2
3
4
5
6


Correlation circle
Expert 1
Expert 3
Expert 2
Dim 2 (9.51 %)
Dim 1 (85.78 %)
y1
z3
x2
x3
z2
z4
y2
x1
z1


Confidence ellipses around the categories of oaktype
1
2
wine3
wine2
wine4
wine1
wine6
model$call$X[, var] 2
model$call$X[, var] 1
Dim 2 (9.51%)
Dim 1 (85.78%)


Partial axes
Expert 1
Expert 3
Expert 2
Dim2.Expert 3
Dim1.Expert 1
Dim1.Expert 2
Dim1.Expert 3
Dim2.Expert 1
Dim2.Expert 2
Dim 2 (9.51 %)
Dim 1 (85.78 %)

Chapitre 3

Méthodes de classifications : arbre supervisés et non-supervisés

L'objectif général cette fois-ci est de tenter de regrouper les individus en groupes assez homogènes.

3.1 Distance et dissimilarité

Les notions fondamentales seront donc celles de dissimilarité et de distance. Comme leur nom l'indique, à chaque paire d'individus, on cherche à attacher un nombre réel qui témoigne à quel points ces individus sont écartés. On définira d'abord cette notion entre les individus, puis entre les groupes d'individus.

3.1.i ...entre individus

Comme précédemment, les individus sont notés par un numéro $i = 1, \ldots, n$.

Définition 3.1.1. Une **dissimilarité** est une fonction d, qui associe à deux individus $i, j \in \{1, \ldots, n\}$ un nombre réel. Elle satisfait les conditions suivantes

D1- $d(i,i) = 0$

D2- $d(i,j) = d(j,i)$

D3- $d(i,j) \geq 0$.

Si, de plus, elle satisfait l'inégalité du triangle :

D4- $d(i,j) \leq d(i,k) + d(k,j)$

d sera appelé une **distance**. ★

La dissimilarité (ou distance) qui sera choisie dépend énormément du goût du type d'expérience qu'on fait. Il est en fait recommandé d'essayer plusieurs distances différentes et de voir quelle partie du résultat de la classification est stable.

Lorsque les données sont de type présence/absence d'un caractère, il y a deux dissimilarités classiques. La matrice des données s'écrira toujours X et les données qui correspond au $i^{\text{ème}}$ individu se trouve à la $i^{\text{ème}}$ ligne de la matrice X. Elles donnent le vecteur ligne $\underline{x}_{(i)} = (x_{i1}, x_{i2}, \ldots, x_{ip})$.

Exemple 3.1.2. On examine plusieurs d'animaux de petite taille avec les variables suivantes :

Var1 : présence/absence de mandibules ;

Var2 : présence/absence d'ailes ;

Var3 : 6 ou 8 pattes.

La matrice des données a la forme :

$$X = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

(C'est une caractérisation pour le moins naïve...) ♣

À partir d'une telle matrice on tire les nombres suivants :

— a_{ij} le nombre de variables qui ont la même valeur chez i que chez j.

— b_{ij} le nombre de variables qui ont une valeur différente chez i et chez j.

— d_{ij} le nombre de variables qui ont la valeur 0 chez i et chez j.

Il y a alors 2 dissimilarités "classiques" (parmi un grand nombre) :

La dissimilarité de Jaccard entre deux individus est $d_J(i,j) = 1 - \dfrac{a_{ij}}{a_{ij}+b_{ij}} = \dfrac{b_{ij}}{a_{ij}+b_{ij}}$.

La dissimilarité de Russel & Rao entre deux individus est $d_{RR}(i,j) = 1 - \dfrac{a_{ij}+d_{ij}}{a_{ij}+b_{ij}+d_{ij}} = \dfrac{b_{ij}}{a_{ij}+b_{ij}+d_{ij}}$.

Exemple 3.1.3. (suite de l'exemple 3.1.2) Voici quelqes calculs pour les dissimilarités ci-dessus (sur les lignes la paire d'individu, sur la colonne la dissimilarité) :

$\backslash$	$d_J(i,j) = 1 - \dfrac{a_{ij}}{a_{ij}+b_{ij}}$	$d_{RR}(i,j) = 1 - \dfrac{a_{ij}+d_{ij}}{a_{ij}+b_{ij}+d_{ij}}$
$(1,2)$	$1-\frac{2}{2+1} = \frac{1}{3}$	$1-\frac{2+0}{2+1+0} = \frac{1}{3}$
$(1,3)$	$1-\frac{1}{1+2} = \frac{2}{3}$	$1-\frac{1+0}{2+1+0} = \frac{2}{3}$
$(2,3)$	$1-\frac{2}{2+1} = \frac{1}{3}$	$1-\frac{2+1}{2+1+1} = \frac{1}{4}$
$(3,4)$	$1-\frac{2}{2+1} = \frac{1}{3}$	$1-\frac{2+1}{2+1+1} = \frac{1}{4}$
$(3,5)$	$1-\frac{0}{0+3} = 1$	$1-\frac{0+0}{0+3+0} = 1$
$(3,6)$	$1-\frac{2}{2+1} = \frac{1}{3}$	$1-\frac{2+2}{2+1+2} = \frac{1}{5}$

La dissimilarité R&R a tendance a rapproché les individus qui n'ont pas un caractère en commun.

♣

Plus généralement, on pourrait s'imaginer faire une dissimilarité avec n'importe quels $\alpha, \beta \in \mathbb{R}_{>0}$ et $\delta \in \mathbb{R}_{\geq 0}$ par $d_{\alpha,\beta,\delta}(i,j) = 1 - \dfrac{\alpha a_{ij} + \delta d_{ij}}{\alpha a_{ij} + \beta b_{ij} + \delta d_{ij}} = \dfrac{\beta b_{ij}}{\alpha a_{ij} + \beta b_{ij} + \delta d_{ij}}$.

Pour le cas des données qualitatives plus compliquées, il est (parmi plusieurs choix) possible d'utiliser des distances basées sur des loi du χ^2.

Il est aussi souhaitable de traiter les données quantitatives.

Exemple 3.1.4. Les notes de 12 étudiants à 3 cours sont mises dans un (matrice des données) :

$$X = \begin{pmatrix} 5.5 & 5 & 5.5 \\ 5 & 5 & 4.5 \\ 5 & 6 & 5.5 \\ 4.5 & 5.5 & 5 \\ 5 & 4.5 & 5 \\ 5.5 & 4.5 & 4 \\ 4.5 & 5 & 5 \\ 5.5 & 5.5 & 5 \\ 5 & 5.5 & 4 \\ 4.5 & 6 & 5.5 \\ 4.5 & 4.5 & 6 \\ 4.5 & 4 & 5 \end{pmatrix}$$

Les moyennes sont toutes les trois aux alentours de 5. ♣

À partir d'une telle matrice on tire des distances en mesurant la "longueur" du vecteur $\underline{\mathrm{x}}_{(i)-(j)} = \underline{\mathrm{x}}_{(i)} - \underline{\mathrm{x}}_{(j)}$.

— La distance Euclidienne : $d_E(i,j) = \|\underline{\mathrm{x}}_{(i)-(j)}\| = \|\underline{\mathrm{x}}_{(i)} - \underline{\mathrm{x}}_{(j)}\| = \sqrt{\sum_{k=1}^{p}(x_{ik} - x_{jk})^2}$.

— La distance de Manhattan : $d_M(i,j) = \|\underline{\mathrm{x}}_{(i)-(j)}\|_1 = \|\underline{\mathrm{x}}_{(i)} - \underline{\mathrm{x}}_{(j)}\|_1 := \sum_{k=1}^{p} |x_{ik} - x_{jk})|$.

— La distance de Mahalanobis : $d_V(i,j) = \|\underline{\mathrm{x}}_{(i)-(j)}\|_V := \sqrt{\underline{\mathrm{x}}_{(i)-(j)}^{\mathsf{T}} V_X^{-1} \underline{\mathrm{x}}_{(i)-(j)}}$.

Exemple 3.1.5. (suite de l'exemple 3.1.4) Voici quelques exemples de calculs :

i,j	vecteur $\mathbf{x}_{(i)-(j)}$	Manhattan	Euclidienne	Mahalanobis
1,2	(0.5 , 0.0 , 1.0)	1.5	1.1180	0.7737
1,3	(0.5 , -1.0 , 0.0)	1.5	1.1180	0.6410
1,4	(1.0 , -0.5 , 0.5)	2.0	1.2247	0.9675
1,5	(0.5 , 0.5 , 0.5)	1.5	0.8660	0.5716
1,6	(0.0 , 0.5 , 1.5)	2.0	1.5811	0.8144
3,5	(0.5 , 1.0 , 0.5)	2.0	1.2247	0.6837
3,6	(-0.5 , 0.5 , 0.5)	1.5	0.8660	0.4484
3,7	(0.0 , 0.5 , 1.5)	2.0	1.5811	0.8144

La méthode Manhattan est la plus simple à calculer, et celle de Mahalanobis la plus longue. (Mais pour un ordinateur ça ne change pas significativement les choses.) ♣

Parmi les autres distances imaginables, il est bon de mentionner la distance “d'édition” entre deux mots : c'est le plus petit nombre de caractères qu'il faut changer, ajouter ou supprimer pour passer d'un mot à l'autre : par exemple, “patate” est à distance 1 de “patte”, “patates” et à distance 2 de “potage”. Cette distance est utilisée par les correcteurs orthographiques.

Pour ceux qui font de la génétique, une distance naturelle (qui correspond aux changements qui sont possibles sur le code génétique) est définie comme suit : quatre types de mouvement sont possibles

— changer une lettre,
— ajouter une lettre,
— supprimer une lettre,
— et prendre un bloc de lettres de taille quelconque pour le remettre ailleurs.

On attribue à chacun de ces mouvements un coût. Puis on cherche le coût minimal pour passer d'une séquence à l'autre.

3.1.ii ...entre classes

L'idée de regrouper les individus est appelée mathématiquement faire une “partition” de l'ensemble des individus.

Définition 3.1.6. Une **partition** d'une ensemble $E = \{1,\dots,n\}$ est une division de cet ensemble en k sous-ensembles $C_i \subset E$, appelés **classes**, de sorte que

P1- tout élément appartient à une classe, *i.e.* la réunion des C_i soit tout E, *i.e.* $\cup_{i=1}^{k} C_i = E$;

P2- aucun élément n'appartient à deux classes, *i.e.* l'intersection de deux classes distinctes est vide, *i.e.* $i \neq j \Rightarrow C_i \cap C_j = \varnothing$.

★

Exemple 3.1.7. Si $E = \{1,2,3,4,5\}$ alors

— $C_1 = \{1,2\}$ et $C_2 = \{3,4,5\}$ est une partition de E.
— $C_1 = \{1\}$, $C_2 = \{2\}$, $C_3 = \{3\}$, $C_4 = \{4\}$ et $C_5 = \{5\}$ est une partition de E.
— $C_1 = \{1,3,5\}$ et $C_2 = \{2,4\}$ est une partition de E.
— $C_1 = \{1,2,3,4,5\}$ est une partition de E.
— $C_1 = \{1,5\}$, $C_2 = \{2,4\}$ et $C_3 = \{3\}$ est une partition de E.

Par contre, $C_1 = \{1,3,5\}$, $C_2 = \{2,4\}$ et $C_3 = \{3\}$ N'est PAS une partition de E (l'élément 3 appartient à deux classes). ♣

Le nombre de partition d'un ensemble à k éléments est très très grands. Plus précisément,

Théorème 3.1.8

Le nombre de partitions d'un ensemble à k éléments est $\frac{1}{e}\sum_{n=0}^{\infty}\frac{n^k}{n!}$.

En particulier, le nombre de partitions d'un ensemble à 9 éléments est 21 147 et celui d'un ensemble à 20 éléments 51 724 158 235 372 $\simeq 51 \times 10^{12}$! Ainsi, il n'est pas possible d'espère de pouvoir essayer toutes les partitions possibles, puis de prendre celle qui convient le mieux. La méthode

qui sera décrite ici est progressive (ce qui lui donnera la hiérarchie). On commence par rassembler les éléments les plus proches. Mais avant de pouvoir continuer, il faut parler de distance entre les classes.

De nouveau, il y a une pléthore de méthodes possibles, mais aucune n'a de raisons d'être choisies plutôt qu'une autre, même si certaines sont plus adaptées à certaines situations. De manière générale, il est préférable d'en utiliser plusieurs pour comparer.

Voici les trois plus classiques :

— Le saut minimum : $d(C_i,C_j) = \min_{a\in C_i, b\in C_j} d(a,b)$.

— Le saut maximum : $d(C_i,C_j) = \max_{a\in C_i, b\in C_j} d(a,b)$.

— Le saut moyen : $d(C_i,C_j) = \frac{1}{|C_i|}\sum_{a\in C_i}\frac{1}{|C_j|}\sum_{b\in C_j} d(a,b)$.

Exemple 3.1.9. Supposons qu'on ait 5 individus et une matrice des dissimilarités donnée par

$$D = \begin{pmatrix} 0 & 7 & 6 & 3 & 2 \\ 7 & 0 & 1 & 10 & 5 \\ 6 & 1 & 0 & 9 & 4 \\ 3 & 10 & 9 & 0 & 5 \\ 2 & 5 & 4 & 5 & 0 \end{pmatrix}$$

— Alors pour la partition $C_1 = \{1,2\}$ et $C_2 = \{3,4,5\}$, il n'y a qu'une distance à calculer. Elle est de

$$\begin{aligned} 1 &\quad \text{pour le saut minimum} \\ 10 &\quad \text{pour le saut maximum et} \\ \tfrac{6+3+2+1+10+5}{6} = \tfrac{27}{6} = 4.5 &\quad \text{pour le saut moyen} \end{aligned}$$

— Pour la partition $C_1 = \{1,3,5\}$ et $C_2 = \{2,4\}$, il n'y aussi qu'une seule distance non-nulle. Elle est de

$$\begin{aligned} 1 &\quad \text{pour le saut minimum} \\ 9 &\quad \text{pour le saut maximum et} \\ \tfrac{7+1+5+3+9+5}{6} = 5 &\quad \text{pour le saut moyen.} \end{aligned}$$

— Pour la partition $C_1 = \{1,5\}, C_2 = \{2,4\}$ et $C_3 = \{3\}$, les nouvelles matrices des distances seront (selon qu'on utilise le saut minimum, maximum ou moyen) :

(min)	C_1	C_2	C_3
C_1	0	3	4
C_2	3	0	1
C_3	4	1	0

(max)	C_1	C_2	C_3
C_1	0	7	6
C_2	7	0	9
C_3	6	9	0

(moy)	C_1	C_2	C_3
C_1	0	5	5
C_2	5	0	5
C_3	5	5	0

— Pour la partition $C_1 = \{1,2,3,4,5\}$, il n'y a pas de distance à calculer puisqu'il n'y a qu'une seule classe (sa distance à elle-même est toujours 0).

— Et pour $C_1 = \{1\}, C_2 = \{2\}, C_3 = \{3\}, C_4 = \{4\}$ et $C_5 = \{5\}$, la matrice des dissimilarités sera la même puisque chaque classe n'a qu'un individu.

♣

3.1.iii Méthode de Ward

...en construction...

3.2 Classification hiérarchique

Une classification est hiérarchique si elle permet de donner une hiérarchie. Cette hiérarchie est la plupart du temps représentée par un arbre (penser aux arbres phylogénétiques) En "coupant" l'arbre à une certaine hauteur, un certains nombres de groupes dont "l'homogénéité" est (au pire) une certaine quantité (qui dépend de la hauteur de la coupure).

La méthode pour produire une classification hiérarchique ici présentée est "générique". En effet, une fois qu'un choix de distance et de distance entre les classes est fait, elle est même totalement indépendante de tout ce qui précède.

Voici sans plus attendre l'algorithme :

Données : Une matrice des distances/dissimilarités et une méthode pour calculer les distances entre les classes.

Étape 0 : Tous les individus sont placé dans un classe qui ne contient qu'un individu.

Étape n : À la fin d'une étape n, une partition des individus est là.

Étape $n+1$: De la partition de l'étape n, une matrice des distances entre les classes est calculée, disons $D^{[n]}$. La plus petite entrée de $D^{[n]}$ est identifiée, disons d_{min}. Une des classes C_i et C_j telles que la distance entre C_i et C_j est d_{min} sont fusionnées (*i.e.* ces deux classes sont remplacées par une seule classe $C_{fus} = C_i \cup C_j$.

L'algorithme se termine lorsqu'il ne reste plus qu'une seule classe.

À noter que la première étape de l'algorithme est indépendante du type de distance entre les classes employée : en effet, chaque classe ne contient qu'un seul individu et il serait très bizarre que deux individus (considérés comme des classes) aient une distance différente que lorsque qu'ils sont considérés comme des individus.

La manière la plus grossière [1] d'en tirer un arbre est de faire comme suit : ...

Pour illustrer cet algorithme, les trois exemples suivants seront basés sur la même matrice (des dissimilarités, ce n'est pas la même qu'à l'exemple 3.1.9) :

$$D = \begin{pmatrix} 0 & 7 & 6 & 3 & 2 \\ 7 & 0 & 1 & 5 & 10 \\ 6 & 1 & 0 & 9 & 4 \\ 3 & 5 & 9 & 0 & 5 \\ 2 & 10 & 4 & 5 & 0 \end{pmatrix}$$

Exemple 3.2.1. D'abord, le saut minimum sera utilisé.

Étape 1 : (cette étape est indépendante du type de distance entre les classes employée !) chaque individu est dans une classe (*i.e.* $C_i = \{i\}$), la matrice des distance est ainsi D. La

1. Encore une fois, plusieurs méthodes existent.

plus petite distance est 1. Elle correspond aux deux éléments à plus petite distance : 2 et 3. Ils sont alors fusionnés, et les classes sont maintenant $\{1\},\{2,3\},\{4\}$ et $\{5\}$.

Étape 2 : Une nouvelle matrice des distances doit être calculée (selon la règle du saut minimum).

\	$\{1\}$	$\{2,3\}$	$\{4\}$	$\{5\}$
$\{1\}$	0	6	3	2
$\{2,3\}$	6	0	5	4
$\{4\}$	3	5	0	5
$\{5\}$	2	4	5	0

La plus petite entrée est 2, et elle correspond aux classes $\{1\}$ et $\{5\}$. Ces classes seront fusionnées pour donner les nouvelles classes $\{1,5\},\{2,3\}$ et $\{4\}$.

Étape 3 : La matrice des distance est

\	$\{1,5\}$	$\{2,3\}$	$\{4\}$
$\{1,5\}$	0	4	3
$\{2,3\}$	4	0	5
$\{4\}$	3	5	0

Sa plus petite entrée est le 3, et en conséquence les classes de $\{1,5\}$ et $\{4\}$ seront fusionnées. On a maintenant deux classes : $\{1,4,5\}$ et $\{2,3\}$.

Étape 4 : Il ne reste que deux classes, et donc une seule possibilité pour faire la fusion. Néanmoins, il est utile de connaître la distance qui les sépare : ici 4.

♣

Exemple 3.2.2. Le même exemple avec le saut maximum :

Étape 1 : Comme à l'exemple précédent : la plus petite distance (celle entre 2 et 3) est 1. Ils sont alors fusionnés, et les classes sont maintenant $\{1\},\{2,3\},\{4\}$ et $\{5\}$.

Étape 2 : Une nouvelle matrice des distances doit être calculée (selon la règle du saut maximum).

\	$\{1\}$	$\{2,3\}$	$\{4\}$	$\{5\}$
$\{1\}$	0	7	3	2
$\{2,3\}$	7	0	9	10
$\{4\}$	3	9	0	5
$\{5\}$	2	10	5	0

La plus petite entrée est (encore) 2, et elle correspond aux classes $\{1\}$ et $\{5\}$. Ces classes seront fusionnées pour donner les nouvelles classes $\{1,5\},\{2,3\}$ et $\{4\}$.

Étape 3 : La matrice des distance est

\	$\{1,5\}$	$\{2,3\}$	$\{4\}$
$\{1,5\}$	0	10	5
$\{2,3\}$	10	0	9
$\{4\}$	5	9	0

Sa plus petite entrée est le 5, et en conséquence les classes de $\{1,5\}$ et $\{4\}$ seront fusionnées. On a maintenant deux classes : $\{1,4,5\}$ et $\{2,3\}$.

Étape 4 : Il ne reste que deux classes, et donc une seule possibilité pour faire la fusion. Néanmoins, il est utile de connaître la distance qui les sépare : ici 10.

♣

Exemple 3.2.3. Finalement, avec le saut moyen :

Étape 1 : la plus petite distance (celle entre 2 et 3) est 1. Ils sont fusionnés, et les classes sont $\{1\},\{2,3\},\{4\}$ et $\{5\}$.

Étape 2 : Une nouvelle matrice des distances doit être calculée (selon la règle du saut moyen).

$\setminus$	$\{1\}$	$\{2,3\}$	$\{4\}$	$\{5\}$
$\{1\}$	0	6.5	3	2
$\{2,3\}$	6.5	0	7	7
$\{4\}$	3	7	0	5
$\{5\}$	2	7	5	0

La plus petite entrée est (encore) 2, et elle correspond aux classes $\{1\}$ et $\{5\}$. Ces classes seront fusionnées pour donner les nouvelles classes $\{1,5\},\{2,3\}$ et $\{4\}$.

Étape 3 : La matrice des distance est

$\setminus$	$\{1,5\}$	$\{2,3\}$	$\{4\}$
$\{1,5\}$	0	6.75	4
$\{2,3\}$	6.75	0	7
$\{4\}$	4	7	0

Sa plus petite entrée est le 4, et les classes $\{1,5\}$ et $\{4\}$ seront fusionnées. On a maintenant deux classes : $\{1,4,5\}$ et $\{2,3\}$.

Étape 4 : Il ne reste que deux classes, et donc une seule possibilité pour faire la fusion. Néanmoins, il est utile de connaître la distance qui les sépare : ici $6.8\overline{3}$.

♣

Dans la situation, le classement final est stable selon la méthode utilisée, ce qui est encourageant.

Bibliographie

[1] H. Abdi D. Valentin, Multiple Factor Analysis, in *Encyclopedia of Measurement and Statistics*, Sage, 2007.

[2] C. Chatfield and A.J. Collins, Introduction to Multivariate Analysis, Chapman and Hall, 1980.

[3] Y. Dodge, The Concise Encyclopedia of Statistics, Springer, 2008.

[4] A. Lazraq, R. Cleroux, *Testing the significance of the successive components in redundancy analysis*, Psychometrika **67** (2002), 411–419.

[5] P. Legendre and L. Legendre, Numerical Ecology, Elsevier 1998.

[6] C. R. Rao, Linear statistical inference and its applications. 2nd edition. Wiley, New York, 1973.

[7] Peter J. Rousseeuw, S*ilhouettes : A graphical aid to the interpretation and validation of cluster analysis*, Journal of Computational and Applied Mathematics, **20** (1987), 53–65.

[8] G. Saporta, Probabilités, analyse de donées et statistique. 3rd edition. Technip, 2011.

[9] A. L. Van den Wollenberg, *Redundancy analysis. An alternative for canonical correlation analysis*, Psychometrika **42** (1977), 207–219.

Index