ANALYSE DES DONNEES

Cours préparé par Djellali Abdelkhalek, Ph.D

Ce cours est destiné aux étudiants de troisième année électromécanique. Il s'agit d'un recueil de différentes publications modifiées et arrangées pour homogénéiser le contenu du module.

Les auteurs, à la base des communications ayant servi à la préparation de ce cours, appartiennent à différentes disciplines : On y trouve des mathématiciens, des statisticiens, des physiciens et des informaticiens.

Annaba, décembre 2020

ANALYSE DES DONNEES

I. GENERALITES SUR L'ANALYSE DE DONNEES

L'analyse des données est une famille de méthodes statistiques dont les principales caractéristiques sont d'être multidimensionnelles et descriptives. Certaines méthodes, pour la plupart géométriques, aident à faire ressortir les relations pouvant exister entre les différentes données et à en tirer une information statistique qui permet de décrire de façon plus succincte les principales informations contenues dans ces données. D'autres techniques permettent de regrouper les données de façon à faire apparaître clairement ce qui les rend homogènes, et ainsi mieux les connaître.

L'analyse des données permet de traiter un nombre très important de données et de dégager les aspects les plus intéressants de la structure de celles-ci. Le succès de cette discipline dans les dernières années est dû, dans une large mesure, aux représentations graphiques fournies. Ces graphiques peuvent mettre en évidence des relations difficilement saisies par l'analyse directe des données ; mais surtout, ces représentations ne sont pas liées à une opinion « a priori » sur les lois des phénomènes analysés contrairement aux méthodes de la statistique classique.

Les fondements mathématiques de l'analyse des données ont commencé à se développer au début du XX^{e} siècle, mais ce sont les ordinateurs qui ont rendu cette discipline opérationnelle, et qui en ont permis une utilisation très étendue. Mathématiques et informatique sont ici intimement liées.

1.1. Définition

Dans l'acception française, la terminologie « analyse des données » désigne un sous-ensemble de ce qui est appelé plus généralement la statistique multivariée. L'analyse des données est un ensemble de techniques descriptives, dont l'outil mathématique majeur est l'algèbre matricielle, et qui s'exprime sans supposer a priori un modèle probabiliste.

Elle comprend l'analyse en composantes principales (ACP), employée pour des données quantitatives, et ses méthodes dérivées : l'analyse factorielle des correspondances (AFC) utilisée sur des données qualitatives (tableau d'association) et l'analyse factorielle des correspondances multiples (AFCM ou

ACM) généralisant la précédente. L'analyse canonique et l'analyse canonique généralisée, qui sont plus des cadres théoriques que des méthodes aisément applicables, étendent plusieurs de ces méthodes et vont au-delà des techniques de description. L'Analyse Factorielle Multiple est adaptée aux tableaux dans lesquels les variables sont structurées en groupes et peuvent être quantitative et/ou qualitatives. La classification automatique, l'analyse factorielle discriminante (AFD) ou analyse discriminante permettent d'identifier des groupes homogènes au sein de la population du point de vue des variables étudiées.

1.2. Les principales méthodes

1.2.1. L'analyse par réduction des dimensions

L'analyse en composantes principales est utilisée pour réduire p variables corrélées en un nombre q de variables non corrélées de telles manières que les q variables soient des combinaisons linéaires des p variables initiales, que leur variance soit maximale et que les nouvelles variables soient orthogonales entre elles suivant une distance particulière. En ACP, les variables sont quantitatives.

Les composantes, les nouvelles variables, définissent un sous-espace à q dimensions sur lequel sont projetés les individus avec un minimum de pertes d'information. Dans cet espace le nuage de points est plus facilement représentable et l'analyse est plus aisée. En analyse des correspondances, la représentation des individus et des variables ne se fait pas dans le même espace.

La mesure de la qualité de représentation des données peut être effectuée à l'aide du calcul de la contribution de l'inertie de chaque composante à l'inertie totale. Dans l'exemple donné sur les deux images ci-contre, la première composante participe à hauteur de 45,89 % à l'inerte totale, la seconde à 21,2 %.

Plus les variables sont proches des composantes et plus elles sont corrélées avec elles. L'analyste se sert de cette propriété pour l'interprétation des axes. Dans l'exemple de la fig.01 les deux composantes principales représentent l'activité majeure et l'activité secondaire la plus fréquente dans lesquelles les Femmes (F) et les Hommes (H) mariés (M) ou célibataires (C) aux Usa (U) ou en Europe de l'Ouest (W) partagent leur journée. Sur la fig.02 est illustré le cercle des corrélations où les variables sont représentées en fonction de leur projection sur le plan des deux premières composantes. Plus les variables sont bien représentées et plus elles sont proches du cercle. Le cosinus de l'angle

formé par deux variables est égal au coefficient de corrélation entre ces deux variables.

De même, plus l'angle engendré par l'individu et l'axe de la composante est petit et mieux l'individu est représenté. Si deux individus, bien représentés par un axe, sont proches, ils sont proches dans leur espace. Si deux individus sont éloignés en projection, ils sont éloignés dans leur espace.

1.2.2. L’analyse factorielle des correspondances

Le but de l'AFC - définie par Jean-Paul Benzécri et ses équipes - est de trouver des liens ou correspondances entre deux variables qualitatives (nominales). Cette technique traite les tableaux de contingence de ces deux variables. En fait, une AFC est une ACP sur ces tableaux dérivés du tableau initial munis de la métrique du Le principe de l'AFC est identique à celui de l'ACP. Les axes explicatifs qui sous-tendent le tableau de fréquences de deux variables qualitatives sont recherchés et présentés dans un graphique.

Il y a au moins deux différences entre une ACP et une AFC : la première est qu'on peut représenter les individus et les variables dans un même graphique, la seconde concerne la similarité. Deux points-lignes sont proches dans la représentation graphique, si les profils-colonnes sont similaires. Par exemple sur le graphique de la fig.03, Paris et les Yvelines ont voté d'une manière similaire, ce qui n'est pas évident quand on regarde le tableau de contingence initial puisque le nombre de votants est assez différent dans les deux départements. De même, deux points-colonnes (dans l'exemple des figures 03 et 04 les points colonnes sont les candidats) sont proches graphiquement si les profils-lignes sont similaires. Dans l'exemple (fig.04), les départements ont votés pour Bayrou et Le Pen de la même manière. Les points-lignes et les points-colonnes ne peuvent pas être comparés d'une manière simple.

En ce qui concerne l'interprétation des facteurs, Jean-Paul Benzécri est très clair :

« ..Interpréter un axe, c'est trouver ce qu'il y a d'analogue d'une part entre tout ce qui est écrit à droite de l'origine, d'autre part entre tout ce qui s'écarte à gauche ; et exprimer, avec concision et exactitude, l'opposition entre les deux extrêmes.....Souvent l'interprétation d'un facteur s'affine par la considération de ceux qui viennent après lui. »

La qualité de la représentation graphique peut être évaluée globalement par la

part du expliquée par chaque axe (mesure de la qualité globale), par l'inertie d'un point projetée sur un axe divisé par l'inertie totale du point (mesure de la qualité pour chaque modalité), la contribution d'un axe à l'inertie totale ou le rapport entre l'inertie d'un nuage (profils lignes ou profils colonnes) projeté sur un axe par l'inertie totale du même nuage.

1.2.3. L'analyse des correspondances multiples

L'analyse factorielle des correspondances multiples est une extension de l'AFC.

L'ACM se propose d'analyser p ($p \geq 2$) variables qualitatives d'observations sur n individus. Comme il s'agit d'une analyse factorielle elle aboutit à la représentation des données dans un espace à dimensions réduites engendré par les facteurs. L'ACM est l'équivalent de l'ACP pour les variables qualitatives et elle se réduit à l'AFC lorsque le nombre de variables qualitatives est égal à 2.

Formellement, une ACM est une AFC appliquée sur le tableau disjonctif complet, ou bien une AFC appliquée sur le tableau de Burt, ces deux tableaux étant issus du tableau initial. Un tableau disjonctif complet est un tableau où les variables sont remplacées par leurs modalités et les éléments par 1 si la modalité est remplie 0 sinon pour chaque individu. Un tableau de Burt est le tableau de contingence des p variables prises deux à deux.

Les corrélations entre les variables à l'intérieur des deux groupes sont représentées par les corrélogrammes du haut, la corrélation entre les deux groupes est expliquée au-dessous. Si la couleur dominante était vert clair aucune corrélation n'aurait été détectée. Sur la fig.07, les deux groupes de variables sont rassemblés dans le cercle des corrélations rapportés aux deux premières variables canoniques.

Enfin l'analyse canonique généralisée au sens de Caroll (d'après J.D.Caroll) étend l'analyse canonique ordinaire à l'étude de p groupes de variables ($p > 2$) appliquées sur le même espace des individus. Elle admet comme cas particuliers l'ACP, l'AFC et l'ACM, l'analyse canonique simple, mais aussi la régression simple, et multiple, l'analyse de la variance, l'analyse de la covariance et l'analyse discriminante.

1.2.3.1. Le positionnement multidimensionnel

Pour utiliser cette technique les tableaux ne doivent pas être des variables caractéristiques d'individus mais des « distances » entre les individus. L'analyste souhaite étudier les similarités et les dissimilarités entre ces individus.

Le positionnement multidimensionnel (« multidimensional scaling » ou MDS) est donc une méthode factorielle applicable sur des matrices de distances entre individus. Cette méthode ne fait pas partie de ce qu'on nomme habituellement l'analyse des données « à la française ». Mais elle a les mêmes caractéristiques que les méthodes précédentes : elle est fondée sur le calcul matriciel et ne demande pas d'hypothèse probabiliste. Les données peuvent être des mesures de p variables quantitatives sur n individus, et dans ce cas l'analyste calcule la matrice des distances ou bien directement un tableau des distances entre individus.

Dans le cas classique dit métrique, la mesure des dissimilarités utilisée est une distance euclidienne. Elle permet d'approximer les dissimilarités entre individus dans l'espace de dimension réduite. Dans le cas non métrique les données sont ordinales, de type rang. L'analyste s'intéresse plus à l'ordre des dissimilarités plutôt qu'à leur étendue. La MDS non métrique utilise un indice de dissimilarité (équivalent à une distance mais sans l'inégalité triangulaire) et permet l'approximation de l'ordre des entrées dans la matrice des dissimilarités par l'ordre des distances dans l'espace de dimension réduite.

Comme en ACP, il faut déterminer le nombre de dimensions de l'espace cible, et la qualité de la représentation, est mesurée par le rapport de la somme de l'inertie du sous-espace de dimension réduite sur l'inertie totale. En fait, MDS métrique est équivalent à une ACP où les objets de l'analyse MDS serait les individus de l'ACP. Dans l'exemple ci-contre, les villes seraient les individus de l'ACP et le positionnement GPS remplacerait les distances inter-villes. Mais l'Analyse MDS prolonge l'ACP, puisqu'elle peut utiliser des fonctions de similarité/dissimilarité moins contraignantes que les distances.

Avec le positionnement multidimensionnel, visualiser les matrices de dissimilarités, analyser des benchmarks et effectuer visuellement des partitionnements dans des matrices de données ou de dissimilarités sont des opérations aisées à effectuer.

1.2.4. L’analyse factorielle multiple

L’analyse factorielle multiple (AFM) est dédiée aux tableaux dans lesquels un ensemble d’individus est décrit par plusieurs groupes de variables, que ces

variables soient quantitatives, qualitatives ou mixtes. Cette méthode est moins connue que les précédentes mais son très grand potentiel d'application justifie une mention particulière.

1.2.4.1. Des exemples d'application

- Dans les enquêtes d'opinion, les questionnaires sont toujours structurés en thèmes. On peut vouloir analyser plusieurs thèmes simultanément.
- Pour une catégorie de produits alimentaires, on dispose, sur différents aspects des produits, de notes données par des experts et de notes données par des consommateurs. On peut vouloir analyser simultanément les données des experts et les données des consommateurs.
- Pour un ensemble de milieux naturels, on dispose de données biologiques (abondance d'un certain nombre d'espèces) et de données environnementales (caractéristiques du sol, du relief, etc.). On peut vouloir analyser simultanément ces deux types de données.
- Pour un ensemble de magasins, on dispose du chiffre d'affaires par produit à différentes dates. Chaque date constitue un groupe de variables. On peut vouloir étudier ces dates simultanément.

1.2.4.2. L'intérêt

Dans tous ces exemples, il est utile prendre en compte, dans l'analyse elle-même et non seulement lors de l'interprétation, la structure des variables en groupes. C'est ce que fait l'AFM qui :

- pondère les variables de façon à équilibrer l'influence des différents groupes, ce qui est particulièrement précieux lorsque l'on est en présence de groupes quantitatifs et de groupes qualitatifs ;
- fournit des résultats classiques des analyses factorielle : représentation des individus, des variables quantitatives et des modalités des variables qualitatives ;
- fournit des résultats spécifiques de la structure en groupe : représentation des groupes eux-mêmes (un point = un groupe), des individus vus par chacun des groupes (un individu = autant de points que de groupes), des facteurs des analyses séparées des groupes (ACP ou ACM selon la nature des groupes).

1.3. Les autres méthodes

Ces méthodes, mises au point plus récemment, sont moins bien connues que les précédentes.

- L'Analyse Factorielle Multiple Hiérarchique (« Hierarchical Multiple Factorial Analysis ») prend en compte une hiérarchie sur les variables variables et non seulement une partition comme le fait l'AFM
- L'Analyse Procustéenne Généralisée (« Generalized Procustean Analysis ») juxtapose au mieux plusieurs représentations d'un même nuage de points.
- L'Analyse Factorielle Multiple Duale (« Dual Multiple Factor Analysis ») prend en compte une partition des individus.
- L'Analyse Factorielle de Données Mixtes (« Factor Analysis of Mixed Data »)[i] est adaptée aux tableaux dans lesquels figurent à la fois des variables quantitatives et qualitatives.

Iconographie des corrélations entre les variables des planètes. Traits pleins : corrélations positives "remarquables". Traits pointillés : corrélations négatives "remarquables".

- L'iconographie des corrélations représente les corrélations entre variables (qualitatives et quantitatives) ainsi que les individus « remarquables ». Cette méthode non supervisée se prête bien à la restitution d'une organisation, qu'elle soit arborescente ou bouclée, hiérarchique ou non. Quelle que soit la dimension des données, variables et individus remarquables sont à la surface d'une sphère ; il n'est donc pas besoin d'interpréter des axes. Plus que sur la position des points, l'interprétation repose essentiellement sur l'organisation des liens.
- L'ACI décompose une variable multivariée en composantes linéairement et statistiquement indépendantes.

1.3.1. L'analyse par classification

La classification des individus est le domaine de la classification automatique et de l'analyse discriminante. Classifier consiste à définir des classes, classer est l'opération permettant de mettre un objet dans une classe définie au préalable. La classification automatique est ce qu'on appelle en exploration de données (« data mining ») la classification non supervisée, l'analyse discriminante fait partie des techniques statistiques connues en exploration de données sous le nom de classification supervisée.

1.3.1.1. La classification automatique

Le but de la classification automatique est de découper l'ensemble des données étudiées en un ou plusieurs sous-ensembles nommés classes, chaque sous-ensemble devant être le plus homogène possible. Les membres d'une classe ressemblent plus aux autres membres de la même classe qu'aux membres d'une autre classe. Deux types de classification peuvent être relevés : d'une part la classification (partitionnement ou recouvrement) « à plat » et d'autre part le partitionnement hiérarchique. Dans les deux cas, classifier revient à choisir une mesure de la similarité/dissimilarité, un critère d'homogénéité, un algorithme, et parfois un nombre de classes composant la partition.

1.3.1.2. La classification « à plat »

La ressemblance (similarité/dissimilarité) des individus est mesurée par un indice de similarité, un indice de dissimilarité ou une distance[i]. Par exemple, pour des données binaires l'utilisation des indices de similarité tels que l'indice de Jaccard, l'indice de Dice, l'indice de concordance ou celui de Tanimoto est fréquente. Pour des données quantitatives, la distance euclidienne est la plus appropriée, mais la distance de Mahalanobis est parfois adoptée. Les données sont soit des matrices de p variables qualitatives ou quantitatives mesurées sur n individus, soit directement des données de distances ou des données de dissimilarité.

Le critère d'homogénéité des classes est en général exprimé par la diagonale d'une matrice de variances-covariances (l'inertie) inter-classes ou intra-classes. Ce critère permet de faire converger les algorithmes de ré-allocation dynamiques qui minimisent l'inertie intra-classe ou qui maximisent l'inertie inter-classes.

Les principaux algorithmes utilisent la ré-allocation dynamique en appliquant la méthode de B.W. Forgy des centres mobiles, ou une de ses variantes : la méthode des k-means, la méthode des nuées dynamiques, ou PAM (« Partitioning Around Medoids (PAM) »). Les méthodes basées sur la méthode de Condorcet, l'algorithme espérance-maximisation, les densités sont aussi utilisées pour bâtir une classification.

Il n'y a pas de classification meilleure que les autres, en particulier lorsque le nombre de classes de la partition n'est pas prédéterminé. Il faut donc mesurer la qualité de la classification et faire des compromis. La qualité de la

classification peut se mesurer à l'aide de l'indice qui est le rapport de l'inertie inter classe sur l'inertie totale, calculé pour plusieurs valeurs du nombre de classe total, le compromis étant obtenu par la méthode du coude.

L'interprétation des classes, permettant de comprendre la partition, peut s'effectuer en analysant les individus qui composent chaque classe. Le statisticien peut compter les individus dans chaque classe, calculer le diamètre des classes - ie la distance maximum entre individus de chaque classe. Il peut identifier les individus proches du centre de gravité, établir la séparation entre deux classes - opération consistant à mesurer la distance minimum entre deux membres de ces classes. Il peut analyser aussi les variables, en calculant par exemple la fréquence de certaines valeurs de variables prises par les individus de chaque classe, ou en caractérisant les classes par certaines valeurs de variables prises par les individus de chaque classe.

1.3.1.3. La classification hiérarchique

Les données en entrée d'une classification ascendante hiérarchique (CAH) sont présentées sous la forme d'un tableau de dissimilarités ou un tableau de distances entre individus.

Il a fallu au préalable choisir une distance (euclidienne, Manhattan, Tchebychev ou autre) ou un indice de similarité (Jacard, Sokal, Sorensen, coefficient de corrélation linéaire, ou autre).

La classification ascendante se propose de classer les individus à l'aide d'un algorithme itératif. À chaque étape, l'algorithme produit une partition en agrégeant deux classes de la partition obtenue à l'étape précédente.

Le critère permettant de choisir les deux classes dépend de la méthode d'agrégation. La plus utilisée est la méthode de Ward qui consiste à agréger les deux classes qui font baisser le moins l'inertie interclasse. D'autres indices d'agrégation existent comme celui du saut minimum (« single linkage ») où sont agrégées deux partitions pour lesquelles deux éléments - le premier appartenant à la première classe, le second à la seconde - sont le plus proches selon la distance prédéfinie, ou bien celui du diamètre (« complete linkage ») pour lequel les deux classes à agréger sont celles qui possèdent le couple d'éléments le plus éloigné.

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

La qualité de la classification est mesurée par le rapport inertie inter-classe sur inertie totale.

Des stratégies mixtes, alliant une classification « à plat » à une classification hiérarchique, offrent quelques avantages. Effectuer une CAH sur des classes homogènes obtenues par une classification par ré-allocation dynamique permet de traiter les gros tableaux de plusieurs milliers d'individus, ce qui n'est pas possible par une CAH seule. Effectuer une CAH après un échantillonnage et une analyse factorielle permet d'obtenir des classes homogènes par rapport à l'échantillonnage.

1.3.2. L'analyse factorielle discriminante

L'analyse factorielle discriminante (AFD), qui est la partie descriptive de l'analyse discriminante, est aussi connue sous le nom d'analyse linéaire discriminante, d'analyse discriminante de Fisher et d'analyse canonique discriminante. Cette technique projette des classes prédéfinies sur des plans factoriels discriminant le plus possible. Le tableau de données décrit n individus sur lesquels p variables quantitatives et une variable qualitative à q modalités ont été mesurées. La variable qualitative permet de définir les q classes et le regroupement des individus dans ces classes. L'AFD se propose de trouver q-1 variables, appelées variables discriminantes, dont les axes séparent le plus les projections des q classes qui découpent le nuage de points.

Comme dans toutes les analyses factorielles descriptives, aucune hypothèse statistique n'est faite au préalable ; ce n'est que dans la partie prédictive de l'analyse discriminante que des hypothèses a priori sont émises.

La mesure de la qualité de la discrimination est effectuée à l'aide du de Wilks qui est égal au rapport du déterminant de la matrice de variances-covariances intra-classe sur le déterminant de la matrice de variances-

covariances totale. Un de Wilks faible indique une discrimination forte par les plans factoriels. Par exemple sur les données Iris, il est de 0.0234 sur les deux premiers facteurs. En outre si la première valeur propre est proche de 1, l'AFD est de qualité.

La corrélation entre les variables et les facteurs permet d'interpréter ceux-ci.

Une AFD est une ACP effectuée sur les barycentres des classes d'individus constituées à l'aide des modalités de la variable qualitative. C'est aussi une

analyse canonique entre le groupe des variables quantitatives et celui constitué du tableau disjonctif de la variable qualitative.

2. L’analyse des données et les régressions

En s'inspirant de ce qu'écrivent Henry Rouanet et ses coauteurs, l'analyse des données descriptive et l'analyse prédictive peuvent être complémentaires, et parfois produire des résultats similaires.

2.1. L’approche PLS

L'approche PLS est plus prédictive que descriptive, mais les liens avec certaines analyses que l'on vient de voir ont été clairement établis.

L'algorithme d'Herman Wold, nommé tout d'abord NILES (« Nonlinear Estimation by Iterative Least SquareS »), puis NIPALS (« Nonlinear Estimation by Iterative Partial Least SquareS ») a été conçu en premier lieu pour l'analyse en composantes principales

En outre, PLS permet de retrouver l'analyse canonique à deux blocs de variables, l'analyse inter batteries de Tucker, l'analyse des redondances et l'analyse canonique généralisée au sens de Carroll. La pratique montre que l'algorithme PLS converge vers les premières valeurs propres dans le cas de l'analyse inter batteries de Tucker, l'analyse canonique à deux blocs de variables et l'analyse des redondances.

2.2. Les régressions

La régression sur composantes principales (PCR) utilise l'ACP pour réduire le nombre de variables explicatives en les remplaçant par les composantes principales qui ont l'avantage de ne pas être corrélées. PLS et PCR sont souvent comparées l'une à l'autre dans la littérature.

Déjà mentionné plus haut dans cet article, l'analyse canonique est équivalente à la régression linéaire lorsqu'un des deux groupes se réduit à une seule variable.

3. Les logiciels

L'analyse des données moderne ne peut être dissociée de l'utilisation des ordinateurs ; de nombreux logiciels permettant d'utiliser les méthodes d'analyse des données vues dans cet article peuvent être cités. SPSS, Statistica,

HyperCube et SAS fournissent des modules complets d'analyse des données ; le logiciel R aussi avec des bibliothèques comme FactoMineR, Ade4 ou MASS ; Braincube, solution d'analyse de données massives pour l'industrie.

II. QUELQUES NOTIONS DE STATISTIQUES DESCRIPTIVES

1. Les tableaux statistiques

1.1. La population statistique

Une population statistique est l'ensemble sur lequel on effectue des observations.

Exemples :
- ensemble de personnes interrogées pour une enquête
- ensemble de parcelles cultivées sur lesquelles on mesure un rendement
- ensemble de pays pour lesquels on dispose de données géographiques ou économiques, ...

1.2. Les paramètres statistiques

Ce sont quelques nombres permettant de résumer numériquement les traits principaux d'une distribution statistique.

Par exemple : la moyenne, l'écart-type, l'étendue sont des paramètres statistiques.

1.3. Le caractère statistique (ou variables statistiques)

C'est ce qui est observé ou mesuré sur les individus d'une population statistique.
Il peut s'agir d'une variable qualitative ou quantitative.

Exemples :
- Taille, poids, salaire, sexe, profession d'un groupe donné d'individus
- Rendement d'un ensemble de parcelles cultivées
- Température maximale et minimale, pluviométrie, ensoleillement, mesurés à un endroit donné tous les jours

1.4. L’individu (ou unités statistiques)

Les individus sont les éléments de la population statistique étudiée. Pour chaque individu, on dispose d'une ou plusieurs observations.

Exemples :

- chacune des personnes interrogées pour une enquête
- chaque parcelle cultivée en vue d'étudier le rendement
- chaque pays pour lequel on étudie des données socio-économiques, ..
- chaque jour de l'année pour lequel on dispose de données météorologiques, ...

1.5. La modalité

Les modalités d'une variable qualitative sont les différentes valeurs que peut prendre celle-ci.

Par exemple les modalités de la variable "situation familiale" sont : célibataire, marié, veuf, divorcé.

Les modalités de la variable "sexe" sont : féminin, masculin (pouvant être codées par exemple 0 et 1).

1.6. L’unité statistique (ou individu)

Les individus sont les éléments de la population statistique étudiée.

Pour chaque individu, on dispose d'une ou plusieurs observations.

Exemples :

- chacune des personnes interrogées pour une enquête
- chaque parcelle cultivée en vue d'étudier le rendement
- chaque pays pour lequel on étudie des données socio-économiques, ...
- chaque jour de l'année pour lequel on dispose de données météorologiques, ...

1.7. La variable continue

C'est une variable quantitative pouvant prendre par nature une infinité de valeurs, généralement tout un intervalle réel.

Exemples :
- tailles, poids, salaires, surfaces cultivées, températures, ...

1.8. La variable discrète

C'est une variable quantitative pouvant prendre par nature un nombre fini (ou dénombrable) de valeurs.

Exemples :
- nombre d'enfants par famille
- nombre de pièces d'un appartement
- nombre de pièces défectueuses dans un lot de pièces mécaniques ...

1.9. La variable qualitative (ou caractère qualitatif)

Une variable statistique est qualitative si ses valeurs, ou modalités, s'expriment de façon littérale ou par un codage sur lequel les opérations arithmétiques telles que moyenne, somme, ... , n'ont pas de sens.

Exemples :
- sexe de la personne interrogée, situation familiale, numéro de son département de naissance, ...
- état du temps constaté à un endroit donné chaque jour (pluvieux, neigeux, beau, venteux, ...)

1.10. La variable qualitative nominale

C'est une variable qualitative dont les modalités ne sont pas ordonnées.

Exemples
- la variable sexe peut être notée M F , 0 1 ,

: ou 1 0

- la variable CSP : on ne peut pas classer les catégories socio- professionnelles selon un ordre préétabli.

1.11. La variable qualitative ordinale

C'est une variable qualitative dont les modalités sont naturellement ordonnées selon un ordre total : on peut dire que selon un certain sens la modalité A est moins forte que la B, qui est moins forte que la C, etc...

Exemples :

- tailles de vêtement 0 1 2 3 ... mais la taille 2 ne signifie pas que le vêtement est 2 fois plus grand que celui de la taille 1 ! Il ne s'agit pas d'une variable quantitative discrète.

1.12. La variable quantitative (ou caractère quantitatif)

Une variable statistique est quantitative si ses valeurs sont des nombres sur lesquels des opérations arithmétiques telles que somme, moyenne, ... ont un sens.

Exemples :

- taille, poids, salaire
- rendement
- note à un examen
- PNB / habitant, espérance de vie, nombre d'habitants d'un ensemble de pays

1.13. La variable statistique (ou caractère statistique)

C'est ce qui est observé ou mesuré sur les individus d'une population statistique. Il peut s'agir d'une variable qualitative ou quantitative.

Exemples :

- taille, poids, salaire, sexe, profession d'un groupe donné d'individus
- rendement d'un ensemble de parcelles cultivées

- température maximale et minimale, pluviométrie, ensoleillement, mesurés à un endroit donné tous les jours.

1.14. Les classes

Intervalles de valeurs d'une variable continue, l'ensemble des classes formant une partition de l'ensemble des valeurs possibles de la variable.

Par exemple, si tous les salaires des employés d'une entreprise se situent entre 750 et moins de 3 000 dollars, on peut construire (par exemple) les classes :

[750 - 900 [, [900 - 1 500 [, [1 500 - 2 250 [, [2 250 - 3 000 [

Chaque valeur observée de la variable doit appartenir à une classe et une seule.

1.15. Les effectifs

Nombre d'individus pour lesquels une variable statistique a pris une valeur donnée. Si, sur 150 familles, 50 ont 2 enfants, on dira que l'effectif n_i correspondant à la valeur $x_i = 2$ de la variable "nombre d'enfants", est 50.

1.16. Les effectifs cumulés

Résultat de l'addition, de proche en proche, des effectifs d'une distribution observée, soit en commençant par le 1er :

$N_1 = n_1$, $N_2 = n_1 + n_2$, ... , $N_i = n_1 + n_2 + ... + n_i$

(effectifs cumulés croissants),

soit en commençant par le dernier :

$N'_K = n_K$, $N'_{K-1} = n_K + n_{K-1}$, ... , $N_i' = n_K + n_{K-1} + ... + n_i$

(effectifs cumulés décroissants).

Exemples :

Nombre d'appels	Nombre de jours	Effectifs cumulés croissants	Effectifs cumulés décroissants
0	2	2	96
1	14	16	94
2	23	39	80
3	24	63	57
4	18	81	33
5	9	90	15
6	6	96	6
Total :	96		

1.17. Les effectifs totaux

C'est le nombre d'observations, d'une série statistique brute, nombre d'individus de la population étudiée.

Il est égal à la somme des effectifs associés aux différentes modalités, valeurs ou classes :

$$n = \sum_{i=1}^{K} n_i$$

1.18. Les fréquences (ou fréquences relatives)

C'est la proportion (ou le pourcentage) d'individus pour lesquels une variable statistique a pris une valeur donnée. Si, sur 150 familles, 50 ont 2 enfants, on dira que la fréquence f_i correspondant à la valeur $x_i = 2$ de la variable "nombre d'enfants", est :

$f_i = \frac{n_i}{n} = \frac{50}{150} = 0.33$ soit 1/3 ou 33.33 %

1.19. Les fréquences cumulées

Résultat de l'addition, de proche en proche, des fréquences d'une distribution observée, soit en commençant par le 1er :

$F_1 = f_1$, $F_2 = f_1 + f_2$, ... , $F_i = f_1 + f_2 + ... + f_i$

(fréquences cumulées croissantes),

Soit en commençant par le dernier :

$F'_K = f_K$, $F'_{K-1} = f_K + f_{K-1}$, ... , $F'_i = f_K + f_{K-1} + ... + f_i$

(Fréquences cumulées décroissantes).

Exemples :

Nombre d'appels	Fréquences en %	Fréquences cumulées croissantes	Fréquences cumulées décroissantes
0	2.08	2.08	100
1	14.58	16.66	97.92
2	23.96	40.62	83.34
3	25.00	65.62	59.38
4	18.75	84.37	34.38
5	9.38	93.75	15.63
6	6.25	100	6.25

2. Les caracteristiques de tendances centrales

2.1. Le mode

C'est la valeur observée d'effectif maximum.

2.1.1. La variable discrète

Classer les données par ordre croissant. Celle d'effectif maximum donne le mode.
Il est fortement conseillé d'utiliser le diagramme en bâtons pour déterminer le mode. En effet, deux valeurs consécutives xi , xi+1 peuvent avoir le même effectif maximum; on parlera d'intervalle modal [xi , xi+1]. Il peut aussi y avoir un mélange de deux populations qui conduit à un diagramme en bâtons où apparaissent deux bosses; on considérera deux modes. Il est déconseillé, sauf raison explicite, d'envisager plus de deux modes.

2.1.2. La variable classée

La classe modale correspond à la classe ayant l'effectif maximum. Il est fortement conseillé d'utiliser l'histogramme pour déterminer le mode. Comme pour le cas discret, on peut avoir deux classes modales. Toutes les valeurs de la

classe pouvant à priori se réaliser, on ne se contentera pas de déterminer la classe modale. Une des valeurs de cette classe sera le mode. Certains auteurs préconisent par simplicité de prendre le centre de la classe modale. Il est préférable cependant de tenir compte des classes adjacentes de la manière suivante:

2.2. La médiane

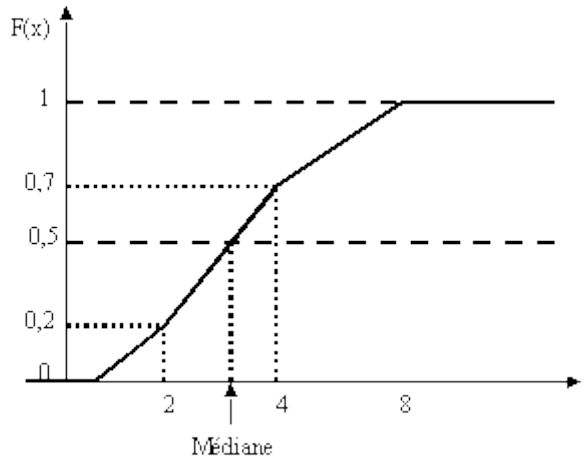
Les valeurs étant rangées par ordre croissant, c'est la valeur de la variable qui sépare les observations en deux groupes d'effectifs égaux.

2.2.1. La variable discrète

La détermination peut s'obtenir à partir du tableau statistique en recherchant la valeur de la variable correspondant à une fonction cumulée égale à n/2 (effectif cumulé) ou ½ (fréquence cumulée). Il est encore plus facile de lire sur les graphiques cumulatifs les abscisses des points d'ordonnée n/2 (effectif cumulé) ou ½ (fréquence cumulée). Si tout un intervalle a pour image n/2 (½ pour la fréquence), on parlera d'intervalle médian (on peut prendre le milieu de l'intervalle comme médiane)

2.2.2. La variable classée

L'abscisse du point d'ordonnée n/2 (½ pour la fréquence)se situe en général à l'intérieur d'une classe. Pour obtenir une valeur plus précise de la médiane, on procède à une interpolation linéaire. La valeur de la médiane peut être lue sur le graphique ou calculée analytiquement.



$$\frac{Mé - 2}{4 - 2} = \frac{0,5 - 0,2}{0,7 - 0,2}$$

D'où la valeur de la médiane.

De manière générale, si a et b sont les bornes de la classe contenant la médiane, F(a) et F(b) les valeurs de la fréquence cumulée croissante en a et b, alors

$$Mé = a + (b - a) \times \frac{0,5 - F(a)}{F(b) - F(a)}$$

2.3. La moyenne arithmétique

Si xi sont les observations d'une variable discrète ou les centres de classe d'une variable classée, la moyenne arithmétiq ue $\bar{\bar{x}}$ est égale à $\sum_{i=1}^{k} \frac{n_i x_i}{n} = \sum_{i=1}^{k} f_i x_i$

La moyenne arithmétique est un paramètre de tendance centrale plus utilisé que les autres de par ses propriétés algébriques:

2.3.1. Lorsque il y a plusieurs populations

Pour les d'effectifs n1, n2, ..., nk, de moyennes respectives $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k$

La moyenne globale = La moyenne des moyennes

$$\bar{\bar{x}} = \frac{\sum_{i=1}^{k} n_i \bar{x}_i}{n}$$

2.3.2. Pour la moyenne arithmétique

Elle conserve les changements d'échelle et d'origine

$$x : (x_i, n_i) \rightarrow y : (y_i = ax_i + b, n_i)$$
$$\bar{x} \rightarrow \bar{y} = a\bar{x} + b$$

2.4. La moyenne géométrique

Si xi sont les observations d'une variable quantitative, la moyenne géométrique est égale à

$$G = \sqrt[n]{x_1^{n_1} \times \ldots\ldots \times x_k^{n_k}}$$

Ce type de moyenne est surtout utilisé pour calculer des pourcentages moyens. r étant un taux d'accroissement, 1+r est appelé coefficient multiplicateur; et le coefficient multiplicateur moyen est alors égal à la moyenne géométrique des coefficients multiplicateurs.

2.5. La moyenne harmonique

Si xi sont les observations d'une variable quantitative, la moyenne harmonique est égale à

$$H = \frac{n}{\sum_{i=1}^{k} \frac{n_i}{x_i}}$$

Il n'est pas évident d'utiliser ce type de moyenne.
Elle intervient lorsqu'on demande une moyenne de valeurs se présentant sous forme de quotient de deux variables x/y (km/h, km/litre,...). Attention, il faut cependant bien décortiquer le problème car il peut aussi s'agir d'une moyenne arithmétique.

2.6. La moyenne quadratique

Si xi sont les observations d'une variable quantitative, la moyenne harmonique est égale à

$$Q = \sqrt{\frac{n_1 x_1^2 + \ldots\ldots + n_k x_k^2}{n}}$$

3. Les caractéristiques de dispersion

3.1. L'étendue

C'est la différence entre la plus grande et la plus petite observation

3.2. La variance et l'écart-type

Si xi sont les observations d'une variable discrète ou les centres de classe d'une variable classée, la variance

V est égale à $\sum_{i=1}^{k} \frac{n_i (x_i - \bar{x})^2}{n} = \sum_{i=1}^{k} f_i (x_i - \bar{x})^2$

On a aussi $V = \sum_{i=1}^{k} \frac{n_i x_i^2}{n} - \bar{x}^2$

c.à.d. moyenne des carrés - carré de la moyenne

On utilise plus couramment l'écart-type qui est la racine carrée de la variance et qui a l'avantage d'être un nombre de même dimension que les données (contrairement à la variance qui en est le carré)

La variance est un paramètre de dispersion plus utilisé que les autres de par ses propriétés algébriques:

a) Pour plusieurs populations d'effectifs $n_1, n_2, \ldots, n_k$, de moyennes respectives $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k$, de variances $V_1, V_2, \ldots, V_k$

Variance globale = variance des moyennes + moyenne des variances

$$V = \sum_{i=1}^{k} \frac{n_i}{n} (\bar{x}_i - \bar{\bar{x}})^2 + \sum_{i=1}^{k} \frac{n_i}{n} V_i$$

où $\bar{\bar{x}}$ représente la moyenne des moyennes

b) changement d'échelle et d'origine

$$x : (x_i , n_i) \rightarrow y : (y_i = a x_i + b , n_i)$$

$$V_x \rightarrow V_y = a^2 V_x$$

3.3. La variance expliquée

C'est la variance des moyennes des distributions conditionnelles : si Y est quantitative, et si X subdivise l'ensemble des individus en K classes d'effectifs n_1 , n_2 , ... , n_K telles que la moyenne de Y sur chaque classe est : $\bar{y}_1$, ... , $\bar{y}_K$,

la variance de Y expliquée par X est :

$\frac{1}{n} \sum_{i=1}^{K} (n_i \bar{y}^2_i) - \bar{y}^2$

3.4. La variance résiduelle

C'est la moyenne des variances des distributions conditionnelles, pondérées par les effectifs. SiY est quantitative, et si X subdivise l'ensemble des individus en K classes d'effectifs n_1 , n_2 , ... , n_K telles que la moyenne de Y sur chaque classe est :

$\overline{y}_1$, ... , $\overline{y}_K$,

avec les variances s^2_1 , s^2_2 , ..., s^2_K , la variance de Y se décompose en :

$$s^2_Y = \left[\frac{1}{n}\sum_{i=1}^{n}(n_i \overline{y}_i^2) - \overline{y}^2\right] + \left[\frac{1}{n}\sum_{i=1}^{n} n_i s_i^2\right]$$

Le premier terme de la somme est la variance de Y expliquée par X, le second la variance résiduelle.

3.5. Le coefficient de variation

$$CV = \frac{\sigma}{\overline{x}}$$

C'est un coefficient qui permet de relativiser l'écart-type en fonction de la taille des valeurs. Il permet ainsi de comparer la dispersion de séries de mesures exprimées dans des unités différentes

4. La régression

4.1. La covariance

On appelle covariance de deux variables statistiques X et Y sur les mêmes n individus le nombre :

$$cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \left(\frac{1}{n}\sum_{i=1}^{n} x_i y_i\right) - \overline{x}\,\overline{y}$$

"Moyenne des produits moins le produit des moyennes"

Ce nombre est positif si X et Y ont tendance à varier dans le même sens, et négatif si elles ont tendance à varier en sens contraire.

Si les données sont groupées en (x_i , y_i) d'effectifs n_i ,

$$\mathrm{cov}(X,Y) = \frac{1}{n}\sum_{i=1}^{K} n_i (x_i - \bar{x})(y_i - \bar{y}) = \left(\frac{1}{n}\sum_{i=1}^{K} n_i x_i y_i\right) - \bar{x}\,\bar{y}$$

4.2. Le rapport de corrélation

C'est

$$e^2_{Y/X} = \frac{\text{variance expliquée}}{\text{variance de Y}}$$

coefficient compris entre 0 et 1 mesurant la part plus ou moins grande de la variabilité d'une variable Y qui peut être expliquée par les variations d'une autre variable X, qualitative, discrète, ou continue découpée en classes.

4.3. La courbe de régression

Si X et Y sont 2 variables quantitatives, la courbe de régression de Y en X est la courbe représentant les moyennes conditionnelles de Y, à X fixé.

La courbe de régression de X en Y représente les moyennes conditionnelles de X, à Y fixé.

5. Exercices corrigés

5.1. Exercice 1

A. Un industriel a commandé à un sous-traitant un lot de 40 pièces dont le diamètre doit mesurer 80 mm et il est convenu que le lot ne sera accepté que si les deux conditions suivantes sont simultanément réalisées :

Première condition **:** l'écart entre 80 mm et la moyenne $\bar{x}$ du lot est inférieur à 0,05 mm.

Deuxième condition : Au moins 60 % des pièces du lot ont un diamètre d tel que $80-0,05 \leq d \leq 80+0,05$ (1)
Les mesures faites sur le lot sont les suivantes :

Mesure de d à 0,05 mm près	79,75	79,80	79,85	79,90	79,95	80	80,05	80,10	80,15	80,20
Effectif	1	2	3	5	6	14	5	2	1	1

1) Calculer la moyenne x des mesures faites
2) Quel est le pourcentage de pièces dont le diamètre d vérifie la double inégalité (1) ?
3) Le lot est-il accepté ou refusé par l'industriel ? Justifier la réponse

B. Un élève a obtenu les notes suivantes : 4;6;3;9;10;8;12;10;19;12;20;12;18 . Calculer sa moyenne

C.

1) Calculer $\sum_{i=0}^{6}(2i+1)$
2) Écrire en utilisant la notation Σ: 3+5+7+9+…15+17

Solution de l'exercice 5.1

A.

1) La moyenne $\bar{x}$ des mesures faites vaut :

$$\bar{x} = \frac{1\times 79,75 + 2\times 79,80 + \ldots\ldots + 1\times 80,15 + 1\times 80,2}{40} = \frac{3198,9}{40} = 79,9725$$

2) Le nombre de pièces dont le diamètre d vérifie la double inégalité (1) est égal à 6+14+5=25, soit un pourcentage égal à 25/40 *100 = 62,5%

3) L'écart entre la moyenne x et 80 mm étant égal à 80−79,9725=0,0275<0,05, et plus de 60 % des pièces ayant un diamètre d vérifiant la double inégalité (1), le lot sera accepté.

B.

La moyenne de l'élève est égale à :

$$\bar{x} = \frac{4+6+3+9+10+8+12+10+19+12+20+12+18}{13} = 11$$

C.

1) $\sum_{i=0}^{6}(2i+1)$ = 2×0+1+2×1+1+2×2+1+2×3+1+2×4+1+2×5+1+2×6+1=49

2) 3+5+7+…+17= $\sum_{i=1}^{8}(2i+1)$

5.2. Exercice 2

Un relevé des durées de communications téléphoniques effectuées dans un central téléphonique a fourni les informations consignées dans le tableau suivant (l'unité de durée est la minute).

Intervalle de durée	[0;2[	[2;4[	[4;6[	[6;8[	[8;10[	[10;12[
Effectif	14	16	25	15	17	13

Travail à faire:

1) Calculer la durée moyenne d'un appel
2) On regroupe les classes par deux, ce qui revient à considérer les classes [0;4[, [4,8[et [8;12[. Calculer la durée moyenne d'un appel pour cette nouvelle série
3) Quelle conclusion pouvez-vous formuler ?

Solution de l'exercice 5.2

1) Pour calculer la moyenne de cette série statistique, on prend en compte le milieu des classes, à savoir :

Intervalle de durée	[0;2[	[2;4[	[4;6[	[6;8[	[8;10[	[10;12[
Milieu des classes	**1**	**3**	**5**	**7**	**9**	**11**
Effectif	14	16	25	15	17	13

La durée moyenne d'un appel vaut donc :

$$\bar{x} = \frac{1\times 14 + 3\times 16 + \ldots\ldots + 9\times 17 + 11\times 13}{100} = \frac{588}{100} = 5{,}88 \text{ minutes}$$

Soit 5 minutes et 0,88×60=52,8 secondes.

La durée moyenne d'un appel vaut donc 5 minutes, 52 secondes et 8 dixièmes.

2) La nouvelle série statistique est donc**:**

Intervalle de durée	[0;4[	[4;8[	[8;12[
Effectif	14+16=30	25+15=40	17+13=30

Pour calculer la moyenne de cette série statistique, on prend en compte le milieu des classes, à savoir :

Intervalle de durée	[0;4[	[4;8[	[8;12[
Milieu des classes	**2**	**6**	**10**
Effectif	30	40	30

La durée moyenne d'un appel calculée à partir de cette série vaut donc :

$$\bar{x} = \frac{2\times 30 + 6\times 40 + 10\times 30}{100} = \frac{600}{100} = 6 \text{ minutes}$$

3) Selon la manière de regrouper les communications téléphoniques (donc seulement la présentation de la série statistique !), les résultats peuvent être différents.

III. QUELQUES NOTIONS DE CALCUL MATRICIEL

1. Représentation d'une matrice

1.1. Définition

On appelle matrice A, la matrice d'ordre (n, p) à éléments dans K. On note a_{ij} l'élément K indexé par (i, j). La matrice A= (a_{ij}) avec $1 \leq i \leq n$ et $1 \leq j \leq p$ est représentée par un tableau rectangulaire

$$A = \begin{pmatrix} a_{11} & \dots & a_{1j} & \dots & a_{1p} \\ \dots & \dots & \dots & \dots & \dots \\ a_{i1} & \dots & a_{ij} & \dots & a_{ip} \\ \dots & \dots & \dots & \dots & \dots \\ a_{n1} & \dots & a_{nj} & \dots & a_{np} \end{pmatrix}$$

L'élément $a_{i\,j}$ est situé sur la ième ligne et la jième colonne par convention.

Les matrices à n lignes et p colonnes à coefficients dans K sont notées $M_{n,\,p}(K)$.

Exemple 1.1

$$A = \begin{bmatrix} 2 & -3 & 5 \\ 4 & 5 & 0 \\ 1 & 7 & 3 \end{bmatrix}$$

2. Matrices carrées

2.1. Définition

On appelle matrice carrée, toute matrice de type (n, n).

Les matrices carrées de dimension n dans K (K=R ou C) sont notées $M_{n,\,n}(K)$ ou $M_n(K)$.

Dans la matrice dénommée A, les termes de la forme a_{ii} constituent la diagonale principale.

Exemples 1.2

Pour $A \in M_2$ des réels, on prend l'exemple

$$A = \begin{bmatrix} 6 & 2 \\ 1 & -3 \end{bmatrix}$$

Pour A ϵ M_4 des réels, on prend l'exemple

$$A = \begin{bmatrix} 2 & -3 & 5 & 6 \\ 6 & 5 & 0 & 7 \\ 1 & 0 & 3 & 2 \\ 4 & 5 & 6 & 1 \end{bmatrix}$$

Pour A ϵ M_2 des complexes, on prend l'exemple

$$A = \begin{bmatrix} 6 & -i \\ 1 & -2+i \end{bmatrix}$$

2.3. Propriété

Pour l'addition et la multiplication, les matrices carrées $M_{n,n}(K)$ possèdent une structure d'anneau.

3. Matrices carrées remarquables

3.1. Matrice diagonale

3.1.1. Définition

On appelle matrice diagonale, la matrice carrée dont tous les éléments sont nuls exceptés ceux de la diagonale principale.

$A = (a_{ij})$ avec $a_{ij} = 0$ pour $i \neq j$

Exemple 1.3

$$A = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 3 \end{bmatrix}$$

3.1.2. Propriété

Si on élève une matrice diagonale à une puissance quelconque, cela induit l'élévation de chaque élément de sa diagonale à la même puissance.

Exemple 1.4

$$A = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 3 \end{bmatrix} \Rightarrow A^3 = \begin{bmatrix} 8 & 0 & 0 \\ 0 & 64 & 0 \\ 0 & 0 & 27 \end{bmatrix}$$

3.2. Matrice unité ou identité

3.2.1. Définition

On appelle matrice unité I, la matrice diagonale ou chaque terme de la diagonale principale est égal à 1

$$A = (a_{ij}) \text{ avec } a_{ij} = 1 \text{ pour } i = j$$

Exemples 1.5

$$I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$I = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

3.2.3. Propriété

La matrice unité I est l'élément neutre du produit matriciel.

$$I\,A = A\,I = A$$

Exemples 1.6

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 7 & 4 \\ 5 & 2 \end{bmatrix} = \begin{bmatrix} 7 & 4 \\ 5 & 2 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 7 & 4 \\ 5 & 2 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 2 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 3 \end{bmatrix} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 3 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 3 \end{bmatrix}$$

3.3. Matrice scalaire

3.3.1. Définition

On appelle matrice scalaire β, la matrice diagonale ou chaque terme de la diagonale principale est égal à β

$$A = (a_{ij}) \text{ avec } a_{ij} = \beta \text{ pour } i = j \;, \quad \beta \epsilon K$$

Exemple 1.7

$$\beta = \begin{bmatrix} \beta & 0 & 0 \\ 0 & \beta & 0 \\ 0 & 0 & \beta \end{bmatrix}$$

3.3.2. Propriété

La matrice scalaire β est un nombre égal à β.

Exemple 1.8

$$A = \begin{bmatrix} 4 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 4 \end{bmatrix} \Rightarrow A = 4 \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = 4$$

3.4. Transposée d'une matrice

3.4.1. Définition

On appelle transposée de la matrice A, la matrice carrée notée A^t ou les lignes de A se substituent à ses colonnes et ses colonnes se substituent à ses lignes.

$$A = (a_{ij}) \quad \text{et} \quad A^t = (a_{ji})$$

Exemple 1.11

$$A = \begin{bmatrix} 2 & -3 & 5 \\ 4 & 5 & 0 \\ 1 & 7 & 3 \end{bmatrix} \quad \text{et} \quad A^t = \begin{bmatrix} 2 & 4 & 1 \\ -3 & 5 & 7 \\ 5 & 0 & 3 \end{bmatrix}$$

3.4.2. Propriété

La transposée de la transposée d'une matrice est égale à la matrice initiale

Exemple 1.12

$$A = \begin{bmatrix} 2 & -3 & 5 \\ 4 & 5 & 0 \\ 1 & 7 & 3 \end{bmatrix}, \quad A^t = \begin{bmatrix} 2 & 4 & 1 \\ -3 & 5 & 7 \\ 5 & 0 & 3 \end{bmatrix}$$

$$\text{et } (A^t)^t = \begin{bmatrix} 2 & -3 & 5 \\ 4 & 5 & 0 \\ 1 & 7 & 3 \end{bmatrix} \Rightarrow A = (A^t)^t$$

La transposée du produit d'un scalaire par une matrice est égale au produit du scalaire par la transposée de la matrice

Exemple 1.13

$$A = \begin{bmatrix} 2 & -3 & 5 \\ 4 & 5 & 0 \\ 1 & 7 & 3 \end{bmatrix} \text{ et } (2A)^t = \begin{bmatrix} 4 & 8 & 2 \\ -6 & 10 & 14 \\ 10 & 0 & 6 \end{bmatrix}$$

$$\text{et } 2(A)^t = 2\begin{bmatrix} 2 & 4 & 1 \\ -3 & 5 & 7 \\ 5 & 0 & 3 \end{bmatrix}$$

3.5. Matrice symétrique

3.5.1. Définition

On appelle matrice symétrique A, la matrice carrée dont les éléments symétrique par rapport à la diagonale sont égaux.

$$A = (a_{ij}) = (a_{ji})$$

Exemple 1.9

$$A = \begin{bmatrix} 2 & 3 & 9 \\ 3 & 4 & 1 \\ 9 & 1 & 3 \end{bmatrix}$$

3.5.2. Propriété

La matrice symétrique A est égale à sa transposée A^t.

Exemple 1.10

$$A = \begin{bmatrix} 2 & 3 & 9 \\ 3 & 4 & 1 \\ 9 & 1 & 3 \end{bmatrix} \text{ et } A^t = \begin{bmatrix} 2 & 3 & 9 \\ 3 & 4 & 1 \\ 9 & 1 & 3 \end{bmatrix} \Rightarrow A = A^t$$

3.6. Matrice antisymétrique

3.6.1. Définition

On appelle matrice antisymétrique A, la matrice carrée dont les éléments symétriques par rapport à la diagonale sont opposés et ceux de la diagonale principale sont nuls.

$$a_{ij} = -a_{ji} \quad \text{et} \quad a_{ii} = 0$$

Exemple 1.11

$$A = \begin{bmatrix} 0 & -3 & 9 \\ 3 & 0 & -1 \\ -9 & 1 & 0 \end{bmatrix}$$

3.6.2. Propriété

La matrice antisymétrique A est égale à $-A^{t.}$

Exemple 1.12

$$A = \begin{bmatrix} 0 & -3 & 9 \\ 3 & 0 & -1 \\ -9 & 1 & 0 \end{bmatrix},\ A^t = \begin{bmatrix} 0 & 3 & -9 \\ -3 & 0 & 1 \\ 9 & -1 & 0 \end{bmatrix} \text{ et } -A^t = \begin{bmatrix} 0 & -3 & 9 \\ 3 & 0 & -1 \\ -9 & 1 & 0 \end{bmatrix} \Rightarrow A = -A^t$$

3.7. Matrice triangulaire

3.7.1. Définition

On appelle matrice triangulaire supérieure, la matrice carrée dont les éléments sont nuls au-dessous de la diagonale principale

$$a_{ii} = 0 \quad \text{pour } i > j$$

Exemple 1.13

$$A = \begin{bmatrix} 2 & 3 & 9 \\ 0 & 7 & 1 \\ 0 & 0 & 5 \end{bmatrix}$$

3.7.2. Définition

On appelle matrice triangulaire inférieure, la matrice carrée dont les éléments sont nuls au-dessus de la diagonale principale

$$a_{ii} = 0 \quad \text{pour } i < j$$

Exemple 1.14

$$A = \begin{bmatrix} 2 & 0 & 0 \\ 1 & 7 & 0 \\ 4 & 3 & 5 \end{bmatrix}$$

4. Matrices non carrées

4.1. Définition

On appelle matrice non carrée, la matrice dont le nombre de ligne est différent de celui des colonnes. Les matrices non carrées sont notées $M_{n,p}(K)$ avec $n \neq p$.

Si n=1, c'est une matrice ligne à p colonnes.

Si p=1, c'est une matrice colonne à n lignes.

Exemples 1.15

La matrice rectangulaire horizontale $A = \begin{bmatrix} 1 & 2 & 6 & 5 \\ 3 & 5 & 9 & 7 \end{bmatrix}$

La matrice rectangulaire verticale $A = \begin{bmatrix} 4 & 8 \\ 9 & 5 \\ 2 & 6 \\ 1 & 5 \end{bmatrix}$

La matrice ligne $A = [2 \;\; 8 \;\; 7 \;\; 5 \;\; 1 \;\; 5]$

La matrice colonne $A = \begin{bmatrix} 2 \\ 8 \\ 6 \\ 0 \\ 9 \end{bmatrix}$

5. Déterminant d'une matrice

5.1. Définition

Soit la matrice carrée $A = (a_{ij}) \quad 1 \leq i, j \leq n \quad \in M_n(\mathbb{K})$.

$$A = \begin{pmatrix} a_{11} & \dots & a_{1j} & \dots & a_{1n} \\ \dots & \dots & \dots & \dots & \dots \\ a_{i1} & \dots & a_{ij} & \dots & a_{in} \\ \dots & \dots & \dots & \dots & \dots \\ a_{n1} & \dots & a_{nj} & \dots & a_{nn} \end{pmatrix}$$

On appelle déterminant de la matrice A, d'ordre n, le tableau carré contenant les éléments de la matrice limité par deux traits verticaux.

$$\det A = \begin{vmatrix} a_{11} & \dots & a_{1j} & \dots & a_{1n} \\ \dots & \dots & \dots & \dots & \dots \\ a_{i1} & \dots & a_{ij} & \dots & a_{in} \\ \dots & \dots & \dots & \dots & \dots \\ a_{n1} & \dots & a_{nj} & \dots & a_{nn} \end{vmatrix}$$

Notation : $|A|$ ou

5.2. Exemple

$$n = 2 \quad |A_2| = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}$$

$$n = 3 \quad |A_3| = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix}$$

On appelle mineur $|M_{ij}|$ de l'élément a_{ij} du déterminant d'ordre n, le déterminant d'ordre $(n-1)$ obtenu en supprimant la ième ligne et la jème colonne de $|A|$.

5.3. Exemple

$$n = 2 \quad |A_2| = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}$$

$|M_{11}| = a_{22}, |M_{12}| = a_{21}, \dots$

$$n = 3 \quad |A_3| = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix}$$

$$|M_{12}| = \begin{vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{vmatrix} = a_{21}a_{33} - a_{23}a_{31}$$

$$|M_{21}| = \begin{vmatrix} a_{12} & a_{13} \\ a_{32} & a_{33} \end{vmatrix} = a_{12}a_{33} - a_{13}a_{32}$$

On appelle cofacteur Δ_{ij} de l'élément a_{ij}, le mineur $|M_{ij}|$ affecté du signe + ou _ suivant la relation : $\Delta_{ij} = (-1)^{i+j}|M_{ij}|$

5.4. Exemple

$$n = 2 \quad |A_2| = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}$$

$\Delta_{11} = (-1)^{1+1} |M_{11}| = +a_{22}$

$$n = 3 \quad |A_3| = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix}$$

$$\Delta_{12} = (-1)^{1+2} |M_{12}| = -(a_{21}a_{33} - a_{23}a_{31})$$

6. Valeurs propres - Vecteurs propres

Soit $\mathcal{E}$ un espace vectoriel sur $\mathbb{K}$ et f un endomorphisme de $\mathcal{E}$.

6.1. Définition

On appelle vecteur propre de f tout vecteur x, non nul de $\mathcal{E}$, vérifiant : $f(x) = \lambda x$.

(Les vecteurs propres sont donc les vecteurs dont la direction est inchangée par l'application f).

Le scalaire $l \in \mathbb{K}$ est appelé valeur propre associée au vecteur x.

6.2. Calcul de valeurs propres et vectrices propres

$$X = \begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{pmatrix}$$

- Si $A = (a_{ij})$ est la matrice de l'application f dans une base B de $\mathcal{E}$ et la matrice unicolonne du vecteur propre x dans B, alors :

  $f(x) = \lambda x \Rightarrow AX = \lambda X \Leftrightarrow (A - \lambda I)X = 0$ (I : Matrice unité d'ordre n)

- Le système homogène ainsi obtenu :

$$\begin{cases} (a_{11} - \lambda)x_1 + a_{12}x_2 + \dots + a_{1n}x_n = 0 \\ a_{21}x_1 + (a_{22} - \lambda)x_2 + \dots + a_{2n}x_n = 0 \\ \dots \\ a_{n1}x_1 + a_{n2}x_2 + \dots + (a_{nn} - \lambda)x_n = 0 \end{cases}$$

  À l'exclusion de la solution triviale $X = 0$, admettra des solutions si le déterminant de $(A - \lambda I) = 0$.

- Les valeurs propres de f (ou de A) sont les scalaires λ tels que :

$$\det(A-\lambda I)=\begin{vmatrix} a_{11}-\lambda & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22}-\lambda & \dots & \dots \\ \dots & \dots & \dots & \dots \\ a_{n1} & \dots & \dots & a_{nn}-\lambda \end{vmatrix}=0$$

L'équation de degré n en λ ainsi obtenu est dite Equation caractéristique.

(Voir exemple "Calcul de valeurs propres" ci-dessous)

- Un vecteur propre x de composantes $(x', x'', \dots)$ associé à la valeur propre λ doit vérifier la relation :

$$AX=\lambda X \Leftrightarrow (A-\lambda I_n)\begin{pmatrix} x' \\ x'' \\ \dots \end{pmatrix}=\begin{pmatrix} 0 \\ 0 \\ \dots \end{pmatrix}$$

(In : Matrice identité à l'ordre n)

Propriété **:** Si la matrice A admet p valeurs propres, distinctes deux à deux, les p vecteurs propres associés sont linéairement indépendants et forment une base de l'espace vectoriel $\mathcal{E}$.

6.2.1. Exemple : Calcul de valeurs propres

Déterminer les valeurs propres de la matrice $A=\begin{pmatrix} 5 & -3 \\ 6 & -4 \end{pmatrix}$

Les valeurs propres de A sont les scalaires l vérifiant :

$$\begin{aligned} \det(A-\lambda I_2)=0 \quad &\Leftrightarrow \begin{vmatrix} 5-\lambda & -3 \\ 6 & -4-\lambda \end{vmatrix} \\ &= -(5-\lambda)(4+\lambda)+18 \\ &= \lambda^2-\lambda-2 \\ &= (\lambda+1)(\lambda-2)=0 \end{aligned}$$

D'où les valeurs propres : $l_1=-1$ et $l_2=+2$

6.2.2. Exemple : Calcul de vecteurs propres

Déterminer les vecteurs propres associés aux valeurs propres de la matrice $A=\begin{pmatrix} 5 & -3 \\ 6 & -4 \end{pmatrix}$. Les vecteurs propres obtenus forment-ils une base de $\mathbb{R}^2$?

En posant x_1 et x_2 les vecteurs propres associés respectivement à l_1 et l_2, nous avons :

- Pour $l_1 = -1$

$(A + I_2)X_1 = 0 \Leftrightarrow \begin{pmatrix} 6 & -3 \\ 6 & -3 \end{pmatrix} \begin{pmatrix} x'_1 \\ x''_1 \end{pmatrix} = 0$

Le système est équivalent à : $6x'_1 - 3x''_1 = 0$

Choix d'un vecteur propre : $x_1 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$

- Pour $l_2 = +2$

$(A - 2I_2)X_2 = 0 \Leftrightarrow \begin{pmatrix} 3 & -3 \\ 6 & -6 \end{pmatrix} \begin{pmatrix} x'_2 \\ x''_2 \end{pmatrix} = 0$

Le système est équivalent à : $3x'_2 - 3x''_2 = 0$

Choix d'un vecteur propre : $x_2 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$

Comme le déterminant $\begin{vmatrix} 1 & 1 \\ 2 & 1 \end{vmatrix} \neq 0$, la famille (x_1, x_2) est une base de R^2.

VI. L'ANALYSE DES DONNEES PROPREMENT DITE

Dans cette partie, on va se limiter à présenter deux méthodes d'analyse de données : L'analyse en composantes principales (A.C.P) et l'analyse factorielle des correspondances (A.F.C). (Ces 2 méthodes sont puisées de *cours gratuits d'informatique*)

1. Les méthodes

1.1. L'analyse en composantes principales (A.C.P)

L'analyse en composantes principales (A.C.P.) est une méthode mathématique d'analyse graphique de données qui consiste à rechercher les directions de l'espace qui représentent le mieux les corrélations entre *n* variables aléatoires (relation linéaire entre elles).

Simplement dit, une A.C.P. permet de trouver des similitudes de comportement entre les classes des données observées.

Même si l'A.C.P. est majoritairement utilisée pour visualiser des données, il ne faut pas oublier que c'est aussi un moyen :

- De décorréler ces données. Dans la nouvelle base, constituée des nouveaux axes, les points ont une corrélation nulle (nous le démontrerons).

- De classifier ces données en amas (clusters) corrélés (dans l'industrie c'est surtout cette possibilité qui est intéressante!).

Remarque: L'A.C.P. est aussi connue sous le nom de "transformée de Karhunen-Loève" ou de "transformée de Hotelling" et peut aussi bien être appliquée sans programmation V.B.A. dans MS Excel que dans des logiciels spécialisés (ou le temps de calcul sera par contre plus bref... et plus précis aussi...).

Lorsque nous ne considérons que deux effets, il est usuel de caractériser leurs effets conjoints via le coefficient de corrélation. Lorsque l'on se place en dimension deux, les points disponibles (l'échantillon de points tirés suivant la loi conjointe) peuvent être représentés sur un plan. Le résultat d'une A.C.P. sur ce plan est de déterminer les deux axes qui expliquent le mieux la dispersion des points disponibles.

Lorsqu'il y a plus de deux effets, par exemple trois effets, il y a trois coefficients de corrélations à prendre en compte. La question qui a donné naissance à l'A.C.P. est : comment avoir une intuition rapide des effets conjoints?

En dimension plus grande que deux, une A.C.P. va toujours déterminer les axes qui expliquent le mieux la dispersion du nuage des points disponibles..

L'objectif de l'A.C.P. est de décrire graphiquement un tableau de données d'individus avec leurs variables quantitatives de grande taille :

individus/variables	$\mathrm{var}_1 \cdots \mathrm{var}_j \cdots \mathrm{var}_p$
ind_1 $\vdots$ ind_i $\vdots$ ind_n	x_{ij}

Tableau: 57.1 - Représentation type d'un tableau A.C.P.

Afin de ne pas alourdir l'exposé de cette méthode et de permettre au lecteur de refaire complètement les calculs, nous travaillerons sur un exemple.

Considérons pour l'exemple une étude d'un botaniste qui a mesuré les dimensions de 15 fleurs d'iris. Les trois variables $(p = 3)$ mesurées sont :

- x_1 : longueur du sépale

- x_2 : largeur du sépale

- x_3 : longueur du pétale

Les données sont les suivantes :

Fleur n°	x_1	x_2	x_3

1	5.1	3.5	1.4
2	4.9	3.0	1.4
3	4.7	3.2	1.3
4	4.6	3.1	1.5
5	5.0	3.6	1.4
6	7.0	3.2	4.7
7	6.4	3.2	4.5
8	6.9	3.1	4.9
9	5.5	2.3	4.0
10	6.5	2.8	4.6
11	6.3	3.3	6.0
12	5.8	2.7	5.1
13	7.1	3.0	5.9
14	6.3	2.9	5.6
15	6.5	3.0	5.8

Tableau: 57.2 - Exemple pratique de données tabulaires A.C.P.

Pour nous un tel tableau de données sera tout simplement une matricée réelle à *n* lignes (les individus) et à *p* colonnes (les variables) :

$$X = \left(x_{ij}\right)_{\substack{i=1\dots n \\ j=1\dots p}} \quad (57.1)$$

Par suite l'indice *i* correspondra à l'indice ligne et donc aux individus. Nous identifierons donc l'individu *i* avec le point ligne $x_{i.} = \left(x_{i1}, \dots, x_{ip}\right)$ qui sera considéré comme un point dans un espace affine (cf. chapitre de Calcul Vectoriel) de dimension *p*. L'indice *j* correspondra à l'indice colonne donc aux variables. Nous identifierons la variable *j* avec le vecteur colonne :

$$\vec{x}_{.j} = \begin{pmatrix} x_{1j} \\ \vdots \\ x_{nj} \end{pmatrix} \quad (57.2)$$

c'est donc un vecteur dans l'espace vectoriel de dimension *n* dans $\mathbb{R}^n$.

Nous nous placerons dans la suite suivant deux points de vue : Soit nous prendrons le tableau de données comme *n* points dans un espace affine de dimension *p*, soit nous prendrons ce tableau comme *p* points d'un espace vectoriel de dimension *n*. Nous verrons qu'il y a des dualités entre ces deux points de vue.

L'outil mathématique que nous allons utiliser ici est l'algèbre linéaire (cf. chapitre d'Algèbre Linéaire), avec les notions de produit scalaire, de norme euclidienne et de distance euclidienne.

Afin de simplifier la présentation, nous allons dans un premier temps considérer que chaque individu, comme chaque variable, a la même importance, le même poids. Nous ne considérerons aussi, que le cas de la distance euclidienne.

Nous allons commencer en centrant les données, c'est-à-dire mettre l'origine du système d'axes au centre de gravité du nuage de points. Ceci ne modifie pas l'aspect du nuage, mais permet d'avoir les coordonnées du point *M* égales aux coordonnées du vecteur $\overrightarrow{GM}$ et donc de ce placer dans l'espace vectoriel pour pouvoir y faire les calculs! Comme nous supposons dans toute la suite que le poids des individus sont identiques, nous prendrons donc $m_i = 1/n$ avec $i = 1...n$.

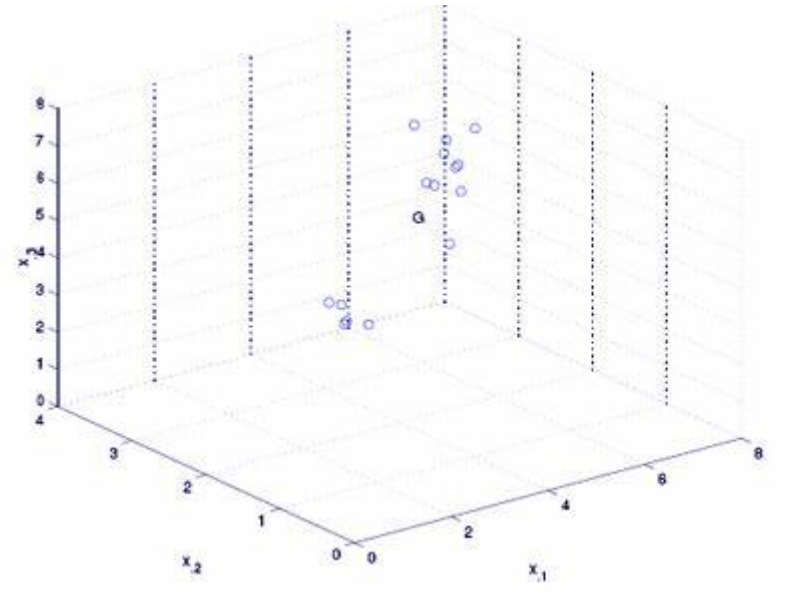
Nous considérons le repère orthonormé $\left(O,\vec{e}_1,\vec{e}_2,...,\vec{e}_p\right)$ dans la bas canonique $\left(\vec{e}_1,\vec{e}_2,...,\vec{e}_p\right)$ de $\mathbb{R}^p$. Soit donc *G* le centre de gravité du nuage de point, Comme nous considérons ici chaque variable, comme chaque individu, ayant le même poids, *G* a alors pour coordonnées dans le repère $\left(O,\vec{e}_1,\vec{e}_2,...,\vec{e}_p\right)$:

$$\overrightarrow{OG} = \begin{pmatrix} \overline{x}_1 \\ \vdots \\ \overline{x}_j \\ \vdots \\ \overline{x}_p \end{pmatrix} \quad (57.3)$$

avec :

$$\overline{x}_j = \frac{1}{n}\sum_{i=1}^{j} x_{ij} \quad (57.4)$$

Nous avons alors pour l'instant sous forme graphique :



(57.5)

Nous appelons "matrice centrée" la matrice :

$$X_c = \begin{bmatrix} x_{11} - \overline{x}_{.1} & \cdots & x_{1j} - \overline{x}_{.j} & \cdots & x_{1p} - \overline{x}_{.p} \\ \vdots & & \vdots & & \vdots \\ x_{i1} - \overline{x}_{.1} & \cdots & x_{ij} - \overline{x}_{.j} & \cdots & x_{1ip} - \overline{x}_{.p} \\ \vdots & & \vdots & & \vdots \\ x_{n1} - \overline{x}_{.1} & \cdots & x_{nj} - \overline{x}_{.j} & \cdots & x_{np} - \overline{x}_{.p} \end{bmatrix} \quad (57.6)$$

Remarque: La matrice des données centrées contient les coordonnées centrées (que nous noterons $^{xc_{ij}}$) des individus dans le repère $\left(G, \vec{e}_1, \vec{e}_2, \ldots, \vec{e}_p\right)$. Nous nous placerons dans la suite toujours dans ce repère pour le nuage de points des individus et nous prendrons $O = G$.

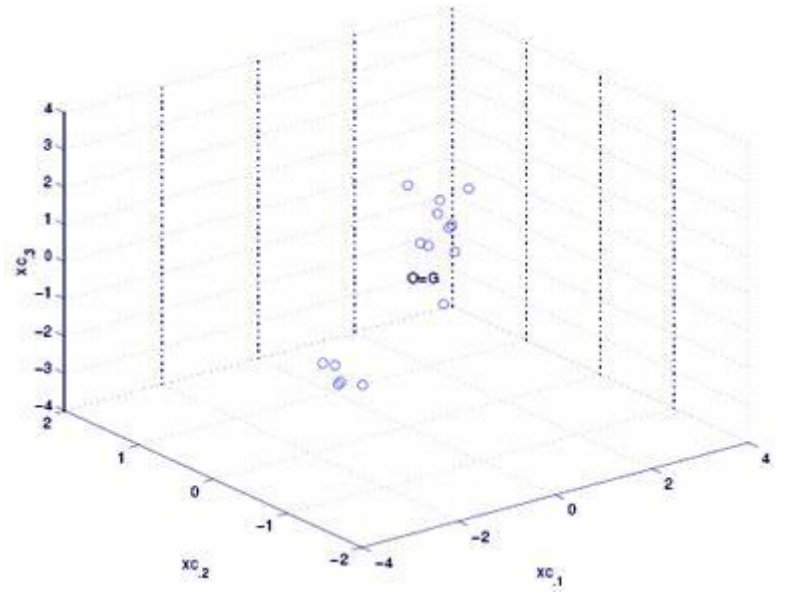
Pour notre exemple, nous avons :

$$G = \overline{x} = (5.91, 3.06, 3.87) \quad (57.7)$$

et pour la matrice centrée :

$$X_c = \begin{bmatrix} -0.8067 & 0.4400 & -2.4733 \\ -1.0067 & -0.0600 & -2.4733 \\ -1.2067 & 0.1400 & -2.5733 \\ -1.3067 & 0.0400 & -2.3733 \\ -0.9067 & 0.1400 & -2.4733 \\ 1.0933 & 0.5400 & 0.8267 \\ 0.4933 & 0.1400 & 0.6267 \\ 0.9933 & 0.0400 & 1.0267 \\ -0.4067 & -0.7600 & 0.1267 \\ 0.5933 & -0.2600 & 0.7267 \\ 0.3933 & 0.2400 & 2.1267 \\ -0.1067 & -0.3600 & 1.2267 \\ 1.1933 & -0.0600 & 2.0267 \\ 0.3933 & -0.1600 & 1.7267 \\ 0.5933 & -0.0600 & 1.9267 \end{bmatrix} \quad (57.8)$$

et sous forme graphique :



(57.9)

Pour donner une importance identique à chaque variable afin que le type d'unités des mesures n'influence pas l'analyse, nous travaillerons avec les données centrées réduites (cf. chapitre de Statistiques). Pour cela, nous noterons d'abord:

$$\sigma_j^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_{ij} - \overline{x}_j\right)^2 = \frac{1}{n}\sum_{i=1}^{n} c_{ij}^2 = \frac{1}{n}\left\| c_j \right\|^2 \quad (57.10)$$

la variance d'échantillon de la variable $x_{.j}$ est donc égale à un facteur 1/*n* près à la norme de cette même variable mais centrée. La matrice des données centrées réduites (sans dimensions) est alors :

$$Y = \begin{bmatrix} (x_{11} - \overline{x}_1)/\sigma_1 & \cdots & (x_{1j} - \overline{x}_j)/\sigma_j & \cdots & (x_{1p} - \overline{x}_p)/\sigma_p \\ \vdots & & \vdots & & \vdots \\ (x_{i1} - \overline{x}_1)/\sigma_1 & \cdots & (x_{ij} - \overline{x}_j)/\sigma_j & \cdots & (x_{ip} - \overline{x}_p)/\sigma_p \\ \vdots & & \vdots & & \vdots \\ (x_{n1} - \overline{x}_1)/\sigma_1 & \cdots & (x_{nj} - \overline{x}_j)/\sigma_j & \cdots & (x_{np} - \overline{x}_p)/\sigma_p \end{bmatrix} \quad (57.11)$$

Si nous notons $D_{1/\sigma}$ la matrice diagonale suivante :

$$D_{1/\sigma} = \begin{bmatrix} 1/\sigma_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1/\sigma_p \end{bmatrix} \quad (57.12)$$

Nous avons alors :

$$Y = X_c D_{1/\sigma} \quad (57.13)$$

Remarque: La moyenne de la variable $y_{.j}$ est nulle et donc sa variance est alors 1 (ce qui revient à dire que la norme de la variable centrée réduite est de norme unitaire comme nous allons de suite le démontrer).

Nous définissons la "matrice des données centrées normées" par :

$$Z = \frac{1}{\sqrt{n}} Y \quad (57.14)$$

Soit encore (il s'agit simplement de l'erreur quadratique moyenne que nous avions introduit dans le chapitre de Statistiques) :

$$z_{ij} = \frac{x_{ij} - \overline{x}_j}{\sqrt{n}\sigma_j} \quad (57.15)$$

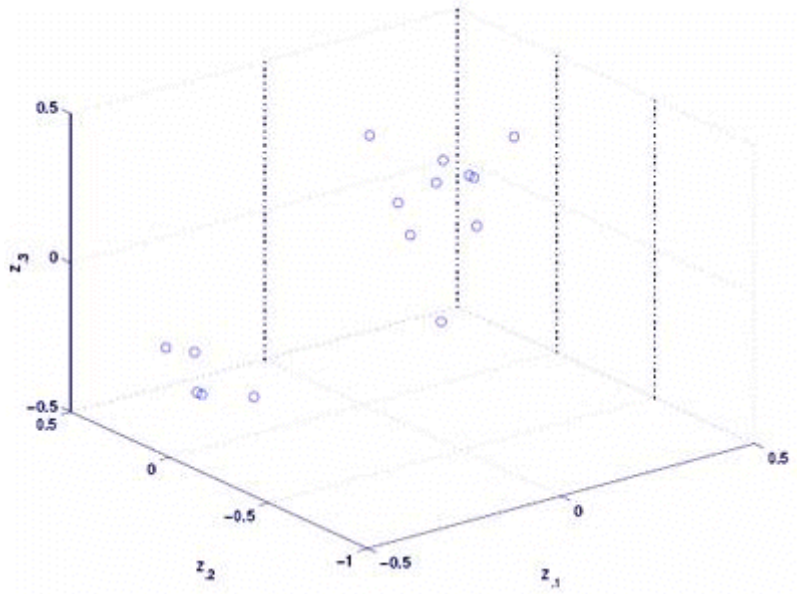
La terminologie vient bien évidemment du fait que la variable (vecteur) $z_{.j}$ est de norme unitaire. En effet :

$$\left\| z_{.j} \right\|^2 = \sum_{i=1}^{n} z_{ij}^2 = \sum_{i=1}^{n} \left(\frac{x_{ij} - \overline{x}_j}{\sqrt{n}\sigma_j} \right)^2 = \frac{1}{\sigma_j^2} \frac{1}{n} \sum_{i=1}^{n} \left(x_{ij} - \overline{x}_j \right)^2 = 1 \quad (57.16)$$

Ce qui donne:

$$Z = \begin{bmatrix} -0.2383 & 0.3572 & -0.3375 \\ -0.2974 & -0.0487 & -0.3375 \\ -0.3565 & 0.1136 & -0.3512 \\ -0.3861 & 0.0324 & -0.3239 \\ -0.2679 & 0.4384 & -0.3375 \\ 0.3230 & 0.1136 & 0.1128 \\ 0.1457 & 0.1136 & 0.0855 \\ 0.2935 & 0.0324 & 0.1401 \\ -0.1201 & -0.6170 & 0.0172 \\ 0.1753 & -0.2110 & 0.0991 \\ 0.1162 & 0.1948 & 0.2902 \\ -0.0315 & -0.2922 & 0.1674 \\ 0.3526 & -0.0487 & 0.2765 \\ 0.1162 & -0.1298 & 0.2356 \\ 0.1753 & -0.0487 & 0.2629 \end{bmatrix} \quad (57.17)$$

Nous avons graphiquement :



(57.18)

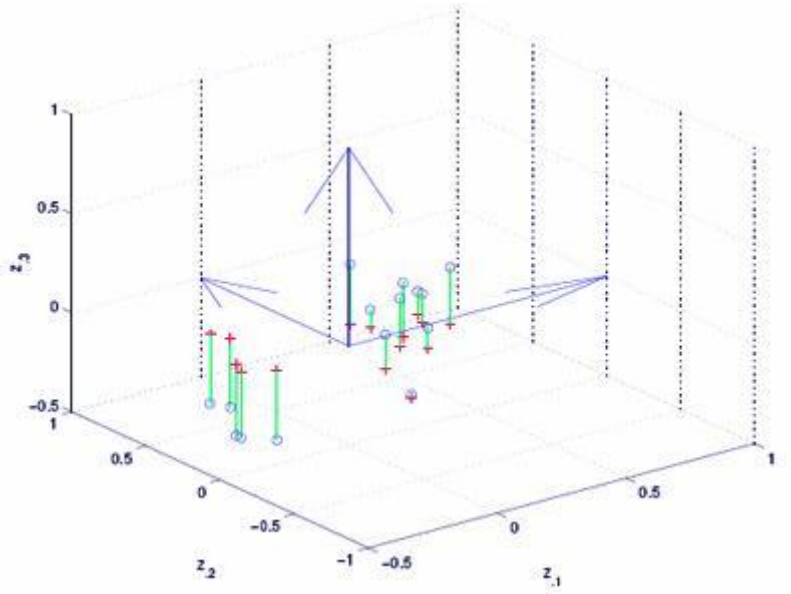
Représenter le nuage de points des données centrées réduites ou centrées normées ne modifie rien à la forme de celui-ci. En effet, la différence entre les deux n'est qu'un changement d'échelle.

L'information intéressante pour les individus est la distance entre les points! En effet plus cette distance sera grande entre deux individus $z_{i\cdot}$ et $z_{i'\cdot}$ plus les deux individus seront différents et mieux on pourra les caractériser. Mais il faut

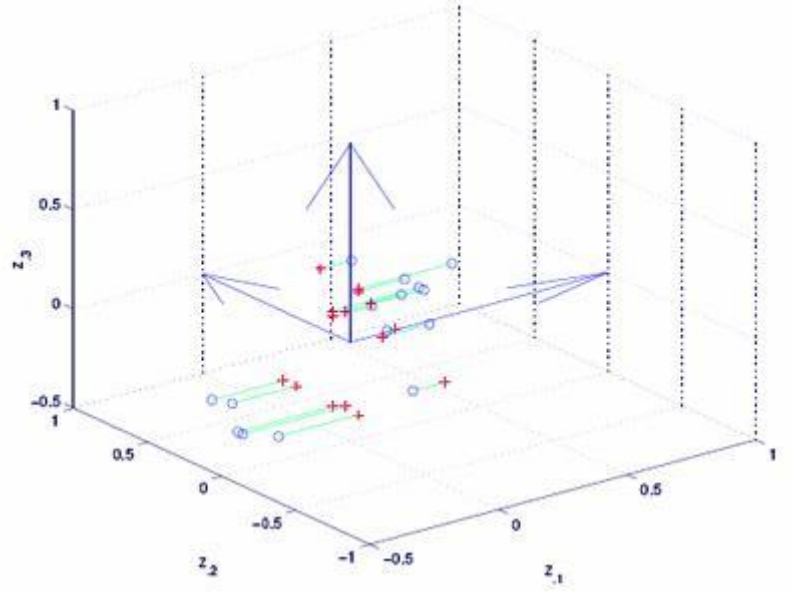
d'abord choisir une distance. Nous prendrons la distance euclidienne (cf. chapitre de Topologie) :

$$d^2(z_{i.}, z_{i'.}) = \left\| \overrightarrow{z_{i.} z_{i'.}} \right\|^2 = \sum_{j=1}^{p} \left(z_{ij} - z_{i'j} \right)^2 \quad (57.19)$$

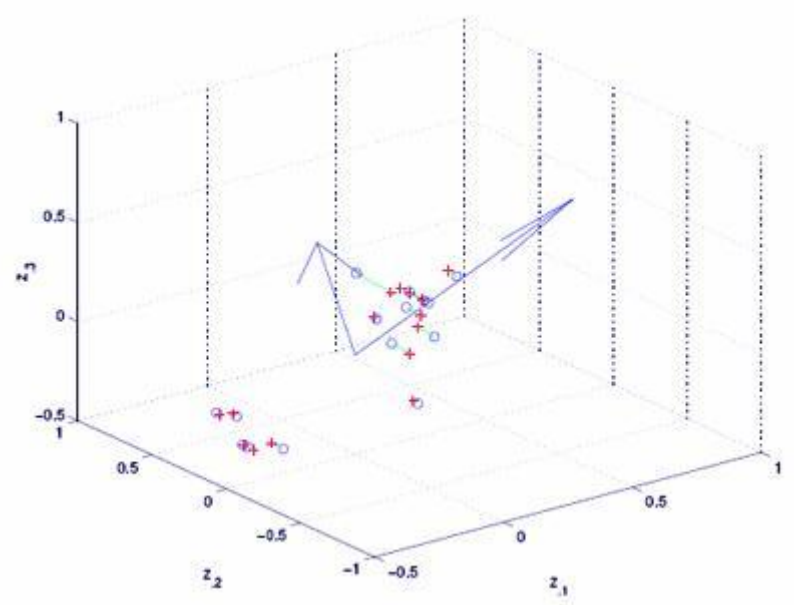
Les figures suivantes montrent les projections orthogonales dans l'espace de ce nuage de points respectivement dans les plans $(O, \vec{e}_1, \vec{e}_2), (O, \vec{e}_2, \vec{e}_3)$ et enfin dans $(O, \vec{u}_1, \vec{u}_3)$ qui est la meilleure projection, appelé "plan factoriel" (ou parfois "diagramme des scores"), dans le sens où elle respecte le mieux les distances entre les individus (in extenso, elle déforme moins le nuage de points dans l'espace). L'objectif de l'A.C.P. est de déterminer ce meilleur plan et nous démontrerons comment.



(57.20)

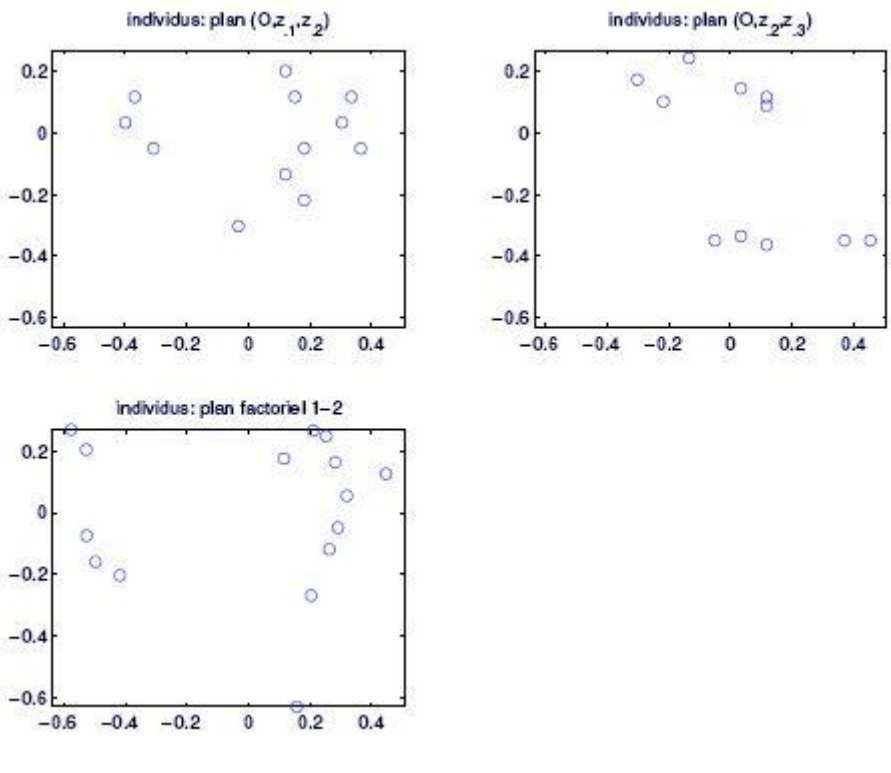


(57.21)



(57.22)

Et la vue plane de chacune des projections :



(57.23)

Avant de déterminer le plan factoriel, nous allons maintenant chercher à détecter les liens possibles entre les variables.

Nous rappelons (cf. chapitre de Statistiques) que la covariance entre deux variables x_j et $x_{j'}$ est donnée par :

$$\mathrm{cov}\left(x_j, x_{j'}\right) = E\left[\left(x_j - \overline{x}_j\right)\left(x_{j'} - \overline{x}_{j'}\right)\right] = \frac{1}{n}\sum_{i=1}^{n}\left(x_{ij} - \overline{x}_j\right)\left(x_{ij'} - \overline{x}_{j'}\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n} c_{ij} c_{ij'} \qquad (57.24)$$

et que le coefficient de corrélation linéaire (cf. chapitre de Statistiques) est :

$$R\left(x_j, x_{j'}\right) = \frac{\mathrm{cov}\left(x_j, x_{j'}\right)}{\sigma_j \sigma_{j'}} \qquad (57.25)$$

Nous noterons par la suite:

$$c_{j,j'} = \left(\mathrm{cov}\left(x_j, x_{j'}\right)\right) \text{ et } \quad r_{j,j'} = \left(R\left(x_j, x_{j'}\right)\right) \qquad (57.26)$$

les matrices des covariances et de corrélations carrées (toutes deux étant pour rappel des matrices carrées et symétriques) avec $j = 1...p, j' = 1...p$.

Nous voyons facilement que la matrices des covariances et au coefficient 1/*n* près, la matrice des produit scalaires canoniques des vecteurs de la matrice des données centrées X_c (en d'autres termes, chaque composante de la matrice des covariances est égale au produit scalaire des variables centrées). Nous en déduisons la relation suivante :

$$c_{j,j'} = \frac{1}{n} X_c^T X_c \qquad (57.27)$$

La matrice des covariances-variances (puisque comme nous l'avons vu dans le chapitre de Statistiques, la diagonale contient les variances) est un outil connu d'interprétation sur ce site. Par contre ce qui est nouveau et va nous être très utile pour déterminer le plan factoriel est la matrice de corrélation linéaire qui peut aussi être écrite sous la forme suivante :

$$R = r_{j,j'} = \frac{1}{n} Y^T Y = Z^T Z \qquad (57.28)$$

Ce qui donne pour notre exemple où nous avons trois variables, la matrice carrée suivante (que les données soient centrées ou non les composantes de la matrice sont identiques):

$$R = \begin{bmatrix} 1 & -0.1609 & 0.8854 \\ -0.1609 & 1 & -.3818 \\ 0.8854 & -.3818 & 1 \end{bmatrix} \qquad (57.29)$$

Pour continuer, toujours dans le but de déterminer le plan factoriel, définissons le concept d'inertie de nuage de point.

Définition: Nous appelons "inertie d'un nuage de points" la quantité :

$$I = \frac{1}{n}\sum_{i=1}^{n} d^2(G, M_i) \qquad (57.30)$$

où *G* est le centre de gravité du nuage de point et M_i le point de $\mathbb{R}^p$ de coordonnées x_i^t.
Remarque: Le carré de la distance est pris par anticipation des développements qui vont suivre.

Ensuite, démontrons que nous avons la relation suivante :

$$\sum_{i=1}^{n}\sum_{i'=1}^{n} d^2\left(M_i, M_i^{'}\right) = 2nI \qquad (57.31)$$

Démonstration:

$$\begin{aligned}
&\sum_{i=1}^{n}\sum_{i'=1}^{n} d^2\left(M_i, M_i^{'}\right) = \sum_{i,i'}\left\|\overrightarrow{M_iM_{i'}}\right\|^2 = \sum_{i,i'}\left\|\overrightarrow{M_iG} + \overrightarrow{GM_{i'}}\right\|^2 \\
&= \sum_{i,i'}\left(\left\|\overrightarrow{M_iG}\right\|^2 + \left\|\overrightarrow{GM_{i'}}\right\|^2 + 2\overrightarrow{M_iG} \circ \overrightarrow{GM_{i'}}\right) \\
&= \sum_{i'}\left(\sum_{i}\left\|\overrightarrow{M_iG}\right\|^2\right) + \sum_{i}\left(\sum_{i'}\left\|\overrightarrow{GM_{i'}}\right\|^2\right) + 2\sum_{i}\left(\overrightarrow{M_iG} \circ \sum_{i'}\overrightarrow{GM_{i'}}\right) \\
&= \sum_{i'}\left(\sum_{i} d^2\left(M_i, G\right)\right) + \sum_{i}\left(\sum_{i'} d^2\left(G, M_{i'}\right)\right) + 2\sum_{i}\left(\overrightarrow{M_iG} \circ \vec{0}\right) \\
&= \sum_{i'} nI + \sum_{i} nI + 0 \\
&= 2nI
\end{aligned} \qquad (57.32)$$

□C.Q.F.D

Nous allons dans toute la suite travailler avec les données centrées normées, in extenso avec la matrice *Z*. Les points M_i auront donc ici comme coordonnées z_i^t.
Le problème est maintenant de trouver le meilleur espace affine de dimension *p* dans le sens où il respecte au mieux les distances entre les points. Pour cela, nous allons rechercher la meilleure droite vectorielle $\Delta_{\vec{u}}$ qui est parfaitement déterminée par le vecteur $\vec{u}$. Appelons H_i la projection orthogonale de M_i sur la droite $\Delta_{\vec{u}}$. Alors notre problème est de trouver la

droite (in extenso le vecteur *u*) qui fasse que la somme des carrés des distances entres les points H_i soit maximale. Nous écrirons le problème sous la forme d'un problème de programmation quadratique :

$$\begin{aligned} &\max \sum_{i,i'} d^2\left(H_i, H_i^{'}\right) \\ &u \in \mathbb{R}^p \\ &\|\vec{u}\| = 1 \end{aligned} \qquad (57.33)$$

Or ici, nous avons :

$$\sum_{i,i'} d^2\left(H_i, H_i^{'}\right) = 2nI \qquad (57.34)$$

En effet, le centre de gravité du nuage de point projeté est aussi l'origine. Par suite, notre problème peut s'écrire :

$$\begin{aligned} &\max I \\ &\vec{u} \in \mathbb{R}^p \\ &\|\vec{u}\| = 1 \end{aligned} \qquad (57.35)$$

Lui même équivalent donc à :

$$\begin{aligned} &\max \sum_{i} d^2\left(O, H_i\right) \\ &\vec{u} \in \mathbb{R}^p \\ &\|\vec{u}\| = 1 \end{aligned} \qquad (57.36)$$

Résolvons donc ce problème :

Tout d'abord, puisque H_i est la projection orthogonale du point M_i sur $\Delta_{\vec{u}}$ nous avons $\overrightarrow{OH_i} = \alpha_i \vec{u}$ pour tout *i* avec $\alpha_i = \overrightarrow{OM_i} \circ \vec{u}$. Par suite les coordonnées des points H_i sur la droite $\Delta_{\vec{u}}$ sont :

$$\begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{pmatrix} = Z \cdot \vec{u} \qquad (57.37)$$

Par suite, nous avons :

$$\sum_{i} d^2\left(O, H_i\right) = \sum_{i} \alpha_i^2 = \left\|Z \cdot \vec{u}\right\|^2 = Z \cdot \vec{u} \circ Z \cdot \vec{u} \qquad (57.38)$$

Ici nous cherchons le vecteur unitaire $\vec{u}$. La matrice *Z* nous est parfaitement connue. Or, nous avons :

$$(Z \cdot \vec{u}) \circ (Z \cdot \vec{u}) = \left(Z^T Z \vec{u}\right) \circ \vec{u} = (R \cdot \vec{u}) \circ \vec{u} \quad (57.39)$$

La matrice de corrélation *R* est symétrique donc, selon le théorème spectral vu dans le chapitre d'Algèbre Linéaire, elle est diagonalisable dans une base orthonormée de vecteurs propres. Ainsi, nous avions démontré dans le théorème spectral que :

$$\Lambda = S^{-1} R S \quad (57.40)$$

est diagonale si *R* est symétrique et *S* orthogonale (qui donc une matrice carrée 3×3 dans notre exemple!). Donc :

$$R = S \Lambda S^{-1} \quad (57.41)$$

et comme *S* avait été démontrée comme orthogonale, nous avons (cf. chapitre d'Algèbre Linéaire) :

$$S^{-1} = S^T \quad (57.42)$$

Donc :

$$R = S \Lambda S^T \quad (57.43)$$

où nous choisissons pour Λ la matrice diagonale des valeurs propres mises en ordre décroissant : $\lambda_1 \geq \lambda_2 \ldots \geq \lambda_p$.

Nous avons donc :

$$(R \cdot u) \circ u = \left(S \Lambda S^T \vec{u}\right) \circ \vec{u} = \left(\Lambda S^T \vec{u}\right) \circ S^T \vec{u} = (\Lambda \vec{w}) \circ \vec{w} = \sum_{j=1}^{p} \lambda_j w_j^2 \quad (57.44)$$

Mais *U* étant orthogonale nous avons par conséquent :

$$\left\{ \vec{w} = S^T \vec{u} \mid \vec{u} \in \mathbb{R}^p \quad \|\vec{u}\| = 1 \right\} \Rightarrow \left\{ \vec{w} \in \mathbb{R}^p \quad \|\vec{w}\| = 1 \right\} \quad (57.45)$$

et ceci provient du fait que la matrice orthogonales est comme nous l'avions démontré dans le chapitre d'algèbre linéaire une isométrie (elle conserve donc la norme!).

Comme les valeurs propres sont dans l'ordre croissant nous avons :

$$\lambda_1 \left(w_1^2 + \sum_{j=2}^{p} \frac{\lambda_j}{\lambda_1} w_j^2 \right) \quad (57.46)$$

Or le terme entre parenthèses est strictement inférieur ou égal à1. Donc :

$$\lambda_1\left(w_1^2+\sum_{j=2}^{p}\frac{\lambda_j}{\lambda_1}w_j^2\right)\leq\lambda_1 \quad (57.47)$$

Soit :

$$\sum_{j=1}^{p}\lambda_j w_j^2\leq\lambda_1 \quad (57.48)$$

Or rappelons que notre objectif est de maximiser cette inégalité. En d'autres termes de chercher w_1 tel que l'égalité soit respectée. Or nous voyons immédiatement que cela est faire si $w_1=1$. Ainsi, une solution de notre problème de maximisation est donc :

$$\vec{w}=\begin{pmatrix}1\\0\\0\\\vdots\\0\end{pmatrix} \quad (57.49)$$

soit puisque $\vec{w}=S^T\vec{u}=S^{-1}\vec{u}\Rightarrow\vec{u}=S\cdot\vec{w}$ qui est alors le premier vecteur propre de *R* (puisque *R* se diagonalise dans cette base) associé à la plus grande valeur propre λ_1. D'où le fait que cette solution soit notée souvent sous la forme $\vec{u}_1=S\cdot\vec{w}_1$ avec $\Lambda=S^{-1}RS$ (il est donc relativement aisé de déterminer *S* avec des logiciels lorsque *R* et Λ sont connus).

Une fois que l'on a trouvée la première droite vectorielle, nous cherchons une deuxième droite dans le sous-espace vectoriel orthogonal à la droite vectorielle qui maximise l'inertie du nuage de point projeté. Nous démontrons, et devinons, que la solution est donnée par la droite vectorielle dirigée par le vecteur propre associé à la deuxième valeur propre de la matrice de corrélation est ainsi de suite...

Ainsi, nous obtenons une nouvelle base $(\vec{u}_1,\ldots,\vec{u}_p)$ dont un des plans constitue le plan factoriel. Cependant il nous faut connaître les composantes de *Z* dans cette base. Comme cette base a été construite sous la condition que *R* y est diagonalisable via la matrice *S* alors cette dernière matrice est l'application linéaire qui va nous permettre d'exprimer *Z* dans la base $(\vec{u}_1,\ldots,\vec{u}_p)$ via la relation :

$$\psi=Z\cdot S \quad (57.50)$$

Ainsi, dans notre exemple les trois valeurs propres sont (cf. chapitre d'Algèbre Linéaire) :

$$\lambda_1 = 2.0303 \quad \lambda_2 = 0.8846 \quad \lambda_3 = 0.0853 \tag{57.51}$$

Remarque: Certains logiciels indiquent les poids en % respectifs et cumulés pour chacune des valeurs propres. Ainsi, nous avons dans le cas présent respectivement les poides suivants en % du total:

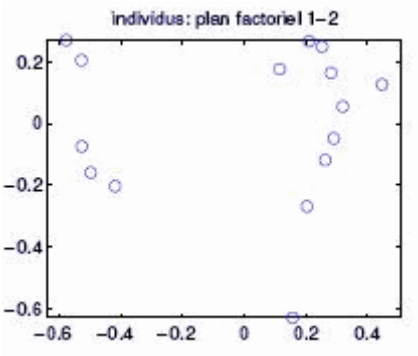
$$P_{\lambda 1} = \frac{2.0303}{\sum_i \lambda_i} = 66.67\% \quad P_{\lambda 2} = \frac{0.8846}{\sum_i \lambda_i} = 29.48\% \quad P_{\lambda 3} = \frac{0.0853}{\sum_i \lambda_i} = 2.84\% \tag{57.52}$$

Nous avons alors comme cordonnées des points M_i dans la base $(\vec{u}_1, \vec{u}_2, \vec{u}_3)$:

$$\psi = \begin{bmatrix} -0.5268 & 0.2045 & -0.0198 \\ -0.4178 & -0.2043 & -0.0564 \\ -0.5259 & -0.0750 & 0.0052 \\ -0.4966 & -0.1604 & 0.0305 \\ -0.5760 & 0.2699 & 0.0161 \\ 0.2525 & 0.2488 & -0.1169 \\ 0.1156 & 0.1757 & -0.0150 \\ 0.2818 & 0.1634 & -0.0916 \\ 0.1578 & -0.6311 & -0.0218 \\ 0.2634 & -0.1194 & -0.0871 \\ 0.2108 & 0.2660 & 0.1739 \\ 0.2039 & -0.2697 & 0.0912 \\ 0.4469 & 0.1261 & -0.0459 \\ 0.2908 & -0.0490 & 0.0712 \\ 0.3196 & 0.0546 & 0.0663 \end{bmatrix} \tag{57.53}$$

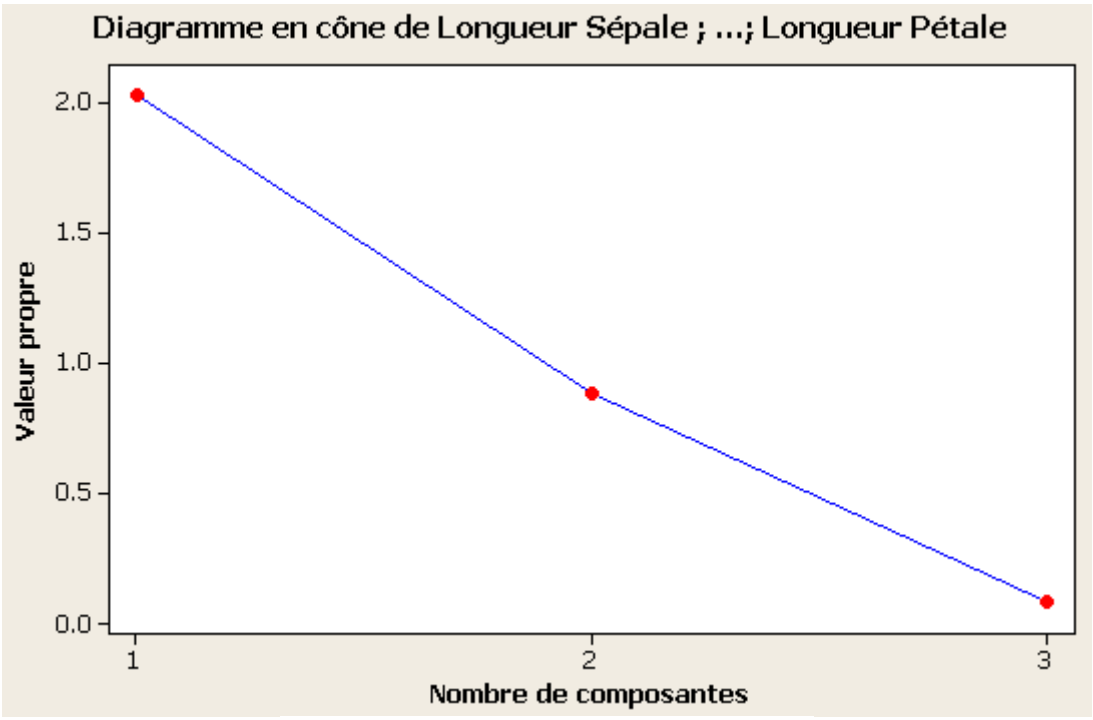
Les coordonnées des projections du nuage de points dans le meilleur plan défini par les vecteurs $(\vec{u}_1, \vec{u}_2)$ sont donc les deux premières colonnes de la matrice précédente (correspondant donc à la longueur du sépale et la largeur du sépale).

Effectivement nous voyons immédiatement que ce sont ces deux colonnes qui maximiseront la somme des normes dans le plan donné:



(57.54)

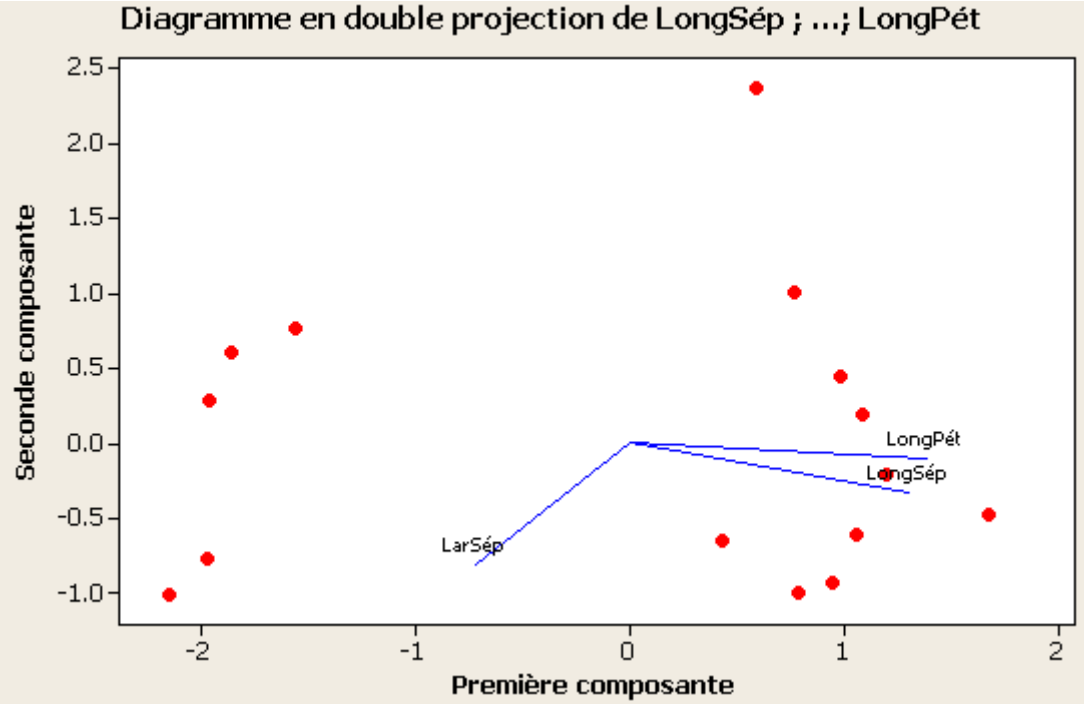
Un logiciel comme Minitab 15.1 (référence dans l'industrie de la gestion de la qualité) donne les informations suivantes pour les valeurs propres (info pas très utile sous forme graphique... à mon avis):



Valeur propre	2.0302	0.8846	0.0853
Proportion	0.677	0.295	0.028
Cumulatif	0.677	0.972	1.000

(57.55)

et le plan factoriel suivant (resterait à savoir comment les valeurs sont calculées car elles ne sont pas identiques à celles que nous avons obtenues ici... mais la forme graphique est bien juste et c'est le principal!):



(57.56)

1.2. L'analyse factorielle des correspondances (A.F.C)

L'analyse factorielle des correspondances, en abrégée AFC, est une méthode statistique d'analyse des données. La technique de l'AFC est essentiellement utilisée pour de grands tableaux de données toutes comparables entre elles (si possible exprimées toutes dans la même unité, comme une monnaie, une dimension, une fréquence ou toute autre grandeur mesurable). Elle peut en particulier permettre d'étudier des tableaux de contingence (ou tableau croisé de co-occurrence). Elle sert à déterminer et à hiérarchiser toutes les dépendances entre les lignes et les colonnes du tableau.

Voyons directement un exemple:

Considérons le tableau suivant des superficies des types de peuplements d'arbres en Picardie en 1984 en hectares:

	Feuillus	Résineux	Mixtes	Total par dép.
L'Aisne (A)	106'500	3'380	1'470	111'350
L'Oise (O)	101'700	10'000	0	111'700

La Somme (S)	45'200	4'350	50	49'600
Total	253'400	17'730	1'520	272'650

Tableau: 57.3 - Tableau de contingence (tableau croisé) de l'A.F.C.

Nous souhaitons analyser s'il existe les degrés de ressemblance et de différence entre les variables. Remarquons, que nous ne cherchons pas à comparer l'égalité des moyennes ou des variances donc les outils statistiques vus dans le chapitre du même nom ne sont pas adaptés à ce genre d'analyse.

Si nous choisissons la distance euclidienne:

$$d^2(x_i x_{i'}) = \left\| \overrightarrow{x_i x_{i'}} \right\|^2 = \sum_{j=1}^{p} \left(x_{ij} - x_{i'j} \right)^2 \quad (57.57)$$

sur les données brutes pour mesurer ces différences entre départements, nous obtenons les écarts suivants :

$$d^2(A,O) = \sum_{j=1}^{3} \left(x_{1j} - x_{2j} \right)^2 = \left(106'500 - 101'700\right)^2 + \left(3'380 - 10'000\right)^2 + \left(1'470 - 0\right)^2$$
$$= 69'025'300 \Rightarrow d(A,O) = \sqrt{69'025'300} = 8'308.1\ [ha] \quad (57.58)$$

et ainsi de suite pour les autres régions. Nous obtenons alors:

$$\begin{aligned} d(A,O) &= 8'308.1\ [ha] \\ d(A,S) &= 61'324.1\ [ha] \\ d(O,S) &= 56'781.8\ [ha] \end{aligned} \quad (57.59)$$

Nous voyons en regardant le tableau et avant tout calcul que les départements de l'Aisne et l'Oise se ressemblent alors que le département de la Somme se diffère nettement. Les distances obtenues mettent en évidence cette observation.

Mais! Pourtant, sur dans le tableau ci-dessus les profils de l'Oise et de la Somme, avec une forêt mixte très faible, sont pourtant très proches en proportion.

Dans ce contexte, nous voyons que la distance euclidienne transcrit les différences de masse entre les départements. En d'autres termes, l'Aisne et l'Oise se ressemblent car leurs superficies sont proches. Pour éliminer l'artefact

lié aux ordres de grandeur, il nous faut transformer les données en pourcentage. Nous obtenons alors:

	Feuillus	Résineux	Mixtes	%Région
Aisne	95.6	3.0	1.3	40.8
Oise	91.0	9.0	0.0	41.0
Somme	91.1	8.8	0.1	18.2

Tableau: 57.4 - Transformation du tableau de contingence en pourcents

Si nous choisissons la distance euclidienne sur les proportions (données relatives), nous obtenons:

$$d^2\left(x_i x_{i'}\right) = \left\|\overrightarrow{x_i x_{i'}}\right\|^2 = \sum_{j=1}^{p}\left(\frac{x_{ij}}{x_{i.}} - \frac{x_{i'j}}{x_{i'.}}\right)^2 = \sum_{j=1}^{p}\left(\frac{x_{ij}}{\sum_j x_{ij}} - \frac{x_{i'j}}{\sum_j x_{i'j}}\right)^2 \quad (57.60)$$

soit:

$$\begin{aligned} d(A,O) &= 7.6\% \\ d(A,S) &= 7.4\% \\ d(O,S) &= 0.2\% \end{aligned} \quad (57.61)$$

Cette fois, l'Oise et la Somme apparaissent bien comme se ressemblant le plus avec leurs forêts. Nous voyons que travailler avec les données relatives semblent donc plus pertinent dans ce cas!

Maintenant, nous allons emprunter une idée des économistes qui lorsqu'ils ont des tableaux du même genre que le précédent calculent ce qu'ils appellent "l'index" ou "élasticité" et qui est donné par:

$$I = \frac{\dfrac{x_{ij}}{x_{i.}}}{\dfrac{x_{.j}}{x_{..}}} = \frac{\dfrac{x_{ij}}{x_{j.}}}{\dfrac{x_{i.}}{x_{..}}} \quad (57.62)$$

Voici un exemple obtenu avec les tableaux croisés dynamiques de MS Excel qui inclut la fonction Index:

Sum	Region			
Item	Alberta	Ontario	Quebec	Total
Binder	1'279.36	5'762.63	2'535.66	9'577.65
Desk	825.00	875.00		1'700.00
Pen	151.24	539.73	1'354.25	2'045.22
Pen Set		2'421.39	1'748.48	4'169.87
Pencil	231.12	1'540.32	363.70	2'135.14
Total	2'486.72	11'139.07	6'002.09	19'627.88

(57.63)

et en activant la fonction Index:

Index	Region		
Item	Alberta	Ontario	Quebec
Binder	1.05	1.06	0.87
Desk	3.83	0.91	-
Pen	0.58	0.47	2.17
Pen Set	-	1.02	1.37
Pencil	0.85	1.27	0.56

(57.64)

Pour voir d'où viennent ces valeurs, regardons par exemple l'article *Desk* dans la région *Alberta* a un rendement de:

$$\frac{x_{21}}{x_{2.}} = \frac{825}{1700} \cong 48.52\% \qquad (57.65)$$

par rapport à toutes les régions ce qui est au-dessus de la valeur de 33.33% qu'aurait comme rendement cette article dans toutes les régions confondues s'il n'y avait pas de préférences de région!

La région *Alberta* a elle un rendement de:

$$\frac{x_{.1}}{x_{..}} = \frac{2486.72}{19627.88} \cong 12.66\% \qquad (57.66)$$

par rapport à toutes les régions ce qui est en-dessous des 33.33% de rendement qu'elle aurait s'il n'y avait de préférences de région. Ainsi, ce tableau d'index permet de savoir si les différences sont significatives!!

Le rapport donne donc:

$$I = \frac{\frac{x_{21}}{x_{2.}}}{\frac{x_{1}}{x_{..}}} \cong 3.83 \qquad (57.67)$$

ce qui montre un fort décalage entre la valeur obtenue et la valeur que nous aurions si les proportions étaient respectées.

C'est donc une sorte de calcul de conformité: si le rapport valait 1, c'est que le rendement régional des ventes de cet article particulier serait conforme au rapport de toutes les ventes de cette région relativement à un marché national. Il n'y aurait alors pas d'anomalies Voyons cela par exemple pour nos arbres où nous avions les effectifs observés:

	Feuillus	Résineux	Mixtes	Total par dép.
L'Aisne (A)	106'500	3'380	1'470	111'350
L'Oise (O)	101'700	10'000	0	111'700
La Somme (S)	45'200	4'350	50	49'600
Total	253'400	17'730	1'520	272'650

Tableau: 57.5 - Tableau de contingence (tableau croisé) de l'A.F.C.

et pour lequel nous obtenons le tableau des index effectifs observés suivant dans MS Excel:

Index	Arbre		
Région	Feuillus	Résineux	Mixtes
La Somme (S)	0.98	1.35	0.18
L'Aisne (A)	1.03	0.47	2.37
L'Oise (O)	0.98	1.38	-

(57.68)

et nous voyons encore clairement à l'aide de ce tableau que ce sont l'Oise et la Somme qui se ressemblent le plus!

Avant de continuer, nous pourrions nous poser la question extrêmement importante suivante: Quels seraient les effectifs théoriques qui auraient été obtenus si les proportions des arbres dans les régions étaient rigoureusement

équivalentes à la proportion d'ensemble (soit de telle manière à ce que les index soient tous unitaires)?

Eh bien simplement en faisant le calcul suivant:

	Feuillus	Résineux	Mixtes
Aisne	=(253'400/272'650)*111'350 =103'488	=(17'730/272'650)*111'350 =7'240	=(1'470/272'650)*111'350 =620
Oise	=(253'400/272'650)*111'700 =103'813	=(17'730/272'650)*111'700 =7'263	=(1'470/272'650)*111'700 =622
Somme	=(253'400/272'650)*49'600 =46'098	=(17'730/272'650)*49'600 =3'225	=(1'470/272'650)*49'600 =276

Tableau: 57.6 - Respect des proportions de l'A.F.C.

Et nous obtenons avec ces nouvelles valeurs le tableau des index des effectifs théoriques suivant dans MS Excel:

Index	Arbre		
Région	Feuillus	Résineux	Mixtes
La Somme (S)	1.00	1.00	1.00
L'Aisne (A)	1.00	1.00	1.00
L'Oise (O)	1.00	1.00	1.00

(57.69)

ce qui montre que les proportions sont maintenant respectées! Paranthèse fermée (mais sur laquelle nous reviendrons un peu plus loin)!

Eh bien quand nous voulons faire de l'analyse factorielle de correspondance, notre relation:

$$d^2\left(x_i x_{i'.}\right) = \sum_{j=1}^{p}\left(\frac{x_{ij}}{x_{i.}} - \frac{x_{i'j}}{x_{i'.}}\right)^2 \qquad (57.70)$$

devient alors:

$$d^2(x_i x_{i'.}) = \sum_{j=1}^{p} \frac{\left(\frac{x_{ij}}{x_{i.}} - \frac{x_{i'j}}{x_{i'.}}\right)^2}{\frac{x_{.j}}{x_{..}}} \quad (57.71)$$

soit:

$$\begin{aligned} d(A,O) &= 0.30 \\ d(A,S) &= 0.39 \\ d(O,S) &= 0.02 \end{aligned} \quad (57.72)$$

Cette fois encore, l'Oise et la Somme apparaissent bien comme se ressemblant le plus.

La distance ci-dessus se nomme la "métrique du Khi-2" car elle ressemble (mais c'est tout!) à la distance utilisée dans le test d'ajustement du même nom (cf. chapitre de Statistiques) mais ici, elle permet seulement de mettre en place une hiérarchie dans le cadre d'un tableau de contingences et d'observer les variables similaires de manière plus aisée!!

2. Remarque importante

Pour calculer rapidement une A.C.P, il est plus pratique d'utiliser la statistique descriptive pour trouver facilement les moyennes, les valeurs centrées, la matrice des variances-covariances, l'inertie totale et les axes principaux.

V. EXERCICES SUR L'A.C.P

Ces quatre exercices résolus appartiennent à la série d'exercices du professeur Henri Immediato de l'université Claude Bernard de Lyon. Nous les avons choisis parce qu'ils développent abondamment l'A.C.P en termes de calculs et l'explique en détails.

Exercice n°1

Calculer et dessiner le premier axe principal du nuage de huit points de R^2 défini par

$$\begin{pmatrix} 2 & 5 & 8 & 3 & -2 & -5 & -8 & -3 \\ -4 & 0 & 4 & 4 & 4 & 0 & -4 & -4 \end{pmatrix}.$$

Solution

Soit $M = \begin{pmatrix} 2 & 5 & 8 & 3 & -2 & -5 & -8 & -3 \\ -4 & 0 & 4 & 4 & 4 & 0 & -4 & -4 \end{pmatrix}$ la matrice des données.
Chaque ligne correspond à une variable, chaque colonne correspond à un individu.
Pour chaque individu, on mesure la valeur des deux variables.
Chaque individu est représenté par un point de R^2.
Chaque variable est représentée par un point de R^8.
Comme rien n'est précisé dans l'énoncé, on supposera que chaque individu possède le **même poids statistique** de $\frac{1}{8}$.

La matrice diagonale D dont tous les éléments de la diagonale sont égaux à $\frac{1}{8}$:

$$D = \frac{1}{8} Id,$$

où Id est la matrice unité de R^8, définit un **produit scalaire** dans R^8.

1°/ Données centrées.

La moyenne d'une variable est le produit scalaire de cette variable par le vecteur 1 :

$$\overline{X} = < X \mid 1 > = {}^tX\, D\, 1$$
$$\overline{Y} = < Y \mid 1 > = {}^tY\, D\, 1$$

Appelons □ la matrice à 8 lignes et 2 colonnes, des données.
□ est la transposée de *M*.
La première colonne est le vecteur *X*, la deuxième colonne est le vecteur *Y*.

$$\square = \begin{pmatrix} 2 & -4 \\ 5 & 0 \\ 8 & 4 \\ 3 & 4 \\ -2 & 4 \\ -5 & 0 \\ -8 & -4 \\ -3 & -4 \end{pmatrix}$$

La matrice des moyennes $\overline{\square} = (\overline{X}\ \overline{Y})$ s'exprime par le produit de matrices :

$\overline{\square} = (\overline{X}\ \overline{Y}) = {}^t1\ D\ \square$.

$$= \frac{1}{8}(1\ \ 1\ \ 1\ \ 1\ \ 1\ \ 1\ \ 1\ \ 1) \begin{pmatrix} 2 & -4 \\ 5 & 0 \\ 8 & 4 \\ 3 & 4 \\ -2 & 4 \\ -5 & 0 \\ -8 & -4 \\ -3 & -4 \end{pmatrix}$$
$$= \frac{1}{8}(0\ \ 0)$$
$$= (0\ \ 0)$$

<u>Les données sont centrées</u>.

La matrice *Z* des données centrées est donc la matrice □ des données : *Z* = □.

2°/ Matrice des variances-covariances.

La matrice *C* des variances-covariances est donnée par la formule :

$C = {}^tZ\ D\ Z$

$$= \frac{1}{8}\begin{pmatrix} 2 & 5 & 8 & 3 & -2 & -5 & -8 & -3 \\ -4 & 0 & 4 & 4 & 4 & 0 & -4 & -4 \end{pmatrix} \begin{pmatrix} 2 & -4 \\ 5 & 0 \\ 8 & 4 \\ 3 & 4 \\ -2 & 4 \\ -5 & 0 \\ -8 & -4 \\ -3 & -4 \end{pmatrix}$$
$$= \frac{1}{8}\begin{pmatrix} 204 & 72 \\ 72 & 96 \end{pmatrix}$$

$$C = \begin{pmatrix} 25{,}5 & 9 \\ 9 & 12 \end{pmatrix}$$

3°/ Valeurs propres.

Les valeurs propres λ de la matrice des variances-covariances sont les nombres réels qui annulent le déterminant de la matrice $C - \lambda Id$, où Id est la matrice unité de R^8.
La somme des valeurs propres est la trace 25,5 + 12 = 37,5 de C.
Le produit des valeurs propres est le déterminant 25,5 × 12 – 9 × 9 = 225 de C.
Les valeurs propres sont solutions de l'équation aux valeurs propres :

$$\lambda^2 - 37{,}5\,\lambda + 225 = 0$$
$$2\,\lambda^2 - 75\,\lambda + 450 = 0$$
$$\lambda_1 = \frac{1}{4}(75 + \sqrt{75^2 - 8 \times 450}) = 30$$
$$\lambda_2 = \frac{1}{4}(75 - \sqrt{75^2 - 8 \times 450}) = 7{,}5$$

4°/ Premier axe principal.

Le premier axe principal est déterminé par un vecteur propre relatif à la plus grande valeur propre λ_1 de la matrice C des variances-covariances.

Rappelons que, pour une matrice quelconque $C = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, si λ est une valeur propre de C, le vecteur $u = \begin{pmatrix} d - \lambda \\ -c \end{pmatrix}$ vérifie :

$$(C - \lambda Id)\,u = \begin{pmatrix} a-\lambda & b \\ c & d-\lambda \end{pmatrix}\begin{pmatrix} d-\lambda \\ -c \end{pmatrix} = \begin{pmatrix} (a-\lambda)(d-\lambda) - bc \\ cd - c\lambda - dc + c\lambda \end{pmatrix} = \begin{pmatrix} Det(C - \lambda\, Id) \\ 0 \end{pmatrix}$$

Or, comme λ est valeur propre, λ vérifie l'équation aux valeurs propres

$$Det\,(C - \lambda\, Id) = \lambda^2 - (a + d)\lambda + (a\,d - b\,c) = 0$$

et la relation précédente se réduit à

$$(C - \lambda\, Id)\,u = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$
$$C\,u = \lambda\,u$$

ce qui montre que le vecteur $u = \begin{pmatrix} d-\lambda \\ -c \end{pmatrix}$ est vecteur propre de la matrice C pour la valeur propre ?.

Le vecteur $\begin{pmatrix} 12-30 \\ -9 \end{pmatrix} = -9\begin{pmatrix} 2 \\ 1 \end{pmatrix}$ est donc vecteur propre de la matrice $C = \begin{pmatrix} 25{,}5 & 9 \\ 9 & 12 \end{pmatrix}$.

Il en est donc de même du vecteur $\begin{pmatrix} 2 \\ 1 \end{pmatrix}$.

La norme euclidienne de ce vecteur de R^2 est $\sqrt{2^2+1^2} = \sqrt{5}$.

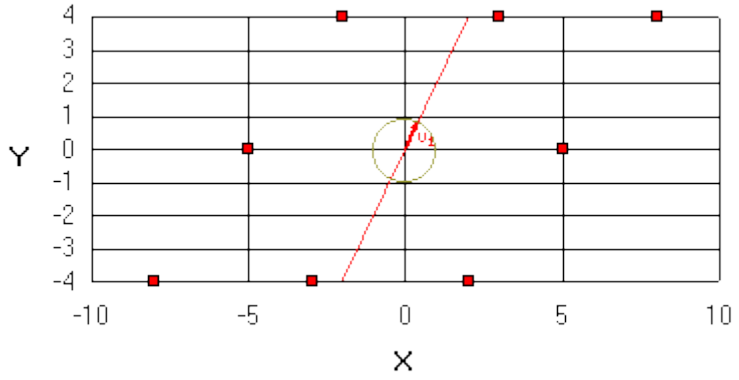
Le premier axe principal est donc défini par le vecteur normé :

$$u_1 = \frac{1}{\sqrt{5}}\begin{pmatrix} 2 \\ 1 \end{pmatrix} = \frac{\sqrt{5}}{5}\begin{pmatrix} 2 \\ 1 \end{pmatrix}.$$

$$\boxed{u_1 = \frac{\sqrt{5}}{5}\begin{pmatrix} 2 \\ 1 \end{pmatrix}.}$$

5°/ Représentation graphique.

Dans R^2, on peut maintenant représenter graphiquement le nuage de points et son premier axe principal :



Propriété.

Le premier axe principal est la **droite de régression orthogonale** : c'est la droite pour laquelle la somme des carrés des distances des points du nuage à la droite est la plus petite possible.

Exercice n°2

Soient deux variables à six valeurs X = (0, 1, 0, 1, 1, 0) et Y = (0, 1, 1, 0, 1, 0). Calculer moyennes, variances, covariance. Dessiner le nuage centré, placer les deux droites de régression et les deux axes principaux. Calculer l'inertie statistique du nuage de points par rapport aux deux axes principaux.	

Solution

Par convention, puisque rien n'est précisé dans le texte, les pondérations des individus et des variables sont implicites : $\frac{1}{6}$ pour les variables, 1 pour les individus.

Autrement dit :

- dans l'espace R^6 des variables, le produit scalaire est défini par la matrice $D = \frac{1}{6} Id$, où Id est la matrice unité de R^6,
- dans l'espace R^2 des individus, le produit scalaire est le produit scalaire euclidien canonique, défini par la matrice unité de R^2.

1°/ Moyennes.

Soit 1 le vecteur de R^6 dont les six coordonnées sont égales à 1.
Soit Ω la matrice des données :

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

Les moyennes des variables s'obtiennent par produit scalaire avec le vecteur 1.
La matrice des moyennes est donc :

$$\overline{\Delta} = (\overline{X} \ \ \overline{Y}) = {}^t1 \, D \, \Omega = \frac{1}{6}(1 \ \ 1 \ \ 1 \ \ 1 \ \ 1 \ \ 1) \begin{pmatrix} 0 & 0 \\ 1 & 1 \\ 0 & 1 \\ 1 & 0 \\ 1 & 1 \\ 0 & 0 \end{pmatrix} = \frac{1}{6}(3 \ \ 3) = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \end{pmatrix}$$

$$\overline{X} = \frac{1}{2};\ \overline{Y} = \frac{1}{2}$$

2°/ Données centrées.

Les données centrées s'obtiennent par la formule :

$$Z = (Id - 1\ {}^t1\ D)$$

où *Id* est la matrice identité de R^6, matrice diagonale à six lignes et six colonnes, dont tous les éléments de la diagonale sont égaux à 1 et les autres à 0.

$$Z = \frac{1}{6}\begin{pmatrix} 5 & -1 & -1 & -1 & -1 & -1 \\ -1 & 5 & -1 & -1 & -1 & -1 \\ -1 & -1 & 5 & -1 & -1 & -1 \\ -1 & -1 & -1 & 5 & -1 & -1 \\ -1 & -1 & -1 & -1 & 5 & -1 \\ -1 & -1 & -1 & -1 & -1 & 5 \end{pmatrix}\begin{pmatrix} 0 & 0 \\ 1 & 1 \\ 0 & 1 \\ 1 & 0 \\ 1 & 1 \\ 0 & 0 \end{pmatrix} = \frac{1}{2}\begin{pmatrix} -1 & -1 \\ 1 & 1 \\ -1 & 1 \\ 1 & -1 \\ 1 & 1 \\ -1 & -1 \end{pmatrix}$$

3°/ Matrice des variances covariances.

La matrice *C* des variances-covariances s'obtient par la formule :

$$C = {}^tZ\ D\ Z$$

On obtient :

$$C = \frac{1}{24}\begin{pmatrix} -1 & 1 & -1 & 1 & 1 & -1 \\ -1 & 1 & 1 & -1 & 1 & -1 \end{pmatrix}\begin{pmatrix} -1 & -1 \\ 1 & 1 \\ -1 & 1 \\ 1 & -1 \\ 1 & 1 \\ -1 & -1 \end{pmatrix} = \frac{1}{12}\begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix} = \begin{pmatrix} s^2(X) & Cov(X,Y) \\ Cov(X,Y) & s^2(Y) \end{pmatrix}$$

Sur cette matrice, on voit les variances et la covariance :

$$s^2(X) = ||\ X_0\ ||^2 = \frac{1}{4};\ s^2(Y) = ||\ Y_0\ ||^2 = \frac{1}{4};\ Cov(X,Y) = \langle X_0 \mid Y_0 \rangle = \frac{1}{12}$$

où X_0 et Y_0 sont les variables centrées, et où normes et produits scalaires sont définis par la matrice *D*.

4°/ Régression linéaire.

a) Régression linéaire de *Y* en *X*.

Dans la régression linéaire de *Y* en *X*, *Y* est la variable à expliquer, *X* la variable explicative.
La matrice des variables explicatives centrées se réduit à un vecteur :

$$X_0 = \frac{1}{2}\begin{pmatrix} -1 \\ 1 \\ -1 \\ 1 \\ 1 \\ -1 \end{pmatrix}$$

La matrice *c* des variances-covariances des variables explicatives se réduit à un nombre

$$c = s^2(X) = \| X_0 \|^2 = \frac{1}{4}$$

et son inverse se réduit à $c^{-1} = 4$.

La variable expliquée centrée est :

$$Y_0 = \frac{1}{2}\begin{pmatrix} -1 \\ 1 \\ 1 \\ -1 \\ 1 \\ -1 \end{pmatrix}$$

Le vecteur *d* des covariances de la variable expliquée et des variables explicatives se réduit à un nombre :

$$d = {}^tX_0\, D\, Y_0 = \langle X_0 \mid Y_0 \rangle = \frac{1}{24}(-1 \;\; 1 \;\; -1 \;\; 1 \;\; 1 \;\; -1)\begin{pmatrix} -1 \\ 1 \\ 1 \\ -1 \\ 1 \\ -1 \end{pmatrix} = \frac{1}{12} = Cov(X, Y).$$

La matrice *b* des coefficients de régression linéaire de *Y* par rapport aux variables explicatives se réduit à un nombre :

$$b = c^{-1} d = \frac{1}{3} = \frac{\langle X_0 | Y_0 \rangle}{\| X_0 \|^2} = \frac{Cov(X, Y)}{s^2(X)}$$

et le terme constant est :

$$a = \overline{Y} - b\ \overline{X} = \frac{1}{2} - \frac{1}{3} \times \frac{1}{2} = \frac{1}{3}$$

La droite régression de *Y* par rapport à *X* a donc pour équation : $y = \frac{1}{3}x + \frac{1}{3}$, soit encore :

$$\boxed{x - 3\,y + 1 = 0}$$

b) Régression linéaire de *X* en *Y*.

Le coefficient de régression linéaire est :

$$b = \frac{Cov(X,Y)}{s^2(Y)} = \frac{1}{3}$$

et le terme constant est :

$$a = \overline{X} - b\ \overline{Y} = \frac{1}{2} - \frac{1}{3} \times \frac{1}{2} = \frac{1}{3}$$

La droite de régression de *X* par rapport à *Y* a donc pour équation : $x = \frac{1}{3}y + \frac{1}{3}$, soit encore :

$$\boxed{3\,x - y - 1 = 0}$$

On remarquera que le produit des coefficients de régression linéaire de *Y* en *X* et de *X* en *Y* est $r_{XY}^2 = \left(\frac{Cov(X,Y)}{s(X)\,s(Y)} \right)^2 = \frac{1}{9}$.

5°/ Axes principaux.

a) Valeurs propres de la matrice des variances-covariances.

$$C = \frac{1}{12} \begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix}$$

La somme des valeurs propres de la matrice *C* est la trace de la matrice : c'est la somme des éléments de la diagonale, elle vaut $\frac{1}{2}$.

Le produit des valeurs propres est le déterminant de la matrice *C* : il vaut $\frac{8}{12^2} = \frac{1}{18}$.

Les valeurs propres sont les racines de l'équation aux valeurs propres :

$$\lambda^2 - \frac{1}{2}\lambda + \frac{1}{18} = 0$$
$$18\lambda^2 - 9\lambda + 1 = 0$$
$$\lambda = \frac{1}{36}(9 \pm \sqrt{81 - 4 \times 18}) = \frac{1}{36}(9 \pm 3)$$
$$\lambda_1 = \frac{1}{3}$$
$$\lambda_2 = \frac{1}{6}$$

b) Vecteurs propres.

On sait désormais, pour l'avoir déjà vérifié dans l'exercice n°1, que le vecteur $\begin{pmatrix} d-\lambda \\ -c \end{pmatrix}$ est vecteur propre de la matrice $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ pour la valeur propre λ.

On peut donc prendre pour vecteur propre pour la valeur propre $\lambda_1 = \frac{1}{3}$, le vecteur $\begin{pmatrix} \frac{1}{4} - \frac{1}{3} \\ -\frac{1}{12} \end{pmatrix} = -\frac{1}{12}\begin{pmatrix} 1 \\ 1 \end{pmatrix}$, ou encore, tout simplement, le vecteur $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$.

Le vecteur normé correspondant est $u_1 = \frac{\sqrt{2}}{2}\begin{pmatrix} 1 \\ 1 \end{pmatrix}$.

On peut prendre pour vecteur propre pour la valeur propre $\lambda_2 = \frac{1}{6}$, le vecteur $\begin{pmatrix} \frac{1}{4} - \frac{1}{6} \\ -\frac{1}{12} \end{pmatrix} = \frac{1}{12}\begin{pmatrix} 1 \\ -1 \end{pmatrix}$, ou encore, tout simplement, le vecteur $\begin{pmatrix} 1 \\ -1 \end{pmatrix}$.

Le vecteur normé correspondant est $u_2 = \frac{\sqrt{2}}{2}\begin{pmatrix} 1 \\ -1 \end{pmatrix}$.

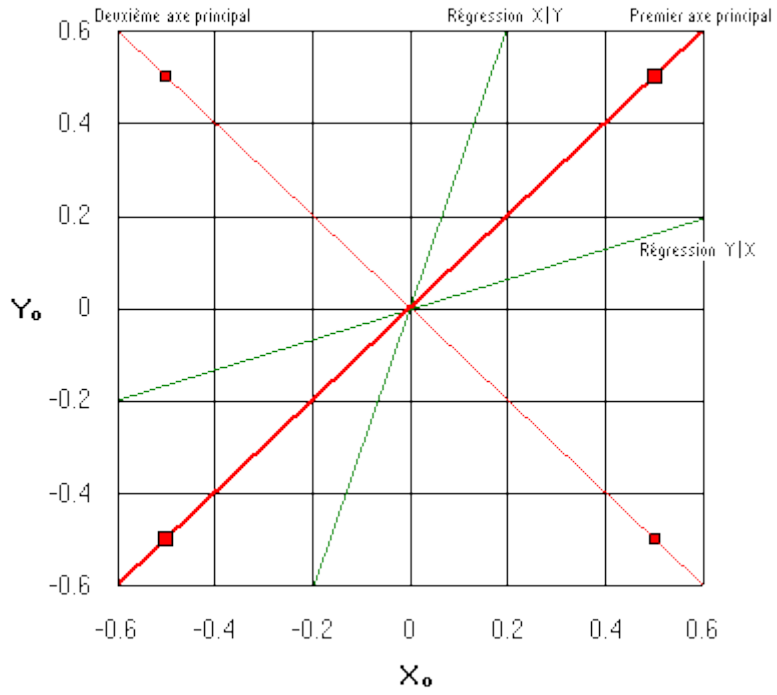
Les deux vecteurs

$$u_1 = \frac{\sqrt{2}}{2}\begin{pmatrix} 1 \\ 1 \end{pmatrix} \text{ et } u_2 = \frac{\sqrt{2}}{2}\begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

forment, dans $\mathbb{R}^2$, une base orthonormée pour le produit scalaire canonique. Leurs directions déterminent les **axes principaux** : ceux-ci sont portés par les deux bissectrices du plan rapporté à deux axes canoniques.

6°/ Représentation graphique.

Portons, sur un même graphique, le nuage centré de points, les axes principaux, les droites de régression centrées (sans terme constant) :



7°/ Inertie statistique.

L'inertie statistique du nuage de points par rapport à un axe défini par un vecteur unitaire *u* est la moyenne des carrés des normes des projetés orthogonaux sur *u*.
Si M_i est un point du nuage, et *G* le centre de gravité du nuage, le vecteur $\overrightarrow{GM_i}$ a, pour coordonnées, la i^e ligne de la matrice *Z* des données centrées, soit $(x_i - \bar{X}, y_i - \bar{Y})$.

Pour un vecteur unitaire $u = \begin{pmatrix} \cos\alpha \\ \sin\alpha \end{pmatrix}$, le projeté orthogonal de $\overrightarrow{GM_i}$ sur *u* est

$$\overrightarrow{Gm_i} = \langle \overrightarrow{GM_i} \mid u \rangle u = ((x_i - \bar{X}) \cos\alpha + (y_i - \bar{Y}) \sin\alpha) \begin{pmatrix} \cos\alpha \\ \sin\alpha \end{pmatrix}.$$

Le carré de la norme de $\overrightarrow{Gm_i}$ est :

$$|| \overrightarrow{Gm_i} ||^2 = (x_i - \bar{X})^2 \cos^2\alpha + (y_i - \bar{Y})^2 \sin^2\alpha + 2 \sin\alpha \cos\alpha (x_i - \bar{X})(y_i - \bar{Y})$$

et l'inertie statistique du nuage de points par rapport à u est :

$$I_S(u) = s^2(X)\cos^2\theta + s^2(Y)\sin^2\theta + 2\sin\theta\cos\theta\,Cov(X,Y) = ||X_0\cos\theta + Y_0\sin\theta||^2$$

la norme étant prise dans R^6.

$X_0\cos\theta + Y_0\sin\theta$ est le vecteur $Z\,u$ de R^6.
En introduisant la matrice $C = {}^tZ\,D\,Z$, des variances-covariances, on obtient la formule :

$$I_S(u) = ||Z\,u||^2 = {}^t(Z\,u)\,D\,(Z\,u) = {}^tu\,({}^tZ\,D\,Z)\,u = {}^tu\,C\,u.$$

$$\boxed{I_S(u) = {}^tu\,C\,u}$$

Pour un axe principal, u est vecteur propre de C pour une valeur propre λ, donc $C\,u = \lambda\,u$

$$I_S(u) = {}^tu\,C\,u = \lambda\,{}^tu\,u = \lambda\,||u||^2 = \lambda.$$

Ainsi :

$$\boxed{\begin{aligned} I_S(u_1) &= \lambda_1 = \frac{1}{3} \\ I_S(u_2) &= \lambda_2 = \frac{1}{6} \end{aligned}}$$

Exercice n°3

Etant données deux variables X et Y mesurées sur n individus, existe-t'il deux réels a et b vérifiant la relation $a^2 + b^2 = 1$, qui maximisent la variance de la variable $a X + b Y$?	

Solution

1e remarque.

Comme la variance de $-(a X + b Y)$ est égale à la variance de $a X + b Y$, on peut limiter l'étude au cas où ***b* est positif**.

2e remarque.

Etant donnés deux réels a et b vérifiant $a^2 + b^2 = 1$, avec b positif, il existe toujours un réel θ compris entre 0 et π tel que l'on ait $a = \cos\theta$ et $b = \sin\theta$.

1°/ Expression de la variance de *a X + b Y*.

$$s^2(a X + b Y) = a^2 s^2(X) + b^2 s^2(Y) + 2 a b Cov(X, Y) = (a \;\; b) \begin{pmatrix} s^2(X) & Cov(X, Y) \\ Cov(X, Y) & s^2(Y) \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix}$$

2°/ Réduction du problème en termes d'approche euclidienne.

La variance de la variable $a X + b Y$ est égale à la variance de la variable centrée $a X_0 + b Y_0$ correspondante.
Soit Z la matrice centrée des données (matrice à n lignes et deux colonnes) : la première colonne est la variable centrée X_0, la deuxième colonne est la variable centrée Y_0.

Dans l'espace $\mathbb{R}^n$ des variables :

$$a X_0 + b Y_0 = Z \begin{pmatrix} a \\ b \end{pmatrix}.$$

Le problème est ainsi ramené à la recherche d'un vecteur unitaire $u = \begin{pmatrix} a \\ b \end{pmatrix}$ de $\mathbb{R}^2$ ($a^2 + b^2 = 1$) tel que la variance de la variable $Z u$ soit maximum.

La matrice diagonale $D_{\frac{1}{n}} = \frac{1}{n} Id_n$ définit le produit scalaire de $\mathbb{R}^n$.
La variance d'une variable centrée est le carré de sa norme pour le produit

scalaire de R^n.

La variance de la variable $Z\,u$ est $||\,Z\,u\,||^2 = <Z\,u \mid Z\,u>^{D_{\frac{1}{n}}} = \frac{1}{n}{}^t(Z\,u)\,Z\,u = \frac{1}{n}{}^t u\,{}^t Z\,Z\,u = {}^t u\,{}^t Z\,D_{\frac{1}{n}}\,Z\,u$.

Or ${}^t Z\,D_{\frac{1}{n}}\,Z = C = \begin{pmatrix} s^2(X) & Cov(X,Y) \\ Cov(X,Y) & s^2(Y) \end{pmatrix}$ est la matrice des variances-covariances de X et Y, donc la variance de Zu est ${}^t u\,C\,u$.
Comme on l'a vu dans l'exercice 2, c'est l'inertie statistique du nuage de points par rapport au vecteur u.

$$s^2(a\,X + b\,Y) = {}^t u\,C\,u = I_S(u)$$

Le problème est donc de trouver dans R^2, un vecteur unitaire u tel que $I_S(u) = {}^t u\,C\,u$ soit maximum.

3°/ Résolution du problème.

Soient u_1 et u_2 une base orthonormale de R^2 formée de vecteurs propres de la matrice C des variances-covariances de X et Y : u_1 et u_2 définissent les axes principaux du nuage de points dans R^2.
On suppose que u_1 définit le premier axe principal, correspondant à la plus grande valeur propre λ_1 de C, et u_2 le deuxième axe principal, correspondant à la plus petite valeur propre λ_2 de C.

Tout vecteur unitaire u peut s'écrire :

$$u = <u \mid u_1>\,u_1 + <u \mid u_2>\,u_2.$$

${}^t u\,C\,u = {}^t(<u \mid u_1>\,u_1 + <u \mid u_2>\,u_2)\,C\,(<u \mid u_1>\,u_1 + <u \mid u_2>\,u_2)$
$= (<u \mid u_1>)^2\,{}^t u_1\,C\,u_1 + (<u \mid u_2>)^2\,{}^t u_2\,C\,u_2 + 2<u \mid u_1><u \mid u_2>\,{}^t u_1\,C\,u_2$

Or u_1 est vecteur propre de C pour la valeur propre λ_1 et u_2 est vecteur propre de C pour la valeur propre λ_2, donc la relation précédente s'écrit :

${}^t u\,C\,u = (<u \mid u_1>)^2\,\lambda_1\,{}^t u_1\,u_1 + (<u \mid u_2>)^2\,\lambda_2\,{}^t u_2\,u_2 + 2<u \mid u_1><u \mid u_2>\,\lambda_2\,{}^t u_1\,u_2$.

Puisque les vecteurs u_1 et u_2 forment une base orthonormale de R^2, leur produit scalaire euclidien ${}^t u_1\, u_2$ est nul, leurs normes euclidiennes ${}^t u_1\, u_1$ et ${}^t u_2\, u_2$ sont égale à 1 et il reste :

$$ {}^t u\, C\, u = (< u \mid u_1 >)^2 \lambda_1 + (< u \mid u_2 >)^2 \lambda_2 $$

Notons θ **l'angle des deux vecteurs *u* et u_1**, on a $< u \mid u_1 > = \cos\theta$ et $| < u \mid u_2 > | = | \sin\theta |$.
La relation précédente s'écrit :

$$ I_S(u) = {}^t u\, C\, u = \lambda_1 \cos^2\theta + \lambda_2 \sin^2\theta. $$

Le problème est donc ramené à trouver un angle θ, compris entre 0 et π, tel que $\lambda_1 \cos^2\theta + \lambda_2 \sin^2\theta$ soit maximum.
Or, on peut écrire :

$$ \lambda_1 \cos^2\theta + \lambda_2 \sin^2\theta = \lambda_1 - (\lambda_1 - \lambda_2) \sin^2\theta, $$

Le terme $(\lambda_1 - \lambda_2) \sin^2\theta$ est un terme positif, nul seulement pour $\theta = 0$ ou π.
$\lambda_1 - (\lambda_1 - \lambda_2) \sin^2\theta$ possède donc pour valeur maximum λ_1, et cette valeur maximum est atteinte lorsque *u* et u_1 sont colinéaires : $u = u_1$ ou $u = -u_1$.

La réponse à notre problème initial est donc positive.

Il existe bien des réels *a* et *b* vérifiant $a^2 + b^2 = 1$ et tels que la variance de $a X + b Y$ soit maximum.

a et *b* sont les coordonnées d'un vecteur unitaire u_1 définissant le premier axe principal du nuage de points.

Par exemple :

$$ u_1 = \frac{1}{\sqrt{\left(s^2(Y) - \lambda_1\right)^2 + \left(Cov(X,Y)\right)^2}} \begin{pmatrix} s^2(Y) - \lambda_1 \\ -Cov(X,Y) \end{pmatrix} $$

$$ \lambda_1 = \frac{1}{2}\left(s^2(X) + s^2(Y) + \sqrt{\left(s^2(X) - s^2(Y)\right)^2 + 4\left(Cov(X,Y)\right)^2} \right) $$

mais on pourrait prendre aussi le vecteur $-u_1$, qui nous donne :

$$a = \frac{s^2(Y) - s^2(X) + \sqrt{\left(s^2(Y) - s^2(X)\right)^2 + 4\left(Cov(X,Y)\right)^2}}{\sqrt{\left(s^2(Y) - s^2(X) + \sqrt{\left(s^2(Y) - s^2(X)\right)^2 + 4\left(Cov(X,Y)\right)^2}\right)^2 + \left(Cov(X,Y)\right)^2}}$$

$$b = \frac{Cov(X,Y)}{\sqrt{\left(s^2(Y) - s^2(X) + \sqrt{\left(s^2(Y) - s^2(X)\right)^2 + 4\left(Cov(X,Y)\right)^2}\right)^2 + \left(Cov(X,Y)\right)^2}}$$

Avec ces valeurs de *a* et *b*, la variance de *a X + b Y* atteint sa valeur maximum □

$_1 = \frac{1}{2}\left(s^2(X) + s^2(Y) + \sqrt{\left(s^2(X) - s^2(Y)\right)^2 + 4\left(Cov(X,Y)\right)^2} \right)$.

Exercice n°4

Deux caractères *X* (PIB par habitant, en dollars) et *Y* (taux d'analphabétisme en pourcentage) ont été mesurés dans 10 pays choisis au hasard dans une liste de pays.
On affecte à chaque pays un "poids" proportionnel à sa population.
Les résultats sont les suivants :

Pays	1	2	3	4	5	6	7	8	9	10
X	330	15 522	1 853	9 857	4 562	864	260	2 264	1 716	155
Y	90,2	0,9	7,7	0,9	0,8	65,0	57,6	5,3	8,3	66,8
Poids	0,006	0,005	0,010	0,020	0,213	0,020	0,658	0,025	0,033	0,011

On notera *D* la matrice diagonale des poids et 1 le vecteur ayant toutes ses coordonnées égales à 1 dans R^{10}.

1°/ En utilisant l'approche euclidienne de la régression, calculer la moyenne de *X* et de *Y*, la matrice des variances-covariance de *X* et *Y*, le coefficient de corrélation linéaire r_{XY}.
2°/ Calculer l'inertie totale du nuage de points par rapport à l'origine.
3°/ Calculer les axes principaux de la distribution statistique.
4°/ Calculer l'inertie du nuage de points par rapport aux axes principaux.
Interprétation.
5°/ Donner les composantes principales de chacun des 10 pays étudiés.
Représenter les dix pays dans un diagramme en composantes principales.
Conclusion ?

Solution

1°/ Approche euclidienne.

La matrice *D* est la matrice diagonale ayant, pour éléments de la diagonale, les poids de chaque pays.

$$D = \begin{pmatrix} 0{,}006 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0{,}005 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0{,}010 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0{,}020 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0{,}213 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0{,}020 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0{,}658 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0{,}025 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0{,}033 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0{,}011 \end{pmatrix}$$

Cette matrice diagonale, donc symétrique, dont tous les éléments de la diagonale sont des nombres strictement positifs, définit un produit scalaire dans l'espace R^{10} des variables.
Le produit scalaire de deux vecteurs u et v de R^{10} est $< u \mid v > = {}^t u\, D\, v$, où ${}^t u$ est la transposée du vecteur colonne u.

a) Moyennes.

La moyenne d'une variable est le produit scalaire de cette variable par le vecteur 1 :

$$\overline{X} = < X \mid 1 > = {}^t X\, D\, 1$$
$$\overline{Y} = < Y \mid 1 > = {}^t Y\, D\, 1$$

Appelons □ la matrice à 10 lignes et 2 colonnes, des données.
La première colonne est le vecteur X, la deuxième colonne est le vecteur Y.

$$□ = \begin{pmatrix} 330 & 90{,}2 \\ 15\,522 & 0{,}9 \\ 1\,853 & 7{,}7 \\ 9\,857 & 0{,}9 \\ 4\,562 & 0{,}8 \\ 864 & 65{,}0 \\ 260 & 57{,}6 \\ 2\,264 & 5{,}3 \\ 1\,716 & 8{,}3 \\ 155 & 66{,}8 \end{pmatrix}$$

La matrice des moyennes $\overline{Δ} = (\overline{X}\ \overline{Y})$ s'exprime par le produit de matrices :

$\overline{\Delta} = (\overline{X}\ \overline{Y}) = {}^t1\ D$.

$$= (1\ \ 1\ \ 1\ \ 1\ \ 1\ \ 1\ \ 1\ \ 1\ \ 1\ \ 1) \begin{pmatrix} 0{,}006 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0{,}005 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0{,}010 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0{,}020 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0{,}213 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0{,}020 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0{,}658 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0{,}025 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0{,}033 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0{,}011 \end{pmatrix}$$

$$\begin{pmatrix} 330 & 90{,}2 \\ 15\,522 & 0{,}9 \\ 1\,853 & 7{,}7 \\ 9\,857 & 0{,}9 \\ 4\,562 & 0{,}8 \\ 864 & 65{,}0 \\ 260 & 57{,}6 \\ 2\,264 & 5{,}3 \\ 1\,716 & 8{,}3 \\ 155 & 66{,}8 \end{pmatrix}$$

$$= (1\,570{,}259 \quad 41{,}1531)$$

b) Données centrées.

$$X_0 = X - 1\ \overline{X}$$
$$Y_0 = Y - 1\ \overline{Y}$$

La matrice des données centrées Z est donc donnée par la formule :

$$Z = \ - \ \overline{\Delta}\,{}^t1 = (Id - 1\ {}^t1\ D)$$

où Id est la matrice identité de R^{10}, matrice diagonale dont tous les éléments de la diagonale sont égaux à 1.
La matrice $(Id - 1\ {}^t1\ D)$ est l'**opérateur de centrage**.

$$Z = \begin{pmatrix} 330 & 90{,}2 \\ 15\,522 & 0{,}9 \\ 1\,853 & 7{,}7 \\ 9\,857 & 0{,}9 \\ 4\,562 & 0{,}8 \\ 864 & 65{,}0 \\ 260 & 57{,}6 \\ 2\,264 & 5{,}3 \\ 1\,716 & 8{,}3 \\ 155 & 66{,}8 \end{pmatrix} - \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} (1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1)$$

$$\begin{pmatrix} 0{,}006 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0{,}005 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0{,}010 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0{,}020 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0{,}213 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0{,}020 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0{,}658 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0{,}025 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0{,}033 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0{,}011 \end{pmatrix} \begin{pmatrix} 330 & 90{,}2 \\ 15\,522 & 0{,}9 \\ 1\,853 & 7{,}7 \\ 9\,857 & 0{,}9 \\ 4\,562 & 0{,}8 \\ 864 & 65{,}0 \\ 260 & 57{,}6 \\ 2\,264 & 5{,}3 \\ 1\,716 & 8{,}3 \\ 155 & 66{,}8 \end{pmatrix}$$

$$= \begin{pmatrix} 330 & 90{,}2 \\ 15\,522 & 0{,}9 \\ 1\,853 & 7{,}7 \\ 9\,857 & 0{,}9 \\ 4\,562 & 0{,}8 \\ 864 & 65{,}0 \\ 260 & 57{,}6 \\ 2\,264 & 5{,}3 \\ 1\,716 & 8{,}3 \\ 155 & 66{,}8 \end{pmatrix} - \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$$

$$(0{,}006\ \ 0{,}005\ \ 0{,}010\ \ 0{,}020\ \ 0{,}213\ \ 0{,}020\ \ 0{,}658\ \ 0{,}025\ \ 0{,}033\ \ 0{,}011) \begin{pmatrix} 330 & 90{,}2 \\ 15\,522 & 0{,}9 \\ 1\,853 & 7{,}7 \\ 9\,857 & 0{,}9 \\ 4\,562 & 0{,}8 \\ 864 & 65{,}0 \\ 260 & 57{,}6 \\ 2\,264 & 5{,}3 \\ 1\,716 & 8{,}3 \\ 155 & 66{,}8 \end{pmatrix}$$

$$= \begin{pmatrix} 0,994 & -0,005 & -0,010 & -0,020 & -0,213 & -0,020 & -0,658 & -0,025 & -0,033 & -0,011 \\ -0,006 & 0,995 & -0,010 & -0,020 & -0,213 & -0,020 & -0,658 & -0,025 & -0,033 & -0,011 \\ -0,006 & -0,005 & 0,990 & -0,020 & -0,213 & -0,020 & -0,658 & -0,025 & -0,033 & -0,011 \\ -0,006 & -0,005 & -0,010 & 0,980 & -0,213 & -0,020 & -0,658 & -0,025 & -0,033 & -0,011 \\ -0,006 & -0,005 & -0,010 & -0,020 & 0,787 & -0,020 & -0,658 & -0,025 & -0,033 & -0,011 \\ -0,006 & -0,005 & -0,010 & -0,020 & -0,213 & 0,980 & -0,658 & -0,025 & -0,033 & -0,011 \\ -0,006 & -0,005 & -0,010 & -0,020 & -0,213 & -0,020 & 0,342 & -0,025 & -0,033 & -0,011 \\ -0,006 & -0,005 & -0,010 & -0,020 & -0,213 & -0,020 & -0,658 & 0,975 & -0,033 & -0,011 \\ -0,006 & -0,005 & -0,010 & -0,020 & -0,213 & -0,020 & -0,658 & -0,025 & 0,967 & -0,011 \\ -0,006 & -0,005 & -0,010 & -0,020 & -0,213 & -0,020 & -0,658 & -0,025 & -0,033 & 0,989 \end{pmatrix} \begin{pmatrix} 330 & 90,2 \\ 15\,522 & 0,9 \\ 1\,853 & 7,7 \\ 9\,857 & 0,9 \\ 4\,562 & 0,8 \\ 864 & 65,0 \\ 260 & 57,6 \\ 2\,264 & 5,3 \\ 1\,716 & 8,3 \\ 155 & 66,8 \end{pmatrix}$$

$$= \begin{pmatrix} -1\,240,259 & 49,0469 \\ 13\,951,741 & -40,2531 \\ 282,741 & -33,4531 \\ 8\,286,741 & -40,2531 \\ 2\,991,741 & -40,3531 \\ -706,259 & 23,8469 \\ -1\,310,259 & 16,4469 \\ 693,741 & -35,8531 \\ 145,741 & -32,8531 \\ -1\,415,259 & 25,6469 \end{pmatrix}$$

c) Matrice des variances-covariances.

La matrice *C* des variances-covariances s'obtient par la formule :

$$C = {}^tZ\, D\, Z$$

On obtient :

$$C = \begin{pmatrix} 5\,437\,527,09125 & -51\,349,1934272 \\ -51\,349,1934272 & 677,32798803 \end{pmatrix} \boxed{?} \begin{pmatrix} 5\,437\,527,1 & -51\,349,2 \\ -51\,349,2 & 677,3 \end{pmatrix}$$

Dans cette matrice, on voit les variances de *X* et de *Y* et la covariance de *X* et *Y* :

$$s^2(X) = ||X_0||^2 = {}^tX_0\, D\, X_0 = 5\,437\,527,1$$
$$s^2(Y) = ||Y_0||^2 = {}^tY_0\, D\, Y_0 = 677,3$$
$$Cov(X, Y) = \langle X_0 \mid Y_0 \rangle = {}^tX_0\, D\, Y_0 = -51\,349,2$$

Les variables centrées réduites $\frac{X_0}{s(X)} = \frac{X_0}{\|X_0\|}$ et $\frac{Y_0}{s(Y)} = \frac{Y_0}{\|Y_0\|}$ sont les vecteurs unitaires de R^{10} portés par X_0 et Y_0.

d) Coefficient de corrélation linéaire.

$$r_{XY} = \frac{Cov(X, Y)}{s(X)\, s(Y)} = \frac{\langle X_0 \mid Y_0 \rangle}{\|X_0\| \|Y_0\|} = \cos(X_0, Y_0) = -\frac{51\,349,2}{\sqrt{5\,437\,527,1} \times \sqrt{677,3}} = -0,8461$$

cos (147° 47' 32" 8) = – 0,8461
Les vecteurs X_0 et Y_0 font donc un angle de 147° 47' 32" 8 dans R ^{10}.

Comme les variables centrées réduites $\frac{X_0}{s(X)} = \frac{X_0}{\|X_0\|}$ et $\frac{Y_0}{s(Y)} = \frac{Y_0}{\|Y_0\|}$, sont colinéaires aux variables centrées dans R ^{10}, leur angle est le même que celui des variables centrées et le produit scalaire des variables centrées réduites est égal au coefficient de corrélation linéaire des variables centrées :

$$r_{XY} = \left\langle \frac{X_0}{\|X_0\|} \middle| \frac{Y_0}{\|Y_0\|} \right\rangle$$

Dans *Z*, les données sont déjà centrées.
Pour les réduire, il faut diviser la première colonne par $s(X) = \| X_0 \|$ et la deuxième colonne par $s(Y) = \| Y_0 \|$.
Cette opération se fait en multipliant à droite la matrice des données centrées *Z* par la matrice diagonale

$$S^{-1} = \begin{pmatrix} \frac{1}{s(X)} & 0 \\ 0 & \frac{1}{s(Y)} \end{pmatrix}.$$

Au total, les données centrées réduites s'obtiennent à partir des données par :

$$Z_r = Z\,S^{-1} = (Id - 1\ {}^t1\ D) \square S^{-1} = \begin{pmatrix} -0,5319 & 1,8846 \\ 5,9831 & -1,5467 \\ 0,1213 & -1,2854 \\ 3,5537 & -1,5467 \\ 1,2830 & -1,5505 \\ -0,3029 & 0,9163 \\ -0,5619 & 0,6320 \\ 0,2975 & -1,3776 \\ 0,0625 & -1,2623 \\ -0,6069 & 0,9855 \end{pmatrix}$$

La matrice des corrélations est la matrice $R = {}^tZ_r\,D\,Z_r = \begin{pmatrix} 1 & r_{XY} \\ r_{XY} & 1 \end{pmatrix}$.

2°/ Inertie totale.

L'inertie totale du nuage de points dans R ² par rapport à l'origine $G = (\overline{X}, \overline{Y})$ est la somme des variances de *X* et de *Y* :

$$I_T = s^2(X) + s^2(Y) = \| X_0 \|^2 + \| Y_0 \|^2$$

On l'obtient en prenant la trace de la matrice *C* des variances-covariances :

$$I_T = 5\ 438\ 204{,}42$$

C'est aussi la somme des valeurs propres de la matrice *C*.

3°/ Axes principaux.

Les axes principaux sont définis par les vecteurs propres de la matrice *C* des variances-covariances.
Pour les calculer, il faut d'abord déterminer les valeurs propres de *C* ; ce sont les solutions λ_1 et λ_2 de l'équation :

$$Dét\ (C - \lambda\ Id) = 0$$

Leur somme est la trace de la matrice *C*, leur produit est le déterminant de la matrice *C* :

$\lambda_1 + \lambda_2 = I_T = || X_0 ||^2 + || Y_0 ||^2$
$\lambda_1 \lambda_2 = || X_0 ||^2 + || Y_0 ||^2 - (< X_0 \mid Y_0 >)^2$

Le discriminant de l'équation aux valeurs propres $Dét\ (C - \lambda\ Id) = 0$ est :

$(\lambda_1 + \lambda_2)^2 - 4\ \lambda_1 \lambda_2 = (Tr\ (C))^2 - 4\ Dét\ (C)$
$= 5\ 438\ 204{,}42^2 - 4 \times 1\ 046\ 249\ 618{,}95$
$= 2{,}956\ 988\ 230\ 69 \times 10^{13} > 0$

Sa racine carrée est 5 437 819,62802

Les valeurs propres sont :

$\lambda_1 = \frac{1}{2}(5\ 438\ 204{,}42 + 5\ 437\ 819{,}63) = 5\ 438\ 012{,}02$
$\lambda_2 = \frac{1}{2}(5\ 438\ 204{,}42 - 5\ 437\ 819{,}63) = 192{,}40$

Si λ est une valeur propre, on a (voir Cours) :

$$\begin{pmatrix} s^2(X) - \lambda & Cov(X,Y) \\ Cov(X,Y) & s^2(Y) - \lambda \end{pmatrix} \begin{pmatrix} s^2(Y) - \lambda \\ -Cov(X,Y) \end{pmatrix} = \begin{pmatrix} (s^2(X) - \lambda)(s^2(Y) - \lambda) - (Cov(X,Y))^2 \\ (s^2(Y) - \lambda)\ Cov(X,Y) - (s^2(Y) - \lambda)\ Cov(X,Y) \end{pmatrix} =$$
$$\begin{pmatrix} Det(C - \lambda\ Id) \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} = 0$$

donc le vecteur $\begin{pmatrix} s^2(Y)-\lambda \\ -Cov(X,Y) \end{pmatrix}$ est un **vecteur propre de la matrice *C* des variances-covariances pour la valeur propre** □.
Le carré de la norme de ce vecteur pour le produit scalaire canonique de R ² est donné par :

$$(s^2(Y)-\square\ -Cov(X,Y))\begin{pmatrix} s^2(Y)-\lambda \\ -Cov(X,Y) \end{pmatrix} = (s^2(Y)-\square)^2 + (Cov(X,Y))^2$$

On peut donc prendre pour vecteur normé relatif à la valeur propre □, le

vecteur $u = \frac{1}{\sqrt{(s^2(Y)-\lambda)^2+(Cov(X,Y))^2}}\begin{pmatrix} s^2(Y)-\lambda \\ -Cov(X,Y) \end{pmatrix}$, ou son opposé.

Ainsi, le premier axe principal du nuage de points est défini par le vecteur unitaire

$$u_1 = \frac{1}{\sqrt{(677{,}33-5\,438\,012{,}02)^2+(-51\,349{,}19)^2}}\begin{pmatrix} 5\,438\,012{,}02-677{,}33 \\ -51\,349{,}19 \end{pmatrix} = \begin{pmatrix} 0{,}999\,955 \\ -0{,}009\,443 \end{pmatrix}$$

$$\boxed{u_1 = \begin{pmatrix} 0{,}999\,955 \\ -0{,}009\,443 \end{pmatrix}}$$

Il est pratiquement confondu avec l'axe des *X*.
Le premier axe principal est la **droite de régression orthogonale**.
Le deuxième axe principal, orthogonal au premier, est pratiquement confondu avec l'axe des *Y* :

$$u_2 = \begin{pmatrix} 0{,}009\,443 \\ 0{,}999\,955 \end{pmatrix}$$

4°/ Inertie par rapport aux axes principaux.

L'inertie par rapport aux axes principaux est égale à la valeur propre correspondante :

□ 1 = 5 438 012,02 par rapport au premier axe principal,
□ 2 = 192,40 par rapport au deuxième axe principal.

Ces valeurs s'interprétent en termes de pourcentage de variance expliquée :

Le premier axe principal explique $\frac{5\,438\,012{,}02}{5\,438\,012{,}02 + 192{,}40}$ = 99,9965 % de la variance totale.

Le deuxième axe principal explique $\frac{192{,}40}{5\,438\,012{,}02 + 192{,}40}$ = 0,0035 % de la variance totale.

En d'autres termes, la dispersion observée des points dans le nuage est due presque uniquement aux variations de la PIB et pratiquement pas aux variations du taux d'analphabétisme.

5°/ Coordonnées factorielles.

La matrice *V* des vecteurs propres a, pour colonnes, les vecteurs propres u_1 et u_2 :

$$V = \begin{pmatrix} 0{,}999\,955 & 0{,}009\,443 \\ -0{,}009\,443 & 0{,}999\,955 \end{pmatrix}$$

S'agissant d'une matrice orthogonale, son inverse est égale à sa transposée :

$$V^{-1} = {}^tV = \begin{pmatrix} 0{,}999\,955 & -0{,}009\,443 \\ 0{,}009\,443 & 0{,}999\,955 \end{pmatrix}$$

Soit (x_i y_i), les coordonnées centrées d'un pays.
Les coordonnées factorielles du pays sont données par

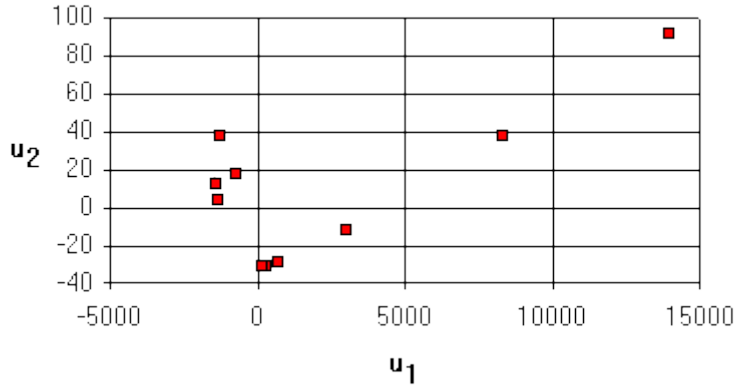
$$(x_i \quad y_i)\, V = (x_i \quad y_i) \begin{pmatrix} 0{,}999\,955 & 0{,}009\,443 \\ -0{,}009\,443 & 0{,}999\,955 \end{pmatrix}.$$

Ces formules, pour $i \in [1, 10]$ peuvent se condenser en une seule :

$$L = Z\,V$$

$$L = \begin{pmatrix} -1\,240{,}259 & 49{,}0469 \\ 13\,951{,}741 & -40{,}2531 \\ 282{,}741 & -33{,}4531 \\ 8\,286{,}741 & -40{,}2531 \\ 2\,991{,}741 & -40{,}3531 \\ -706{,}259 & 23{,}8469 \\ -1\,310{,}259 & 16{,}4469 \\ 693{,}741 & -35{,}8531 \\ 145{,}741 & -32{,}8531 \\ -1\,415{,}259 & 25{,}6469 \end{pmatrix} \begin{pmatrix} 0{,}999\,955 & 0{,}009\,443 \\ -0{,}009\,443 & 0{,}999\,955 \end{pmatrix} = \begin{pmatrix} -1\,240{,}666 & 37{,}333 \\ 13\,951{,}493 & 91{,}495 \\ 283{,}044 & -30{,}782 \\ 8\,286{,}748 & 38{,}000 \\ 2\,991{,}987 & -12{,}100 \\ -706{,}452 & 17{,}177 \\ -1\,310{,}355 & 4{,}073 \\ 694{,}048 & -29{,}300 \\ 146{,}045 & -41{,}475 \\ -1\,415{,}437 & 12{,}281 \end{pmatrix}$$

Portons les dix points sur un diagramme en coordonnées rectangulaires.



Le point (0, 0) correspond au centre de gravité *G* de la distribution.

Sur ce diagramme, on voit nettement apparaître **deux groupes de pays** :

— ceux qui ont la plus faible première composante principale (négative), ce qui correspond à la **PIB (Production Intérieure Brute) la plus faible** : ceux-là ont une deuxième composante principale entre 0 et 40, sans lien apparent avec la PIB par habitant, qui est la première composante principale.
— les autres, qui se répartissent approximativement sur une **droite** : la deuxième composante principale dépend linéairement de la PIB par habitant.

Il est toujours délicat d'interpréter l'analyse en composantes principales, mais dans le cas présent, la première composante principale est pratiquement la variable *X* : c'est la PIB par habitant qui a le plus d'influence sur la dispersion des pays dans le graphique.
Le taux d'analphabétisme joue, par conséquent, un rôle moindre.
Le coefficient de corrélation linéaire négatif $r_{XY} = -0{,}85$ montre que le taux d'analphabétisme, pourcentage d'habitants illettrés dans la population de plus de 15 ans, a tendance à diminuer quand la PIB par habitant augmente, mais le taux de détermination $r_{XY}^2 = 0{,}72$ montre que le prédicteur linéaire du taux d'analphabétisme par la PIB n'explique que 72 % de la variation du taux d'analphabétisme.
Il y a donc certainement d'autres facteurs en jeu, notamment culturels : il y a des analphabètes riches et des intellectuels pauvres.