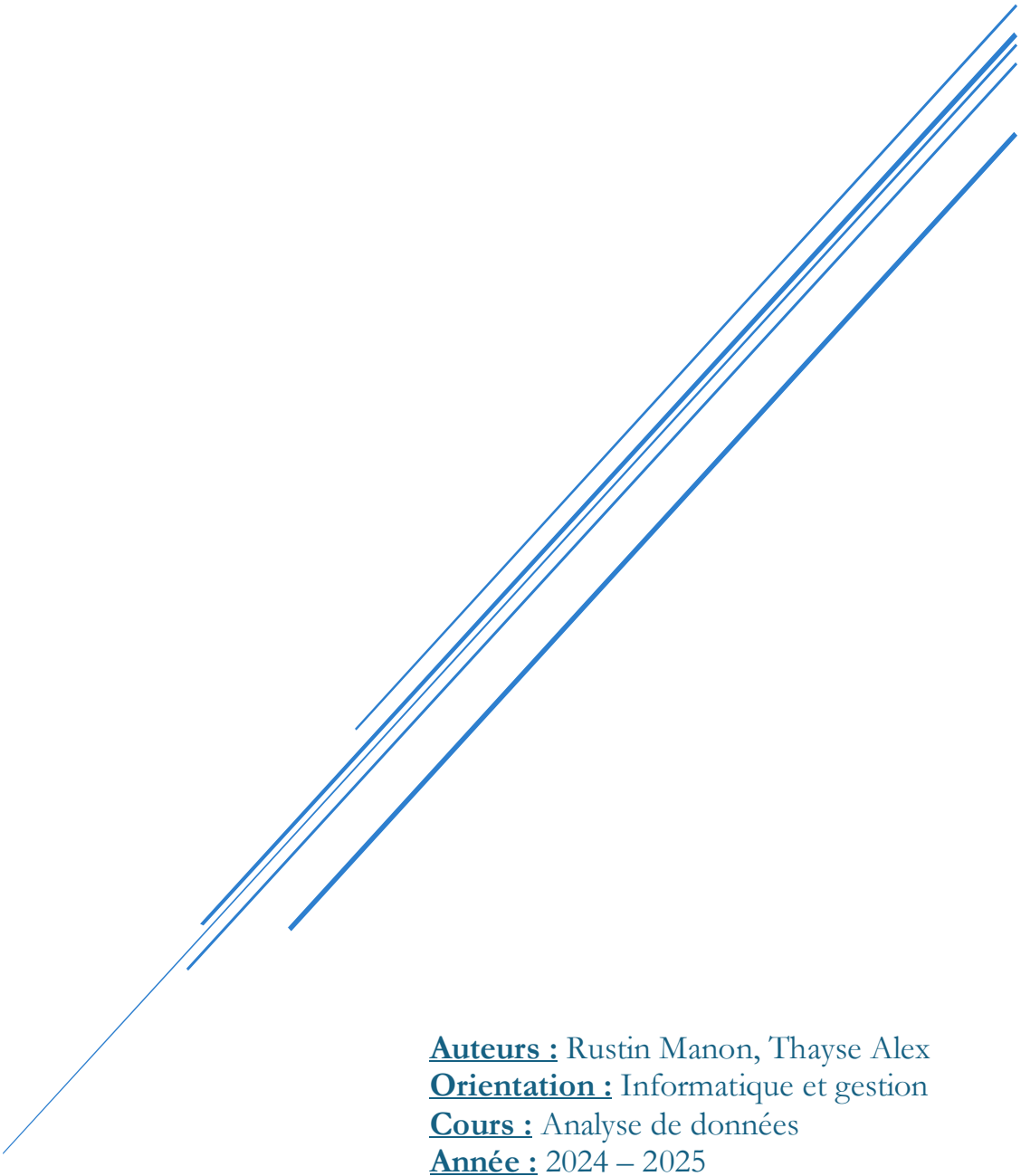
UMONS
University of Mons

POLYTECH MONS

ANALYSE DE DONNÉES

Réduction de Dimension et Analyse de la Variance



Auteurs : Rustin Manon, Thayse Alex
Orientation : Informatique et gestion
Cours : Analyse de données
Année : 2024 – 2025
Professeur : Siebert Xavier
Assistant : Baeckelant Timothy

Table des matières

Table des Figures

Table des Tableaux

Analyse des Données – Réduction de la dimension

1. Introduction

L'objectif de cette partie est de réaliser une analyse en composantes principales (ACP) sur un jeu de données préalablement sélectionné. L'ACP s'avère particulièrement intéressante lorsque les données étudiées contiennent un nombre important de variables quantitatives, où les individus étudiés sont par conséquent représentés dans un espace à haute dimension. Cette analyse permet dès lors de revenir à un espace de dimension réduite tout en conservant autant que possible la variabilité des données. Ceci permet de réduire la complexité des données multidimensionnelles, et ainsi faciliter leur visualisation tout en préservant le plus d'informations possible. Ainsi, on cherche à définir de nouvelles variables qui seront des combinaisons linéaires des variables initiales. Ces variables sont appelées « composantes principales ».

Il existe différentes manières d'appliquer l'ACP, en réduisant la dimensionnalité tout en préservant au mieux la structure des données d'origine. Par exemple, la projection de X sur un sous-espace de dimension réduite (souvent de dimension 2 pour la visualisation des données) et l'approximation de X par une matrice de rang inférieur sont toutes les deux utilisées entre autres pour la compression des données.

2. Exploration des données

Dans le cadre de ce projet, les données que nous analyserons sont constituées d'un ensemble de mesures de feuilles de plantes provenant de 40 espèces différentes. Au total, 339 observations[1] sont reprises dans cette analyse. Chaque observation appartient à une des 40 espèces de feuilles et est caractérisée par 16 attributs.

Pour obtenir un bref aperçu du jeu de données, nous nous intéressons au nombre de variables qu'il contient. Dans notre cas, il y a un total de 16 variables pour chaque observation, dont 8 sont des caractéristiques de forme et 6 sont des caractéristiques de texture. La plupart de ces variables sont numériques, comme *Eccentricity* et *Aspect Ratio*. Seule la variable *Class* est catégorielle, elle permet de distinguer les différentes espèces de plantes. Dans le cadre de l'analyse en composantes principales, seules les variables quantitatives sont requises. Notre analyse ne prendra donc pas en compte la variable *Class*. Par ailleurs, *Specimen Number* est une variable qui ne porte pas d'information continue utile pour l'analyse. En effet, il s'agit de l'identifiant d'un spécimen. Or, l'ACP cherche à réduire la dimension des données en fonction des relations entre les variables continues. Un numéro d'identification n'apporte pas de valeur significative puisqu'il n'a pas de lien avec les autres caractéristiques.

A priori, un traitement des valeurs manquantes est requis avant de procéder à l'analyse en composantes principales. Toutefois, le jeu de données analysé ici ne contient aucune valeur manquante et ne nécessite donc pas de suppression de données.

3. Prétraitement des données

3.1. Standardisation des données

L'analyse en composantes principales nécessite de travailler avec des données centrées-réduites. En effet, il est très fréquent que les variables d'un jeu de données ne soient pas mesurées sur des échelles similaires, ce qui peut engendrer des problèmes dans l'analyse ; les variables de forte variance auront plus de poids dans le calcul de la distance euclidienne que les variables de petite variance. Pour permettre une comparaison pertinente entre les variables, bien qu'elles soient exprimées sur des échelles d'unité différentes, il est possible de remédier au problème en égalisant leur moyenne à 0 (opération de centrage) et leur variance à 1 (opération de réduction).

[1] Valeur obtenue à l'aide du code.

Pour centrer les données, il faut soustraire les coordonnées du centre de gravité des données. Cette opération permet de remédier au problème ci-dessus sans changer la forme du nuage de points. En effet, seule une translation est appliquée de telle sorte que le centre de gravité coïncide avec l'origine du repère :

$$X \to X - \bar{X}$$

Dès lors, les colonnes de la matrice centrée $X - \bar{X}$ sont de moyenne nulle. Soit pour chaque colonne j, on a

$$\frac{1}{n}\sum_{i=1}^{n}(x - \bar{x})_{ij} = 0$$

La réduction des données consiste quant-à-elle à les diviser par leur écart-type pour ainsi obtenir des valeurs dont la variance vaut 1. Cette opération permet d'imposer que les variables aient la même importance. Dès lors, les données ne sont plus aplaties. Centrer-réduire les données modifie donc les distances entre les individus.

$$X \to B = M(X - \bar{X})$$

Avec

$$M = \begin{pmatrix} 1/\sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{pmatrix}$$

Pour réaliser cette étape, nous utilisons la fonction StandardScaler. Les moyennes sont dès lors nulles aux erreurs de troncature près[2] et les écarts-type unitaires.

3.2. Vérification de l'adéquation de l'ACP

Des tests comme le test de Bartlett ou la mesure de l'adéquation de Kaiser-Meyer-Olkin (KMO) peuvent être utilisés pour déterminer si l'analyse en composantes principales est appropriée pour notre jeu de données. Ils permettent de vérifier si les corrélations entre variables sont suffisamment élevées pour justifier une réduction de dimensionnalité. Dans le cadre de ce projet, nous nous permettons d'omettre cette étape par manque de temps, on suppose donc que l'ACP est appropriée pour ce jeu de données.

4. Analyse en composantes principales (ACP)

4.1. Principe de l'ACP

Lorsqu'on effectue une analyse en composantes principales, on étudie le nuage de points représentant nos données. Pour représenter des points dans un espace à n dimensions ($n > 2$) sur un plan, l'ACP suggère de réaliser des projections orthogonales. Néanmoins, le choix de ces projections orthogonales a un impact sur la quantité d'informations perdues. Il faut donc trouver des projections orthogonales qui minimisent les pertes d'informations, c'est-à-dire qui maximisent l'inertie des points. L'inertie permet de mesurer la dispersion du nuage des individus. Il s'agit donc d'une généralisation de la notion de variance au cadre multivarié pour quantifier la dispersion des données.

Rechercher ces axes d'inertie maximale revient à créer de nouvelles variables avec la plus grande variance. Dès lors, le repère initial est remplacé par un nouveau système de représentation où chaque nouvel axe est orienté de telle sorte à extraire un maximum d'inertie du nuage de points : le premier axe prend en charge le maximum d'inertie, le deuxième axe extrait le plus possible de l'inertie non prise en compte par le premier axe, et ainsi de suite. Cette transformation repose sur la diagonalisation de la matrice des variances-covariances.

[2] On précise « aux erreurs de troncature près » pour mettre en évidence le fait que les calculs informatiques utilisent une précision numérique finie, ce qui empêche de représenter tous les nombres réels de manière parfaitement exacte. Ceci permet de justifier que les moyennes ne sont pas parfaitement nulles, même si elles s'en rapprochent très fortement.

4.2. Réalisation de l'ACP

Initialement, nous effectuons l'analyse en composantes principales sous python à l'aide de la fonction $PCA()$ issue de la librairie sklearn.decomposition. Ceci nous donne dès lors un accès à d'autres valeurs intéressantes pour la suite.

4.3. Détermination des composantes principales

4.3.1. Matrice de covariance

La covariance observée entre deux variables x et y est donnée par

$$cov(x,y) = \sigma_{xy} = \sum_{i=1}^{n} p_i(x_i - \bar{x})(y_i - \bar{y})$$

Elle joue un rôle crucial lors d'une analyse en composantes principales puisqu'elle mesure le degré de corrélation entre plusieurs variables. Pour résumer les covariances existantes et identifier les relations entre les variables, on définit la matrice de covariance :

$$V = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1p} \\ \sigma_{21} & & & \\ \vdots & & \ddots & \\ \sigma_{p1} & & & \sigma_p^2 \end{bmatrix}$$

La matrice de covariance nous permet de déterminer la relation linéaire entre les variables. Il s'agit d'une matrice carrée symétrique. Le signe des éléments de la matrice nous indique la direction de la relation tandis que la magnitude exprime si les variables sont fortement ou faiblement corrélées. Un signe positif indique que les variables ont une relation linéaire positive : lorsque l'une augmente, l'autre a tendance à augmenter. Un signe négatif signifie quant-à-lui que les variables ont une relation linéaire négative : lorsque l'une augmente, l'autre diminue. Si la valeur de la covariance est proche de 0, cela signifie qu'il n'y a pas de relation linéaire significative entre les variables.

Dans le cas où les données sont centrées-réduites, la matrice de covariance n'est rien d'autre que la matrice de corrélation.

4.3.2. Heatmap

Sous python, il est possible de représenter visuellement la matrice de corrélation sous forme de heatmap. Cette représentation nous permet de mettre en évidence la corrélation entre deux variables. Les couleurs sont utilisées pour indiquer l'intensité de la corrélation. Des couleurs proches du jaune représentent une corrélation négative tandis que des couleurs proches du bleu représentent une corrélation positive. Une ACP fonctionne bien lorsqu'il y a des corrélations élevées entre certaines variables. Au contraire, lorsque les coefficients de corrélation sont assez faibles (proches de 0), les variables ne varient pas ensemble de manière linéaire. En d'autres termes, une variation dans l'une de ces variables ne permet pas de prédire de manière fiable une variation dans l'autre. Pour le jeu de données analysé, nous avons obtenu la heatmap de corrélation représentée à la Figure 1. On remarque que certaines variables sont fortement corrélées entre elles, tels que *Lobedness* et *Maximal Indentation Depth* (corrélation forte d'environ 0,95). Cela signifie que ces deux variables partagent beaucoup d'information et peuvent être regroupées dans une même composante principale, pour éviter la redondance. C'est également le cas pour *Average Intensity* avec *Average Contrast, Smoothness*, et *Entropy*, avec une corrélation allant jusqu'à 0,98. A l'inverse, certaines variables ont des corrélations plus faibles : le coefficient de corrélation entre *Solidity* et *Average Intensity* n'atteint que 0,085. Ces variables pourraient être

représentées par des composantes principales différentes dans l'ACP, car elles n'ont pas de relation forte qui les regrouperait sur une même dimension.

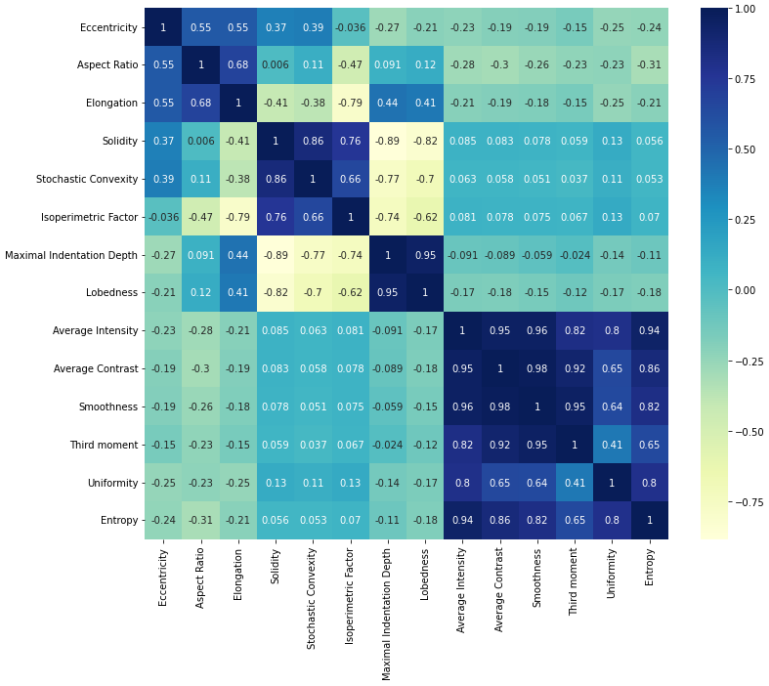


Figure 1 : Heatmap de la matrice de corrélation

4.4. Choix du nombre de composantes principales

4.4.1. Critère de choix du nombre de composantes

A partir des valeurs présentées précédemment, il est possible de déterminer le nombre de composantes principales optimal pour réduire la dimension de nos données. Pour ce faire, nous observons la proportion de variance expliquée pour chacune des composantes principales. La variance expliquée dans l'ACP indique la quantité d'information retenue après la réduction de dimensionnalité. Elle représente la proportion de la variabilité des données initiales capturée par chaque composante principale. En construisant le *scree plot*, il est possible d'observer l'évolution de la proportion de variance expliquée en fonction du nombre de composantes. A partir de ce graphique, on peut déterminer le nombre de composantes principales optimal de deux manières :

- Choisir le nombre de composantes principales de sorte à atteindre un certain pourcentage de variance expliqué, fixé arbitrairement (par exemple 85%) ;
- Repérer le « coude » dans le graphe des valeurs propres : c'est à partir de ce point que l'ajout de nouvelles composantes principales n'apporte plus énormément d'informations en termes de variance expliquée.

Dans le cadre de cette analyse, nous déterminons le nombre de composantes principales optimal sur base d'une version améliorée du *scree plot* traditionnel. En effet, nous prenons notre décision sur base d'un graphique qui présente à la fois la variance expliquée par chaque composante et la variance cumulée (Fig. 2). Ainsi, nous pouvons identifier visuellement le point à partir duquel l'ajout de nouvelles composantes n'améliore plus significativement la variance expliquée. Dans notre cas, une analyse en 2 composantes n'est pas une bonne idée (on atteint seulement 70% de la variance expliquée). Le gain d'informations en intégrant 3 axes est conséquent et nous permet d'atteindre 85% de la variance expliquée. Néanmoins, le passage de la 3ème composante à la 4ème composante n'apporte pas beaucoup d'informations supplémentaires, bien qu'il nous permettrait d'expliquer presque 90% de la variance expliquée. Nous remarquons par ailleurs que le gain d'informations à partir de la 11ème composante est totalement négligeable : près de 100% de la variance

expliquée est atteint avec seulement 10 composantes. Nous choisissons donc de prendre 3 composantes principales lors de l'ACP.

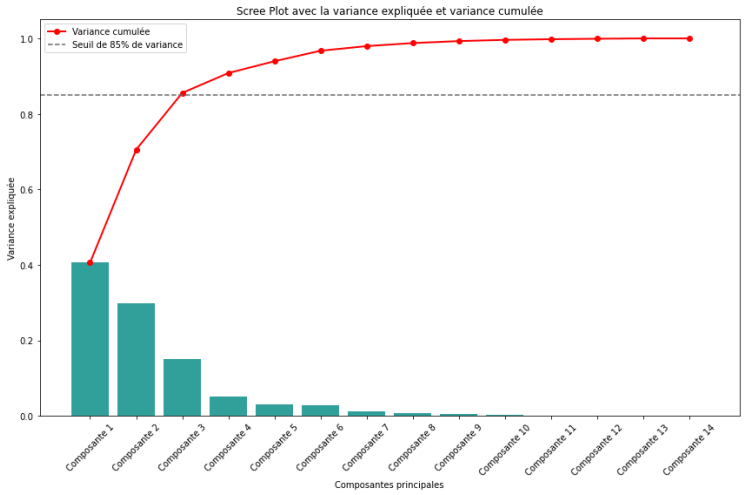


Figure 2 : Graphique de la variance expliquée par chaque composante et de la variance cumulée

5. Interprétation des résultats

Pour interpréter et visualiser les résultats de l'ACP, il est nécessaire de mettre en évidence la contribution des variables aux axes ainsi que la manière dont les individus se positionnent par rapport à ceux-ci.

5.1. Résultats relatifs aux individus

5.1.1. Score des individus

Pour visualiser la répartition des individus par rapport aux nouvelles dimensions identifiées par l'ACP, nous représentons le score des individus sur les composantes principales dans les différents plans factoriels concernés (Fig. 3). La première composante principale capture dans notre cas la plus grande variance (40,60%) et la deuxième composante principale en capture une partie supplémentaire (29,95%) orthogonale à PC1. Ensemble, elles capturent près de 70% de l'inertie totale. La troisième composante principale capture une petite partie de la variance, plus faible (15,02%), mais importante pour représenter le plus fidèlement possible les données.

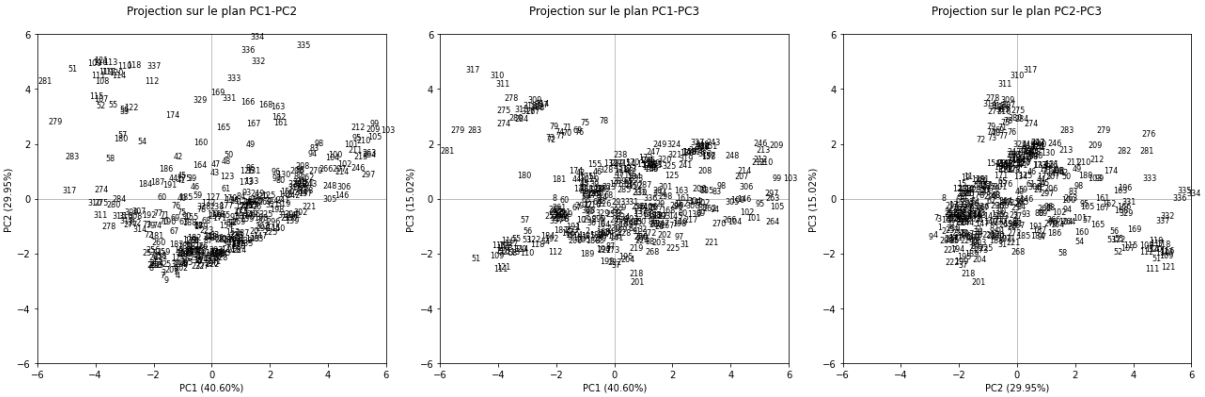


Figure 3 : Représentation du score des individus dans les premiers plans factoriels

Dans la projection sur le premier plan factoriel, nous remarquons que le nuage est assez dispersé. Une partie conséquente des individus ont un score élevé (entre 3 et 6) ou très faible (entre -3 et -6), ce qui signifie que ces individus sont fortement influencés par les variables qui contribuent le plus à cette première composante principale. De la même manière, cette observation reste valable pour la deuxième composante principale,

bien que cela soit dans une mesure plus faible mais toujours non négligeable. Par ailleurs, si nous nous intéressons à la troisième composante, nous remarquons que les scores des individus sont plus faibles que ceux relatifs aux deux premières composantes. Néanmoins, ils restent toujours non-négligeables puisque certains scores atteignent tout de même des valeurs plus élevées que 4.

Concernant les proximités entre les individus, bien que les plans considérés capturent près de 85 % de l'inertie totale, elles doivent être interprétées avec précaution. En effet, deux points proches sur le graphique peuvent correspondre à des individus qui sont en réalité éloignés. Pour bien interpréter ces proximités, il est essentiel de prendre en compte les qualités de représentation des individus.

5.1.2. Contribution des individus et qualité des représentations

La **contribution d'un individu i** à un axe α est donné par

$$CTR_\alpha(i) = \frac{(Score\ de\ i\ sur\ l'axe\ \alpha)^2}{n\lambda_\alpha} = \frac{(f_{i\alpha})^2}{n\lambda_\alpha}$$

Cet indicateur nous permet d'appréhender la manière dont l'individu est projeté sur l'axe considéré, et donc de comprendre à quel point une observation particulière contribue à une dimension principale de la variance dans un jeu de données.

La **qualité de représentation d'un individu i** par un axe k se retrouve sous la forme

$$QLT_k(i) = \frac{(Score\ de\ i\ sur\ l'axe\ k)^2}{\sum_j (Score\ de\ i\ sur\ l'axe\ j)^2}$$

Le cos² mesure la proportion de la variance expliquée par chaque composante principale dans l'ACP. Chaque individu est projeté sur les axes principaux, qui représentent les directions de la plus grande variance dans les données. Le cosinus de l'angle entre un individu et un axe principal indique la similarité entre eux, et le cos² donne la qualité de la projection de l'individu sur cet axe. Plus le cos² est élevé, mieux l'individu est représenté par cet axe, ce qui permet de quantifier l'importance d'un axe dans la représentation des individus.

ID	COORD_PC1	COORD_PC2	CTR_PC1	CTR_PC2	COS²_PC1	COS²_PC2	SUM_COS²
15	-1,2228	-2,0643	0,0008	0,0030	0,2307	0,6574	0,8881
16	-0,5984	-1,8179	0,0002	0,0023	0,0837	0,7723	0,856
17	-1,0766	-2,3604	0,0006	0,0039	0,1627	0,7819	0,9446
18	-0,8948	-1,8067	0,0004	0,0023	0,1739	0,7088	0,8827
19	-1,2995	-1,9416	0,0009	0,0026	0,2486	0,5550	0,8036
20	1,3081	-1,2743	0,0009	0,0011	0,3683	0,3495	0,7178
21	2,9082	0,5605	0,0044	0,0002	0,7683	0,0285	0,7968
22	-0,5309	-1,9934	0,0001	0,0028	0,0427	0,6019	0,6446
23	1,2814	-0,2885	0,0008	0,0001	0,3795	0,0192	0,3987
24	-0,1803	-1,4835	0,0000	0,0015	0,0093	0,6282	0,6375
25	1,1352	-0,2197	0,0007	0,0000	0,4413	0,0165	0,4578

Tableau 1 : Caractérisation des facteurs à l'aide des individus – Coordonnées, contributions et cos²

Après avoir effectué l'analyse et les différents calculs requis, nous obtenons le Tableau 1 qui reprend le score d'une partie des individus[3] ainsi que leur contribution et la qualité de leur représentation pour les deux premières composantes principales. Une dernière colonne présente la somme des qualités de représentation des individus par l'axe 1 et l'axe 2. Elle permet de donner une qualité de représentation sur les deux premiers facteurs. Idéalement, la qualité de représentation sur les axes principaux sélectionnés doit se rapprocher de 1.

[3] Nous n'avons pris qu'une partie des individus pour alléger le rapport. Il ne s'agit pas des premiers individus car les suivants (à partir du 15éme) nous semblaient plus intéressants à analyser.

Dans notre cas, la plupart des observations reprises dans le Tableau 1 sont globalement bien représentées par les deux premières composantes principales puisque la qualité de représentation sur ces axes est assez proche de 1. Par exemple, l'observation 17 a une contribution forte pour la deuxième composante principale ($\cos^2 = 0{,}7819$) et une contribution plus faible pour la première composante principale ($\cos^2 = 0{,}1627$) : la somme des qualités de représentation relative aux axes concernés est très élevée ($\cos^2 = 0{,}9446$), ce qui suggère que l'observation est bien capturée dans l'espace réduit. Si on considère l'observation 20, la somme des $\cos^2$ (0,7178) est un peu plus faible que d'autres observations, ce qui pourrait suggérer que cette observation est moins bien représentée globalement par les deux premières composantes.

Certaines observations, comme l'observation 23, semblent néanmoins assez mal représentées. En effet, la somme des $\cos^2$ n'atteint pas les 50%, ce qui suggère que la qualité de représentation de ces observations dans l'espace des deux premières composantes est modérée : bien que la première composante principale explique une part significative de la variance, il reste encore une part importante de variance non expliquée par ces deux axes. Ceci rejoint notre analyse, un peu plus haut, du nombre de composantes principales optimal pour le jeu de données considéré. Nous avions en effet utilisé 3 composantes principales dans l'analyse initiale, il est donc probable qu'une partie de la variance restante est expliquée par la troisième composante principale, qui pourrait potentiellement mieux représenter cette observation.

5.2. Résultats relatifs aux variables

5.2.1. Cercle de corrélation

Le cercle de corrélation, ou graphique de chargement, est un outil utilisé dans l'ACP afin de visualiser la relation entre les variables d'origine et les composantes principales conservées. Les axes présents correspondent aux composantes principales, et chaque variable est représentée par un vecteur. Il permet d'obtenir une vision immédiate de la caractérisation des facteurs à l'aide des variables. Pour l'interprétation du cercle de corrélation, la longueur d'un vecteur, caractéristique de la projection d'une variable sur un axe, indique la contribution de cette variable à la composante principale relative à cet axe. Plus la flèche est proche du cercle, meilleure est la qualité de la représentation de la variable. L'angle formé par deux vecteurs donne quant-à-lui une indication sur la corrélation entre les différentes variables.

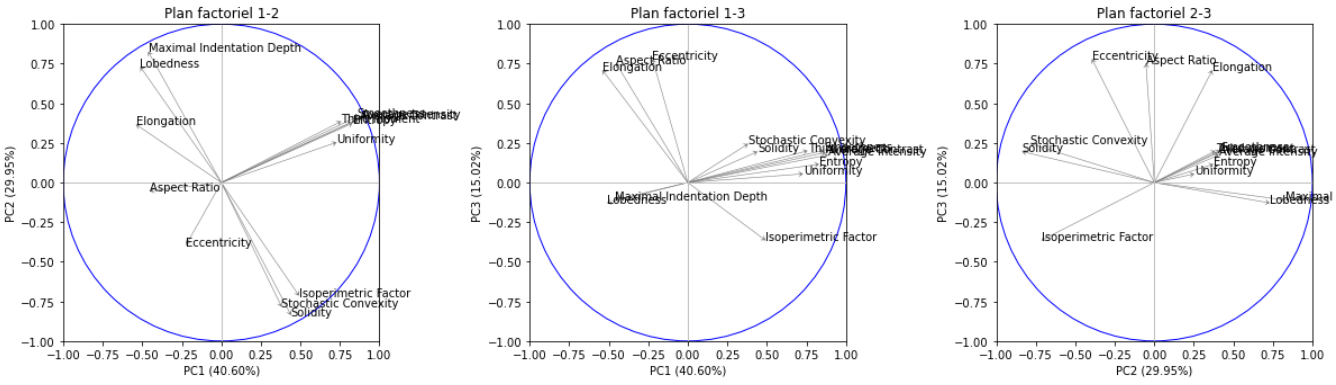


Figure 4 : Cercles de corrélation dans différents plans factoriels

Dans notre cas, les cercles de corrélation pour notre jeu de données, dans les plans factoriels relatifs aux 3 composantes principales conservées, sont repris à la Figure 4. Cette visualisation nous permet de mieux comprendre la contribution des variables aux différentes composantes principales. Par exemple, la variable *Eccentricity* a des contributions assez faibles pour les deux premières composantes (faible longueur du vecteur) mais une contribution forte pour la troisième composante principale, ce qui justifie le fait de considérer au moins 3 composantes principales. C'est également le cas pour les variables *Aspect Ratio* et *Elongation*, dont l'importance de la contribution pour la troisième composante principale est largement visible sur les cercles de corrélation comprenant PC3. On remarque néanmoins sur ces derniers que la troisième composante principale n'est utile que pour certaines variables (*Eccentricity*, *Aspect Ratio*, *Elongation* et dans une autre mesure *Isoperimetric Factor*). Pour les autres variables, leur contribution pour la troisième composante principale est assez faible. En réalité, pour la majorité des variables, leur contribution aux deux

premières composantes principales est assez forte, ce qui justifie les valeurs des variances capturées par ces deux composantes. Par ailleurs, cela confirme que ces variables, qui contribuent à ces deux composantes, sont cruciales pour une analyse efficace de notre jeu de données

A titre indicatif, l'angle sur ces cercles de corrélation donne quant à lui une indication sur la corrélation entre les différentes variables. Un angle assez petit induit des projections similaires – les deux variables ont dès lors des contributions similaires aux composantes principales, elles sont positivement corrélées. Par exemple, les vecteurs représentant les variables *Eccentricity* et *Aspect Ratio* pointent dans des directions similaires sur le cercle de corrélation de la troisième composante principale, ce qui suggère que ces variables sont fortement corrélées entre elles dans l'espace de la troisième composante. A l'inverse, si l'angle est assez grand (de l'ordre de 180° par exemple), les projections sont opposées, indiquant des contributions opposées – les variables sont négativement corrélées (comme *Maximal Indentation Depth* et *Isoperimetric*). Le cas intermédiaire, soit quand l'angle s'approche des 90°, induit qu'il est peu probable que les variables soient corrélées (comme *Lobedness* et *Uniformity*).

6. Limites de l'analyse

Certaines valeurs pourraient biaiser l'analyse en composantes principales si elles sont extrêmes ou aberrantes, car ces valeurs influencent fortement les moyennes et les covariances qui sont à la base des composantes principales. Il est dès lors possible de supprimer ces valeurs aberrantes afin d'obtenir un meilleur classifieur – la variance expliquée par les 3 premières composantes principales pourrait augmenter si on supprime les valeurs aberrantes. La visualisation 3D des données sur les 3 premières composantes principales (Fig. 5) nous permet de mieux identifier ces valeurs aberrantes. Dans notre cas, un certain nombre de valeurs sont assez éloignées du nuage principal mais nous considérons qu'il ne s'agit pas de valeurs aberrantes du fait de leur nombre. En effet, il n'y a pas que quelques valeurs extrêmes isolées que l'on pourrait retirer. Dans notre cas, il faudrait retirer un nombre conséquent de valeurs (plus d'une vingtaine de valeurs), ce qui réduirait la représentativité de l'ensemble des données.

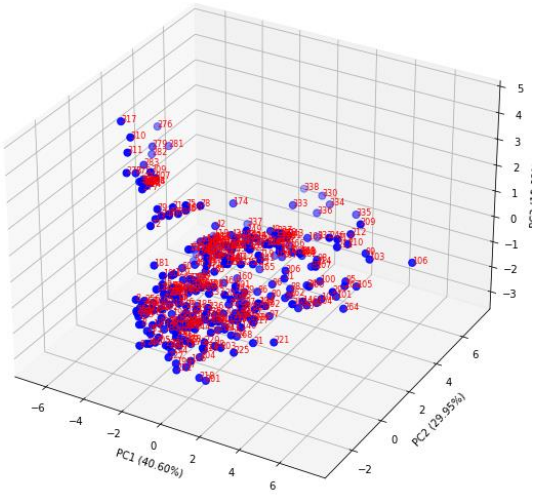


Figure 5 : Projection des individus sur les 3 premières composantes principales

Analyse des Données – Analyse de la Variance

1. Motivation

L'objectif de l'Analyse de la variance (ANOVA) est de déterminer si les moyennes (ou espérances mathématiques) de différents groupes proviennent d'une même population, autrement dit, si les différences observées entre les groupes sont significatives ou simplement dues au hasard. Les données nécessaires pour une ANOVA incluent une variable quantitative, mesurable numériquement, et une ou plusieurs variables qualitatives, chacune pouvant comporter plusieurs modalités. Par exemple, la consommation de carburant est une variable quantitative, tandis que le type de carburant représente une variable qualitative divisant les données en différents groupes. Dans ce contexte, l'ANOVA permet de vérifier si le type de carburant influence de manière significative la quantité de carburant consommée.

2. Structure des données

Le jeu de données, avec lequel ce travail est réalisé, est nommé « Diet » et est disposé sous forme de tableau de taille (78,6). Il contient des informations sur les régimes alimentaires suivis par différentes personnes et leur impact potentiel sur la perte de poids. L'objectif de cette ANOVA est donc de déterminer si le régime alimentaire suivi par une personne a un effet significatif sur sa perte de poids, et si le genre de la personne influence également ce résultat.

Les données sont structurées autour de 4 variables principales :

- *gender* : le genre de la personne (variable qualitative avec deux modalités : homme et femme) ;
- *diet* : le type de régime suivi (variable qualitative avec trois modalités : régime 1, régime 2 et régime 3) ;
- *preweight* : le poids initial (variable quantitative continue mesurée en unités de poids) ;
- *weight6weeks* : le poids mesuré après 6 semaines de régime (variable quantitative continue mesurée en unités de poids).

Pour compléter l'analyse, il est nécessaire d'ajouter une nouvelle variable permettant de calculer la perte de poids (*perte_poids*), en faisant la différence entre *weight6weeks* et *preweight*. Cette nouvelle variable constitue la variable quantitative principale dans cette ANOVA.

Dans cette ANOVA, la variable *diet* est le facteur principal de l'analyse pour tester son effet sur la perte de poids, tandis que *gender* est une variable de contrôle supplémentaire qui permet de prendre en compte son influence dans l'analyse. On parle de variable de contrôle dans le cas où la variable pourrait avoir un impact sur le résultat de l'analyse mais n'est pas la variable principale de la recherche. Cela permet d'éviter que les variations de cette variable n'interfèrent dans notre démarche, et donc de s'assurer ici que les résultats reflètent essentiellement l'effet du type de régime suivi.

Pour visualiser ces données, nous avons formé 6 groupes classés en croisant les 3 types de régime suivi et les 2 genres. Ces groupes sont représentés sous forme de diagrammes en boîte à moustaches permettant de visualiser la distribution de la perte de poids pour chaque combinaison. Un diagramme en boîte se compose de :

- **Une boîte** : elle va du premier quartile (Q1) au troisième quartile (Q3) et contient une ligne marquant la médiane des données.
- **Des moustaches** : Lignes sortant de la boîte qui représentent l'étendue des valeurs des données à l'extérieur des quartiles, exceptés les valeurs extrêmes représentées par des points isolés.

Ces diagrammes permettent de voir le centre et la distribution des données, et ainsi d'identifier les valeurs aberrantes. Dans notre cas, ils permettent d'illustrer la variation de la perte de poids entre les 6 groupes, mettant en évidence l'impact potentiel des régimes alimentaires et du genre sur les résultats. Les diagrammes obtenus (Figure 6) nous permettent de déduire que la distribution de la perte de poids varie en fonction des régimes alimentaires et du genre. Par exemple, dans le régime C, la médiane de la perte de poids semble plus faible chez les hommes que chez les femmes, ce qui suggère que les hommes perdent moins de poids en moyenne avec ce régime. De plus, l'intervalle interquartile des hommes est plus large que celui des femmes, ce qui pourrait indiquer une plus grande dispersion des données chez les hommes. En revanche, les régimes A et B montrent des médianes relativement proches, mais les intervalles interquartiles des hommes semblent toujours plus larges que ceux des femmes, ce qui indique que les hommes présentent une plus grande variabilité dans la perte de poids sous ces régimes. Ces observations visuelles peuvent être explorées davantage à l'aide de tests statistiques, notamment l'ANOVA, pour évaluer si les différences entre les groupes sont significatives.

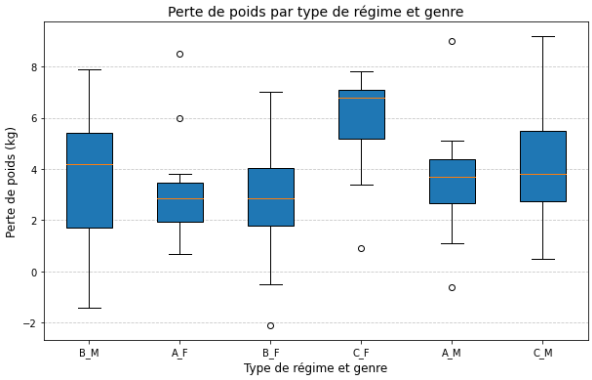


Figure 6 : Diagrammes en boîte

3. Hypothèses du modèle statistique

Un modèle statistique possible pour l'analyse de la variance est :

$$X_{ij} = \mu_i + \epsilon_{ij}$$

- μ_i : perte de poids moyenne des personnes suivant le régime i ;
- ϵ_{ij} : variabilité de la perte de poids de la personne j par rapport à μ_i ;

L'utilisation de ce modèle nécessite de supposer que les échantillons suivent des lois normales :

$$X_{ij} \overset{\mathcal{L}}{=} \mathcal{N}(\mu_i, \sigma^2) \qquad \text{indépendants}$$

D'une autre manière, les hypothèses du modèle sont l'homoscédasticité et la normalité. L'homoscédasticité suppose que la variance est la même dans chaque groupe, tandis que la normalité suppose que les x_{ij} et donc les ϵ_{ij} sont des variables gaussiennes (normales) indépendantes. Il est donc nécessaire d'effectuer des tests statistiques sur les données pour vérifier ces hypothèses.

3.1. Normalité

Pour vérifier la distribution des données, nous utilisons plusieurs représentations visuelles (Histogrammes et QQplot) et des tests statistiques.

Histogramme

La visualisation des différents groupes sous forme d'histogrammes permet d'avoir une vue générale sur la forme de la distribution. Il s'agit d'une première inspection visuelle des données, bien que subjective.

Sous python, nous utilisons la commande plt.hist(...) de pyplot. Deux représentations visuelles[4] sont présentes sur les Figures 6 et 7. Concernant le groupe des femmes qui suivent le régime de type B (Figure 8), la distribution semble approximativement normale : on retrouve en effet une seule valeur centrale (perte de poids de 2 kgs à une fréquence de 3) avec une symétrie autour, certes très approximative. Par ailleurs, des valeurs négatives sont présentes sur l'histogramme, ce qui pourrait suggérer qu'une personne a pris du poids plutôt que d'en perdre. Ces valeurs pourraient être des données aberrantes qu'il serait pertinent d'examiner plus en détail. Cependant, il est important de souligner que les valeurs négatives ne sont pas nécessairement aberrantes, cela dépend du contexte des données. Concernant le groupe des femmes qui suivent le régime de type C (Figure 7), les données ne semblent pas être distribuées normalement : Les données n'apparaissent pas symétriquement réparties autour d'une valeur centrale. Au contraire, les valeurs augmentent de manière plus progressive, ce qui suggère une forme de distribution différente de la normale Néanmoins, ces observations restent très subjectives et ne peuvent pas, à elles-seules, vérifier la distribution des données. Obligatoirement, nous devons passer par des tests statistiques qui viendront infirmer ou confirmer nos observations. Néanmoins, cette visualisation nous permet d'avoir un premier regard sur la distribution des données.

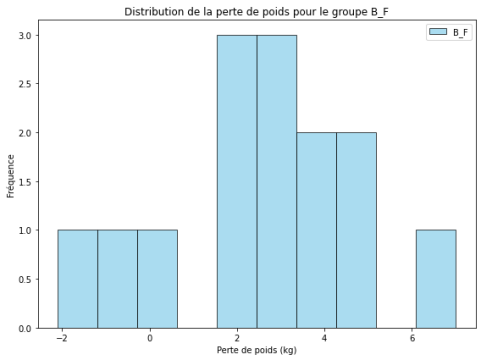


Figure 8 : Histogramme pour le groupe B_F

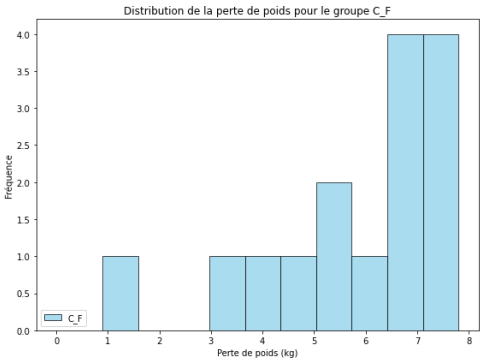


Figure 7 : Histogramme pour le groupe C_F

QQ-plot

Le QQplot est un outil graphique permettant d'évaluer si une série de données suit une distribution spécifique. Dans notre cas, nous cherchons à déterminer si notre jeu de données suit une distribution normale. Le QQplot compare les valeurs des quantiles de la loi empirique à ceux de la loi normale centrée réduite. Les quantiles empiriques sont représentés sur l'axe des ordonnées alors que les quantiles théoriques (distribution normale) sont représentés sur l'axe des abscisses. On peut conclure que la distribution empirique suit une loi de distribution normale si les points sont alignés sur la première bissectrice. Cela permet d'identifier visuellement si notre distribution semble normale.

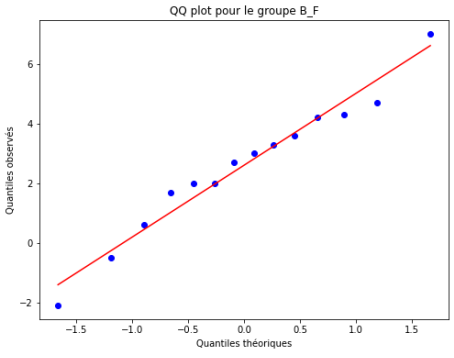


Figure 9 : QQplot pour le groupe B_F

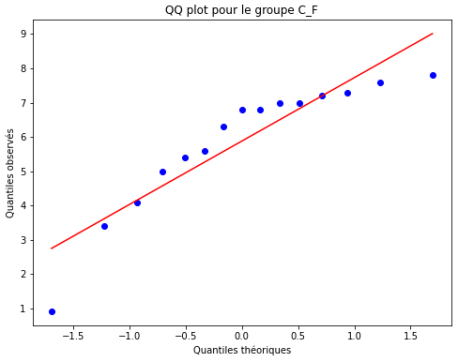


Figure 10 : QQplot pour le groupe C_F

[4] Nous analysons dans le rapport uniquement deux représentations sur les six, qui nous semblent intéressantes, pour éviter d'alourdir davantage le rapport.

Les QQplot pour les groupes B_F et C_F sont présentés respectivement sur les Figures 9 et 10. La représentation visuelle pour le groupe des femmes qui suivent le régime de type C semble indiquer que les données sont distribuées normalement : les points sont en effet alignés sur la première bissectrice – la distribution empirique semble suivre une distribution gaussienne. Cette approximation est d'autant plus valable si l'on néglige les valeurs extrêmes. Les points qui se retrouvent aux extrémités de la droite ont très peu de chances d'être sur la droite elle-même car la quantité de données est assez faible à cet endroit. Concernant le QQplot du groupe des femmes qui suivent le régime de type C, les points semblent former une courbe par rapport à la ligne droite, ce qui suggère que la distribution n'est pas gaussienne. Ces observations confirment celles relatives aux histogrammes. Cependant, cette analyse reste subjective du fait qu'elle soit visuelle.

Test de « Shapiro-Wilks »

Le test de Shapiro-Wilk est un test statistique appliqué à chaque groupe permettant de confirmer l'hypothèse de normalité. Ce test compare les valeurs des données avec des valeurs suivant une loi gaussienne. Plus la statistique est proche de 1, plus les données sont proches d'une distribution normale. L'hypothèse nulle stipule que chacun des groupes suit une distribution normale et l'hypothèse alternative exprime le contraire, soit que chacun des groupes ne suit pas une distribution normale. L'interprétation repose sur l'analyse de la p-value qui est la probabilité de commettre une erreur de première espèce, c'est-à-dire de rejeter à tort l'hypothèse nulle alors qu'elle est vraie. Pour un risque d'erreur de première espèce de 0,05, si la p-value est supérieure à 0,05, on ne rejette pas l'hypothèse nulle et on peut conclure que les données suivent une distribution normale. En revanche, si la p-value est inférieur à 0,05, l'hypothèse nulle est rejetée et on conclut que les données ne suivent pas une distribution normale.

En appliquant le test de Shapiro-Wilk, on remarque que les p-values des groupes B_M, B_F, A_M, C_M sont supérieures à 0,05. Nous pouvons conclure que les données de ces groupes suivent une loi normale. De plus, les statistiques de Shapiro-Wilk pour ces groupes sont proches de 1. Le résultat du test pour les groupes A_F et C_F fournit une p-value inférieure à 0,05. Nous devrions conclure que leurs distributions ne suivent pas une loi normale. En complément, les QQ-plots de ces groupes montrent certains écarts par rapport à la droite de normalité (les points semblent former une courbe par rapport à la droite), nous laissant supposer que ces distributions s'écartent de la loi normale.

Néanmoins, il est important de noter que la taille réduite des échantillons peut influencer les résultats de ce test. Par conséquent, bien que certains tests indiquent des déviations de la normalité pour certains groupes, nous pouvons raisonnablement émettre l'hypothèse que toutes les distributions sont approximativement normales.

3.2. Homoscédasticité :

L'hypothèse nulle relative à cette condition est donnée par

$$H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_r^2$$

avec r le nombre de groupe.

L'hypothèse nulle de ce test permet de vérifier l'égalité des variances entre les groupes.

H_1 : Les variances ne sont pas toutes égales.

Graphique de dispersion des résidus

Le graphique de dispersion des résidus est un outil couramment utilisé pour vérifier l'hypothèse d'homoscédasticité dans les modèles de régression, c'est-à-dire l'hypothèse selon laquelle la variance des résidus est constante à travers toutes les valeurs des variables explicatives. Dans ce graphique, les résidus sont tracés en fonction des différents groupes.

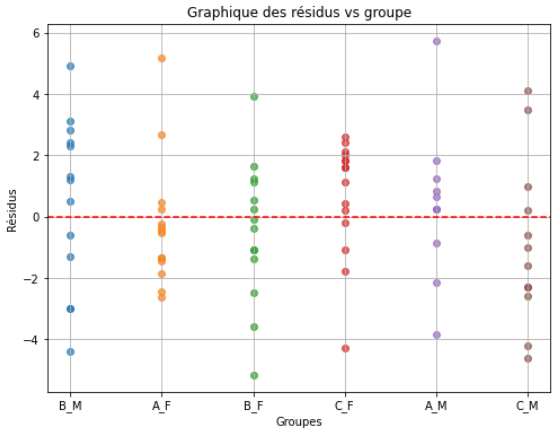


Figure 11 : Résidus pour chaque groupe

Dans notre cas, les résidus semblent se répartir de manière assez uniforme pour l'ensemble des groupes (Figure 11), ce qui suggère que la variance des résidus pourrait être constante, indiquant une homoscédasticité. Cependant, bien que cette observation visuelle soit utile, un test statistique spécifique est nécessaire pour confirmer rigoureusement cette hypothèse.

Test de Levene

Le test de Levene est le test le plus courant et polyvalent pour tester l'homoscédasticité. Il est notamment recommandé lorsqu'il y a plusieurs groupes étudiés. Il est par ailleurs robuste aux déviations par rapport à la normalité. Comme le test de Shapiro-Wilk sur A_F et C_F montre que leurs distributions ne suivent pas une loi normale, nous utilisons dans un premier temps ce test pour tester l'homoscédasticité. L'interprétation des résultats repose sur la p-value. Pour un risque d'erreur de première espèce de 0,05, si la p-value est supérieur à 0,05, l'hypothèse nulle ne peut être rejetée et on peut conclure que les variances sont les mêmes dans tous les groupes. Si la p-value est inférieur à 0,05, on rejette l'hypothèse nulle et on conclut que les variances des différents groupes ne sont pas toutes égales.

$$\text{Statistique utilisée : } \frac{SC_A}{SC_{res}}$$

Dans notre cas, comme la p-value (0,6765) est bien supérieure à 0,05, cela suggère que les variances des groupes ne sont pas significativement différentes, ce qui nous permet de conclure que les variances des groupes sont homogènes (homoscédasticité).

Test de Bartlett

Comme le test de Bartlett est particulièrement puissant lorsque les données suivent une distribution normale, nous l'utilisons ici pour perfectionner l'analyse. Néanmoins, le test de Bartlett est très sensible à la non-normalité des données. Si les données ne suivent pas une distribution normale, le test peut être faussement significatif. C'est la raison pour laquelle nous avons utilisé le test de Levene précédemment.

En appliquant le test de Bartlett, on constate que la p-value est supérieure à 0,05. Cela implique que les variances des différents groupes sont égales. Ce test nous amène aux mêmes conclusions que celles du test de Levene.

4. ANOVA à un facteur

4.1. ANOVA globale

L'objectif de l'ANOVA est de tester l'égalité des moyennes de plus de deux échantillons. L'analyse à un facteur signifie que l'on s'intéresse uniquement à une seule variable influente. Comme les hypothèses du modèle (indépendance des observations, normalité des résidus et homoscédasticité) sont respectées, l'ANOVA peut être appliquée.

Le test d'hypothèse se formule comme suit :

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_r$$

avec r le nombre de groupes.

$$H_1 : \exists i, j : \mu_i \neq \mu_j$$

Il est important de noter que l'hypothèse alternative est souvent formulée en termes de « différence », mais l'ANOVA ne précise pas où ces différences se produisent. Elle précise seulement que ces différences existent.

L'ANOVA se base sur la construction d'un intervalle de confiance suivant une distribution de Fisher-Snedecor (F). La statistique F est calculée en faisant le rapport entre la variabilité inter-groupes et la variabilité intra-groupes. La variabilité inter-groupes est mesurée par la moyenne des carrés entre les groupes MC_A, qui calcule la variation des moyennes de chaque groupe par rapport à la moyenne globale tandis que la variabilité intra-groupes est mesurée par la moyenne des carrés à l'intérieur des groupes MC_{Res}, qui représente la variation au sein de chaque groupe par rapport à sa propre moyenne.

La statistique étudiée F est donc le rapport entre ces deux variabilités :

$$F = \frac{MC_A}{MC_{Res}}$$

Une variabilité inter-groupes élevée par rapport à une variabilité intra-groupes suggère que les moyennes des groupes sont significativement différentes. En effet, cela signifie qu'il existe une grande différence entre les groupes par rapport à une faible variation au sein des groupes. Dans ce cas, on peut affirmer que la variabilité inter-groupes n'est pas due à une variabilité intra-groupes mais à une influence du facteur étudié.

L'ANOVA à un facteur montre qu'il y a une différence significative entre les moyennes des groupes car la p-value est inférieure à 0,05 (p-value obtenue : 0,0078). Cela indique qu'au moins un des groupes est différent des autres en termes de perte de poids moyenne. Cependant, l'ANOVA ne permet pas de savoir quel groupe diffère des autres. Pour identifier précisément les groupes responsables de cette différence, nous devons effectuer un test post-hoc.

4.2. Tests post-hoc

Les tests post-hoc ont pour objectif de déterminer quelles moyennes diffèrent significativement les unes des autres. Ils sont effectués uniquement si l'ANOVA a révélé une différence globale significative entre les groupes (rejet de l'hypothèse nulle). Cela s'applique dans notre cas.

Le test choisi ici est celui de Tukey, également appelé test de la « différence franchement significative ». Ce test est largement utilisé car il offre un bon compromis entre contrôle du taux d'erreur de type I (probabilité d'identifier une différence inexistante) et puissance statistique (capacité à détecter des différences réelles). Le principe consiste à ranger les moyennes par ordre croissant et d'ensuite procéder aux comparaisons multiples. La procédure commence par la comparaison de la moyenne la plus élevée avec la plus faible, puis avec les autres moyennes dans l'ordre croissant. Ce processus s'arrête dès qu'une différence n'est plus significative. Ce test réduit l'accumulation d'erreurs liées aux comparaisons multiples en limitant les comparaisons à celles qui sont nécessaires.

Dans notre cas, une partie des résultats est présente dans le Tableau 2. La différence entre les moyennes des groupes A_F et A_M est très faible, et la p-value (0,9899) est largement supérieure à 0.05, ce qui nous permet de conclure qu'il n'y a pas de différence significative entre ces groupes. Les conclusions sont similaires pour les moyennes des groupes A_F et B_F, A_F et B_M. En revanche, la quatrième comparaison (A_F et C_F) révèle une différence significative (p-value largement inférieure à 0,05), ce qui suggère que le groupe C_F se distingue des autres en termes de moyenne. En effectuant cette analyse pour l'ensemble des tests post-hoc,

on remarque qu'il existe une différence significative entre le groupe C_F et les groupes A_F et B_F tandis que les autres ne présentent pas de différence significative entre eux. Nous interprétons ces résultats de la manière suivante : Les femmes qui suivent le régime C perdent plus de poids que celles qui suivent les régimes A et B. Ces observations peuvent être confirmées en analysant en parallèle les diagrammes en boîte à moustache correspondant.

Groupe 1	Groupe 2	Différence des moyennes	p-value	Intervalle de confiance (95%)		Différence significative
A_F	A_M	0,6	0 ,9899	-2,2748	3,4748	False
A_F	B_F	-0,4429	0,9962	-3,0671	2,1814	False
A_F	B_M	0,4269	0,9971	-2,2474	3,1012	False
A_F	C_F	2 ,83	0,0233	0,2498	5,4102	True

Tableau 2 : Résultats des test Post-Hoc

5. ANOVA à deux facteurs

L'ANOVA à deux facteurs permet de déterminer si chaque facteur exerce une influence significative sur la variable étudiée (effets principaux) et s'il existe une interaction entre ces facteurs. L'application de cette méthode permet d'évaluer les effets des types de régime et du sexe sur la perte de poids, ainsi que l'interaction entre ces deux facteurs. Cette interaction signifie que l'effet d'un facteur (par exemple, le type de régime) sur la perte de poids peut dépendre de la catégorie de l'autre facteur (par exemple, le sexe). Le test fournit les p-values correspondant aux effets des facteurs pris isolément (effets principaux) ainsi qu'à leur interaction. Ces p-values permettent de déterminer si ces effets sont statistiquement significatifs.

	Somme des carrés	Degrés de liberté	Statistique F	p-value associée
C(Diet)	71,0070	2	6,3132	0,0030
C(Gender)	0,1355	1	0,0241	0,8771
C(Diet) :C(Gender)	25,1407	2	2,2353	0,1143

Tableau 3 : Résultats issus de l'ANOVA à deux facteurs

Les p-values issues de ce test (Tableau 3) nous permettent d'observer que l'interaction entre les types de régime et le genre n'est pas significative (p-value = 0,1143 supérieure à 0,05). On remarque néanmoins que le type de régime a un effet significatif sur la perte de poids (p-value = 0,002977), ce qui n'est pas le cas du genre (p-value = 0,877067). Cela signifie que la variation de la perte de poids est principalement expliquée par le type de régime, et non par le genre, et qu'il n'y a pas d'effet combiné significatif entre les deux facteurs.

La somme des carrés pour l'interaction (C(diet) :C(gender)) attire néanmoins notre attention puisqu'elle nous parait relativement importante (25,1407). Néanmoins, cela ne suffit pas pour déterminer si l'interaction est statistiquement significative. La p-value élevée de 0,114336 suggère que, bien qu'il y ait une certaine variation expliquée par l'interaction entre le régime et le genre, cet effet n'est pas assez fort pour être jugé significatif dans le cadre de cette analyse.

Le graphique d'interaction (Figure 12) semble indiquer que les hommes réagissent de manière plus homogène aux différents régimes tandis que les femmes montrent une forte progression de leur perte de poids lorsqu'elles passent au régime C, ce qui pourrait expliquer une partie de l'interaction observée dans l'ANOVA. L'effet du régime sur la perte de poids semble donc dépendre clairement du genre : Pour les hommes, la perte de poids reste relativement stable à travers les régimes, avec une légère augmentation du régime A au régime C (lignes relativement proches mais légèrement inclinées, cela peut suggérer une interaction faible). Pour les femmes, lorsque l'on passe des régimes B et A au régime C, on observe une forte augmentation (lignes très inclinées). En outre, les hommes semblent perdre davantage de poids que les femmes dans les régimes A et B. Cependant, pour le régime C, les femmes ont une perte de poids supérieure à celle des hommes.

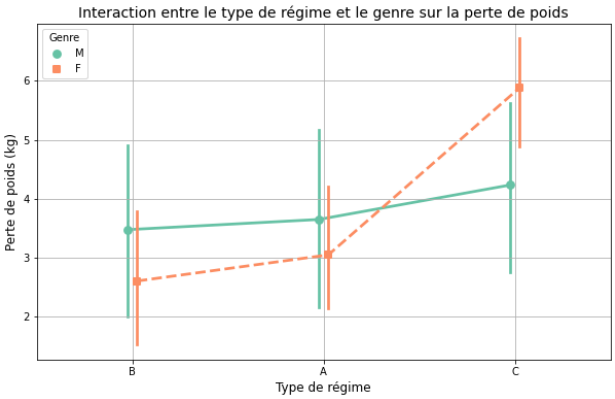


Figure 12 : Graphique d'interaction entre le type de régime et le genre sur la perte de poids

Lien entre les résultats de l'ANOVA à deux facteurs et le graphique d'interaction

Concernant l'effet du type de régime, les résultats de l'ANOVA confirment les observations qui ressortent du graphique de l'interaction : Le type de régime a un effet significatif (p-value = 0,002977), ce qui est cohérent avec le graphique, où les régimes A et C entraînent une perte de poids plus importante.

Pour l'effet du genre, les résultats semblent se contredire. En effet, les résultats de l'ANOVA indiquent que le genre n'a pas d'effet significatif (p-value = 0,877067), ce qui signifierait que les hommes et les femmes ne diffèrent pas de manière significative en termes de perte de poids. Or, sur le graphique, il semble que le genre a un effet dans certains cas (notamment avec le régime C, où les femmes montrent une perte de poids marquée).

L'interaction entre régime et genre n'est pas significative si l'on considère les résultats de l'analyse (p-value = 0,114336), or le graphique suggère que l'effet du régime diffère entre les hommes et les femmes (notamment pour le régime C) puisque les lignes se croisent.

Pour approfondir l'analyse et comprendre les incohérences, nous avons défini à l'avance des contrastes (comparaisons à priori) qui nous ont permis de tester des hypothèses spécifiques.

6. ANOVA à contrastes

	Statistique F pour le contraste	**Valeur critique de la statistique F (95%)**	**Différence significative**
Interaction entre les genres et le type de régime (C vs A/B)	26,22265720052	3,9739	Oui
A_F vs C_F	55,54167188784773	3,9739	Oui
B_F vs C_F	74,2848371864822	3,9739	Oui
A_F vs A_M	2,496597769934095	3,9739	Non
B_F vs B_M	5,246445082704702	3,9739	Oui
C_F vs C_M	18,804312758754218	3,9739	Oui
A_F vs B_F	1,3601079687679536	3,9739	Non
A_M vs B_M	0,2077420481638956	3,9739	Non
B_M vs C_M	3,9679016167819983	3,9739	Non

Tableau 4 : Résultats des comparaisons à posteriori

Les résultats de l'ANOVA à contrastes, réalisés avant l'ANOVA à un facteur, sont présentés dans le Tableau 4. Ils révèlent une interaction significative entre le genre et le type de régime, particulièrement marquée pour le régime C. Dans une moindre mesure, une interaction est également observée pour le régime B (F = 5,2464, légèrement supérieure à F = 3,9739). En revanche, aucune différence significative n'est détectée pour le régime de type A entre les femmes et les hommes (comparaison A_F vs A_M).

Chez les femmes, la perte de poids est fortement influencée par le type de régime. Des différences significatives sont observées entre les groupes A_F vs C_F et B_F vs C_F, soulignant l'effet marqué du régime C. En revanche, la comparaison A_F vs B_F ne montre aucune différence significative, suggérant que les régimes A et B ont des effets similaires. Chez les hommes, le type de régime n'a pas d'impact significatif sur la perte de poids. Aucune différence significative n'est détectée entre les groupes B_M vs C_M ou A_M vs B_M, ce qui suggère une réponse homogène des hommes aux différents régimes.

Ces résultats confirment que l'effet du type de régime est différencié selon le genre, particulièrement en ce qui concerne le régime C. Cela pourrait refléter des mécanismes distincts dans la manière dont les hommes et les femmes réagissent aux régimes, comme l'indique également le graphique d'interaction.

Lien avec le graphique d'interaction et les tests post-hoc

	ANOVA à contrastes – statistique F	Test post-hoc – p-value	Graphique d'interaction – évolution de la perte de poids
A_F vs C_F	55,54167188784773	0,0233	La perte de poids augmente du régime de type A au régime de type C
B_F vs C_F	74,2848371864822	0,0052	La perte de poids augmente du régime de type B au régime de type C
A_F vs A_M	2,496597769934095	0,9899	La perte de poids est similaire pour les hommes et les femmes
B_F vs B_M	5,246445082704702	0,9312	La perte de poids est similaire pour les hommes et les femmes
C_F vs C_M	18,804312758754218	0,4768	La perte de poids est plus élevée chez les femmes que chez les hommes
A_F vs B_F	1,3601079687679536	0,9962	La perte de poids est similaire chez les femmes pour les régimes de type A et B
A_M vs B_M	0,2077420481638956	1.0	La perte de poids est similaire chez les hommes pour les régimes de type A et B
B_M vs C_M	3,9679016167819983	0,9672	La perte de poids est similaire chez les hommes pour les régimes de type B et C

Tableau 5 : Comparaison des différents résultats

Dans le Tableau 5, nous avons synthétisé les résultats des analyses effectuées (ANOVA à contrastes, tests post-hoc et graphique d'interaction). Globalement, ces méthodes confirment des résultats cohérents pour la majorité des comparaisons de groupes (les cases grises indiquent une différence significative entre deux groupes). Par exemple, l'ANOVA à contrastes, les tests post-hoc et le graphique d'interaction montrent tous que le type de régime influence significativement la perte de poids chez les femmes. Plus précisément, le régime de type C se distingue par une perte de poids significative par rapport aux autres régimes. En effet, on remarque que les p-value associées à l'effet du régime C v.s. l'effet du régime A ou du régime B sont très

faibles (respectivement 0.02 et 0.0052). Cela signifie que le type de régime influence la perte de poids, affirmation avec un niveau de confiance de 95%.

Néanmoins, certaines contradictions émergent, comme pour la comparaison entre les groupes B_F et B_M. Selon l'ANOVA à contrastes, il existe une différence significative entre ces deux groupes. Cependant, cette conclusion est contredite par les résultats des tests post-hoc et par l'interprétation visuelle du graphique d'interaction. En examinant les détails, on observe que pour l'ANOVA à contrastes, la statistique F obtenue est très proche de la valeur théorique (contrairement à des comparaisons plus nettes, comme entre A_F et C_F, où la statistique F était largement supérieure à la valeur théorique). Cela suggère que la différence détectée entre B_F et B_M, bien que statistiquement significative, reste faible. La p-value associée est d'ailleurs très élevée, ce qui suggère que l'influence du régime B sur le genre est négligeable.

7. Conclusions

Les analyses montrent que la perte de poids est principalement influencée par le type de régime, avec un effet significatif observé pour le régime C, qui se distingue par une efficacité supérieure par rapport aux régimes A et B. Le genre, pris isolément, n'a pas d'effet statistiquement significatif global sur la perte de poids.

Bien que l'interaction entre le genre et le type de régime ne soit pas significative à un niveau statistique strict, les résultats suggèrent une réponse différente selon le genre pour certains régimes, notamment le régime C, bien que la p-value soit largement supérieure à 0.05, ce qui suggère que l'influence est négligeable.

Annexes

Annexe 1 :

Décomposition en valeurs propres

L'ACP se base en fait sur la décomposition en valeurs propres de la matrice de covariance. En effet, la projection orthogonale d'un point x_i sur un axe de vecteur directeur u_α (soumis à la contrainte de normalisation $u_\alpha^T u_\alpha = 1$) est donnée par $z_{i\alpha} = < x_i | u_\alpha > = x_i^T u_\alpha$. En regroupant l'ensemble des individus, on peut noter $Z = XU$

où $Z_{(n \to p)}$ correspond aux nouvelles coordonnées ;

$X_{(n \to p)}$ correspond aux anciennes coordonnées ;

$U_{(p \to p)}$ représente la matrice de projection.

Pour maximiser la variance, on note en écriture compacte $z_\alpha = arg \min_{||u_\alpha||=1} u_\alpha^T X^T X u_\alpha$. Il s'agit d'un problème d'extrema liés qui peut être résolu en construisant le lagrangien. L'annulation des dérivées par rapport aux composantes u_α permet d'aboutir au résultat suivant $u_\alpha \lambda_\alpha = X^T X u_\alpha$, expression qui correspond totalement à la définition de la valeur propre de $X^T X$. Pour déterminer le premier axe principal u_1, il faut donc prendre le vecteur propre associé à la plus grande valeur propre de $X^T X$. Le deuxième axe principal correspondra quant-à-lui au vecteur propre associé à la deuxième plus grande valeur propre de $X^T X$. En poursuivant cette démarche, le vecteur U se construit petit à petit, colonne par colonne et permet d'aboutir à la représentation la plus fidèle des données de départ.

Ainsi, l'ACP correspond à la décomposition spectrale de la matrice de covariance. On obtient dès lors p axes orthogonaux sur lesquels on projette le nuage des individus.