

STA211 : Méthodes descriptives pour le pré-traitement des données

Vincent Audigier, Ndèye Niang-Keita

21 février, 2019

- 1 Introduction
- 2 Données German Credit
- 3 Analyse univariée
 - 3.1 Représentations graphiques
 - 3.1.1 Variables qualitatives
 - 3.1.2 Variables quantitatives
 - 3.2 Indicateurs numériques
 - 3.2.1 Indicateurs de tendance centrale
 - 3.2.2 Indicateurs de dispersion
 - 3.2.3 Indicateurs de forme
- 4 Analyse bivariée
 - 4.1 Lien entre variables quantitatives
 - 4.1.1 Représentation graphique
 - 4.1.2 Corrélation
 - 4.2 Lien entre variables quantitatives et qualitatives
 - 4.2.1 Représentation graphique
 - 4.2.2 Rapport de corrélation
 - 4.3 Lien entre variables qualitatives
 - 4.3.1 Table de contingence
 - 4.3.2 Représentation graphique
 - 4.3.3 Khi deux
- 5 Analyse multivariée
 - 5.1 Données quantitatives
 - 5.1.1 Analyse en composantes principales
 - 5.1.2 Application au jeu German Credit
 - 5.2 Données qualitatives
 - 5.2.1 Analyse des correspondances multiples
 - 5.2.2 Application au jeu German Credit
- Références

1 Introduction

Les données constituent la première brique d'un processus de KDD. Il est donc essentiel d'en avoir une bonne compréhension et de s'assurer de leur qualité. Les approches statistiques exploratoires, présentées dans cette partie, sont très utiles pour cela.

Une bonne compréhension signifie comprendre le sens des variables considérées, identifier précisément l'individu statistique, savoir la façon dont les données ont été collectées. La compréhension des données s'effectue de différentes manières, notamment via la lecture des documentations des fichiers de données, par des échanges avec des experts, par l'analyse statistique des données.

S'assurer de la qualité des données appelle à, d'une part, identifier les difficultés techniques que présentent les données et qu'il sera nécessaire de surmonter lors de la fouille : valeurs manquantes, valeurs erronées, valeurs aberrantes, modalités rares, distributions non Gaussiennes, liaisons non-linéaires entre les variables, données multi-groupes. D'autre part, cela appelle à développer des stratégies pour dépasser ces difficultés : soit en utilisant des méthodes statistiques avancées, qui seront développées plus loin dans ce cours, soit en apportant des modifications aux données pour pouvoir y appliquer des méthodes plus classiques, on parle de *pré-traitement*.

L'objet de ce document est de présenter, à travers un exemple, les différentes approches exploratoires permettant de comprendre et d'évaluer la qualité des données en vue de les analyser. On distinguera les approches univariées, qui consistent à s'intéresser à l'étude de chaque variable, les approches bivariées qui portent sur les couples de variables et les approches multivariées qui portent sur les liaisons entre toutes les variables simultanément.

2 Données German Credit

Le jeu de données *German Credit*, utilisé comme illustration, est issu du domaine bancaire. Il décrit 1000 dossiers de crédits à la consommation, dont 300 contiennent des impayés, par 20 variables, dont la variable à expliquer (bon payeur / mauvais payeur). Les autres variables (dites *variables explicatives*) portent sur le type de crédit demandé et sur le profil financier et personnel du demandeur. Parmi elles neuf sont

quantitatives (*i.e.* à valeurs dans l'ensemble des nombres réels), et dix sont qualitatives (*i.e.* à valeurs non-numériques). Une description plus détaillée est disponible ici (<https://archive.ics.uci.edu/ml/machine-learning-databases/statlog/german/german.doc>). Il s'agit d'un jeu de données de référence qui, bien que de taille modeste, présente une certaine richesse dans la nature de ses variables. Les données sont disponibles ici (<https://archive.ics.uci.edu/ml/machine-learning-databases/statlog/german/german.data>).

Un extrait du jeu de données est donné en Table 2.1. La Table contient l'ensemble des variables pour les 6 premiers individus. On y voit par exemple que le premier individu prends la modalité *lt.0* ('`less than 0") pour la variable *Status* (signifiant que le solde de son compte courant est négatif), la valeur 6 pour la variable *Duration* (signifiant qu'il a souscrit un crédit de 6 mois), etc. La dernière variable *Creditability* est la variable à expliquer indiquant si l'individu est un bon ou un mauvais payeur. Par exemple, le premier individu est un bon payeur, tandis que le second est un mauvais payeur.

Ce jeu comporte des données manquantes, mais celles-ci ont été regroupées avec d'autres modalités de variables qualitatives. Par exemple, la variable *Property* comporte une modalité *unknown* (*cf* Table 2.1) regroupant les individus non propriétaires et ceux pour lesquels l'information n'est pas disponible. N'étant donc pas en mesure d'identifier les données manquantes, le jeu de données sera considéré comme complet.

Table 2.1: Extrait du jeu de données German Credit

Status	Duration	History	Purpose	Credit.Amount	Savings		Length.of.current.employment	Instalments
					account/bonds	lt.100		
1	lt.0	6	critical	Radio.Television	1169	Unknown	Unemployed	
2	0.to.200	48	paidDuly	Radio.Television	5951	lt.100	4.to.7	
3	none	12	critical	Education	2096	lt.100	gt.7	

3 Analyse univariée

L'analyse univariée des variables renseigne sur leur distribution, dite *distribution marginale* (par opposition à la *distribution jointe* qui correspond à la distribution d'un ensemble de plusieurs variables). Cette analyse permettra notamment d'identifier des valeurs *aberrantes*, et de connaître des caractéristiques de la distribution (*e.g.* son caractère Gaussien). Rappelons d'abord la définition du terme *aberrant*.

Definition 3.1 (Valeurs aberrantes) Une donnée sera considérée comme aberrante si elle n'est pas issue du même modèle que celui qui tient pour la majorité des données.

Il y a parfois confusion entre les termes *aberrant* et *extrême*, certains auteurs employant le terme *extreme* pour désigner des valeurs aberrantes, au sens de la définition précédente, et réservant le terme *aberrant* pour parler de valeurs erronées, comme celles liées par exemple à des erreurs de saisie. Nous considérerons ici les valeurs erronées comme des valeurs aberrantes particulières.

Pour mettre en oeuvre l'analyse univariée, on aura recours essentiellement à la visualisation graphique, ou de tableaux, et à la lecture d'indicateurs numériques.

Dans cette partie, on notera n le nombre d'individus, X désignera une variable et (x_1, \dots, x_n) l'ensemble des n réalisations de cette variable, appelée *série statistique*. X pourra être tantôt quantitative, tantôt qualitative en fonction du contexte.

3.1 Représentations graphiques

Une variable peut être représentée synthétiquement par des graphiques, ou de façon équivalente par des tableaux, le graphique présentant l'avantage d'être plus facile à lire dès lors que la quantité d'information présentée devient importante. Les représentations graphiques des variables doivent être adaptées à la nature de celles-ci. On distinguera donc l'analyse univariée des variables qualitatives et quantitatives.

3.1.1 Variables qualitatives

On appelle variable qualitative une variable dont les valeurs sont non-numériques, on parle de *modalités*. Attention, il faut bien distinguer la valeur de la variable de son codage. Par exemple, le sexe est une variable qualitative. Ses modalités peuvent être codées *H/F*ou de façon équivalente *0/1*. Dans les deux cas, la variable restera qualitative.

Dans le cas particulier où il existe un ordre entre ses modalités, la variable est dite *qualitative ordinaire* (par opposition à une variable *qualitative nominale*). Notons qu'on pourra parfois découper une variable quantitative en classes (pour gérer la non-linéarité par exemple), on obtiendra ainsi une variable qualitative ordonnée. Les présentations des variables qualitatives nominales et qualitatives ordonnées diffèrent quelque peu.

3.1.1.1 Variables qualitatives nominales

La variable *Purpose* est l'une des variables qualitatives nominales du jeu de données German Credit, ses modalités sont *DomesticAppliance*, *Vacation*, etc (*cf* Table 2.1). On note $\{m_1, \dots, m_k\}$ l'ensemble de ses k modalités. Pour résumer la distribution d'une variable qualitative, on s'intéressera aux *fréquences absolues* et *relatives* de ses modalités.

Definition 3.2 (Fréquence absolue) La *fréquence absolue* (ou *effectif*) de la modalité m_q ($1 \leq q \leq k$) d'une variable X , est le nombre total n_q

d'individus de l'échantillon pour lesquels la variable X vaut m_q

$$n_q = \sum_{i=1}^n \mathbf{1}_{m_q}(x_i)$$

où $\mathbf{1}_{m_q}(x_i)$ vaut 1 si $x_i = m_q$ et 0 sinon.

Definition 3.3 (Fréquence relative) La *fréquence relative* de la modalité m_q d'une variable X est la proportion d'individus f_q à présenter cette modalité

$$f_q = \frac{n_q}{n}$$

Le vecteur des fréquences relatives $(f_1, \dots, f_q, \dots, f_k)$ est appelé *profil* de la variable.

La synthèse de la distribution d'une variable qualitative passe par des tableaux statistiques (*cf*Table 3.1) ou des graphiques représentant l'une ou l'autre de ces quantités (diagramme en barres, ou diagramme circulaire *cf*Figure 3.1).

Table 3.1: Résumé de la variable Purpose du jeu de données German Credit : fréquence absolue et fréquence relative

m_q	NewCar	UsedCar	Other	Furniture.Equipment	Radio.Television	DomesticAppliance	Repairs	Education	Retraining	Business
n_q	234	103	12	181	280	12	22	50	9	97
f_q	0.234	0.103	0.012	0.181	0.28	0.012	0.022	0.05	0.009	0.097

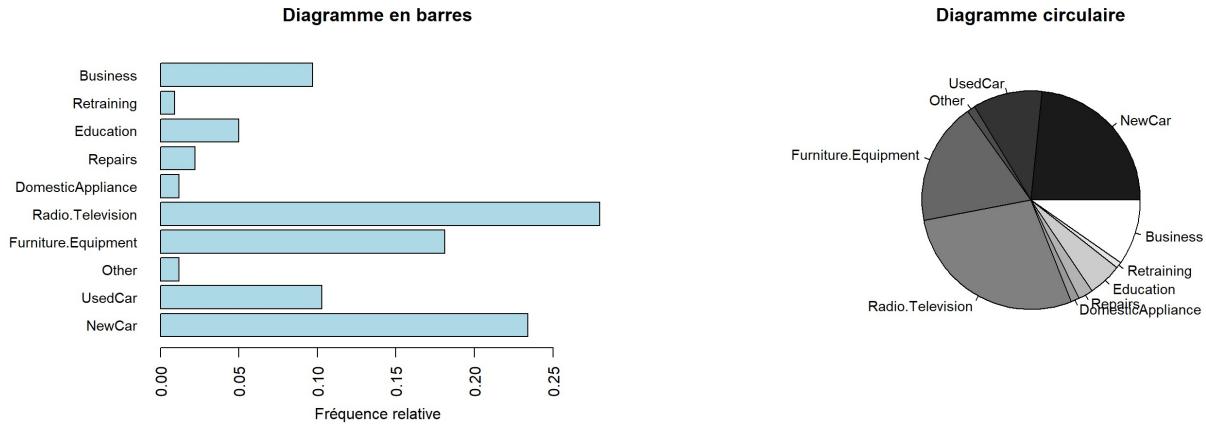


Figure 3.1: Diagramme en barres (horizontales) et diagramme circulaire décrivant la variable Purpose

Le diagramme en barres (verticales ou horizontales) est tel que la longueur de chaque barre est proportionnelle à la fréquence de la modalité associée. Le diagramme circulaire (ou camembert) est tel que chaque portion de superficie est proportionnelle à la fréquence de la modalité associée. Ces différentes présentations de la variable nous permettent d'identifier rapidement quelques modalités rares (Retraining, Domestic Appliance et Other pour la variable *Purpose*). Il pourra être pertinent de regrouper ces dernières, notamment parce qu'un trop grand nombre de modalités sur des variables explicatives d'un modèle (e.g. régression logistique) peut entraîner une grande instabilité sur l'estimation des coefficients du modèle.

3.1.1.2 Variables qualitatives ordinaires

Quand il existe un ordre entre les modalités, on s'intéresse par ailleurs aux *fréquences absolues cumulées* et aux *fréquences relatives cumulées*.

Definition 3.4 (Fréquences absolues cumulées) Pour $q' = 1, \dots, k$, on définit les *fréquences absolues cumulées* par $N_{q'} = \sum_{q=1}^{q'} n_q$. De plus, $N_0 = 0$ et $N_{q'} = N_k$ si $q' > k$.

Definition 3.5 (Fréquences relatives cumulées) Pour $q' = 1, \dots, k$, on définit les *fréquences relatives cumulées* par $F_{q'} = \sum_{q=1}^{q'} f_q$. De plus,

$F_0 = 0$ et $F_{q'} = 1$ si $q' > k$.

Celles-ci nous renseignent respectivement sur le nombre et la proportion d'observations inférieures à une modalité donnée.

Par exemple, la variable Job est une variable qualitative avec 4 modalités ordonnées (1 UnemployedUnskilled : demandeur sans emploi ou non qualifié sans résidence ; 2 UnskilledResident : sans qualification propriétaire ; 3 SkilledEmployee employé qualifié ; 4

Management.SelfEmp.HighlyQualified : manager, auto-entrepreneur, employé très qualifié), on la résume par le tableau 3.2. Par exemple, ce résumé permet de voir rapidement que 85% des demandeurs de crédits sont au mieux des employés qualifiés, tandis que 15% sont hautement

qualifiés.

Table 3.2: Résumé de la variable Job du jeu de données German Credit : fréquences absolues, fréquence relatives, fréquences absolues cumulées et fréquences relatives cumulées

m_q		n_q	f_q	$N_{q'}$	$F_{q'}$
1	UnemployedUnskilled	22	0.022	22	0.022
2	UnskilledResident	200	0.2	222	0.222
3	SkilledEmployee	630	0.63	852	0.852
4	Management.SelfEmp.HighlyQualified	148	0.148	1000	1

3.1.2 Variables quantitatives

3.1.2.1 Variables quantitatives discrètes

Une variable quantitative discrète est une variable à valeurs dans un espace dénombrable, généralement dans un sous-ensemble des entiers naturels, mais il peut aussi s'agir d'un sous ensemble des décimaux. La variable *Duration*, indiquant le nombre de mois sur lesquels porte le crédit est une des variables discrètes du jeu German Credit (cf Table 2.1) à valeurs dans l'ensemble des entiers naturels (NB : on pourrait aussi la considérer comme continue, comme cela sera discuté pour la variable *Age* en Section 3.1.2.2). Par la suite, on notera $\{e_1, e_2, \dots, e_k\}$ les différentes valeurs observées sur la série (sans doublons).

Une variable quantitative discrète pourra être présentée sous forme de tableaux d'une façon similaire à une variable qualitative ordonnée (Table 3.2). Cette représentation ne sera pertinente que si le nombre de valeurs observées est modeste. Ici la variable *Duration* prend 33 valeurs différentes, une telle représentation reste donc relativement adaptée. Notons que quand le nombre de valeurs observées devient élevé, il pourra être parfois intéressant d'utiliser une représentation *tige-et-feuille*. Celle-ci consiste à séparer la partie des dizaines de celle des unités : en face de chaque chiffre des dizaines, le chiffre des unités est répété autant de fois qu'il y a d'observations de la valeur correspondante. Un exemple est donné à titre illustratif en Figure 3.2. Naturellement, il convient d'adapter la présentation en fonction des unités de la variable.

The decimal point is 1 digit(s) to the right of the |

0	444444
0	566+87
1	0001111111112+174
1	555+129
2	00000000111111111111111111111111111111111111111+174
2	67777777777777888
3	000
3	666+38
4	02222222222
4	5555578888888888888888888888888888888888+4
5	44
5	
6	00000000000000
6	

Figure 3.2: Présentation Tige-et-feuille décrivant la variable Duration. On voit que cette présentation reste limitée en présence d'un grand nombre d'individus.

Une représentation graphique par un diagramme en barres sera également adéquate pour représenter les effectifs ou les fréquences des différentes valeurs de la variable. Elle permettra de faire des premières conjectures sur sa distribution. On pourra par ailleurs représenter

graphiquement les effectifs ou fréquences cumulées, via des *diagrammes cumulatifs* (présentés plus loin dans cette section).

Ces différentes présentations sous forme de tableaux ou de graphiques nécessitent d'ordonner préalablement la série statistique.

Definition 3.6 (Statistique d'ordre) Pour une série statistique (x_1, \dots, x_n) , on range ses valeurs dans l'ordre croissant

$x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ le vecteur $(x_{(1)}, \dots, x_{(n)})$ est appelé *statistique d'ordre* de la série

La Table 3.3 reporte un extrait de la statistique d'ordre pour la variable *Duration*. On constate que la plus petite valeur de la série est 4, présente 6 fois. A partir de cette série ordonnée, on peut construire facilement le tableau statistique résumant la variable (Table 3.4).

Table 3.3: 20 premières valeurs de la variable Duration du jeu de données German Credit et statistique d'ordre

	i=1	i=2	i=3	i=4	i=5	i=6	i=7	i=8	i=9	i=10	i=11	i=12	i=13	i=14	i=15	i=16	i=17	i=18	i=19	i=20
x_i	6	48	12	42	24	36	24	36	12	30	12	48	12	24	15	24	24	30	24	24
$x_{(i)}$	4	4	4	4	4	4	5	6	6	6	6	6	6	6	6	6	6	6	6	

Table 3.4: Résumé de la variable Duration du jeu de données German Credit : fréquences absolues, fréquence relatives, fréquences absolues cumulées et fréquences relatives cumulées

e_q	n_q	f_q	$N_{q'}$	$F_{q'}$
4	6	0.006	6	0.006
5	1	0.001	7	0.007
6	75	0.075	82	0.082
7	5	0.005	87	0.087
8	7	0.007	94	0.094

Une représentation des fréquences cumulées est indiquée en Figure 3.3. La fonction représentée porte le nom de *fonction de répartition empirique*, le terme *empirique* faisant référence au fait que cette fonction est simplement déterminée à partir de la série de données.

Definition 3.7 (Fonction de répartition empirique) On appelle *fonction de répartition empirique* F_X la fonction définie par

$$F_X(x) = \begin{cases} 0 & \text{si } x < e_1 \\ F_q & \text{si } e_q \leq x < e_{q+1} \\ 1 & \text{si } x \geq e_k \end{cases}$$

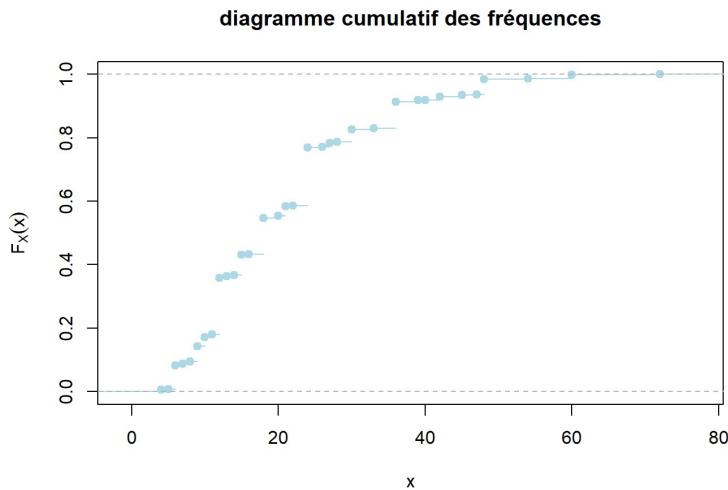


Figure 3.3: Représentation des fréquences cumulées de la variable Duration

La fonction de répartition indique la proportion de valeurs de la variable X inférieures ou égales à une valeur x . Par exemple, la Figure 3.3 indique que 99% des dossiers concernent des durées de crédit inférieure à 50 mois.

3.1.2.2 Variables quantitatives continues

Une variable quantitative est dite *continue* lorsqu'elle est à valeurs dans l'ensemble des nombres réels. En pratique, du fait du codage numérique de la variable avec un nombre fini de chiffres, il ne s'agit pas de valeurs précises mais d'intervalles. Par exemple, l'âge est une variable de nature quantitative continue, mais son codage en années équivaut à considérer des intervalles de temps de longueur égale à une année. D'un point de vue pratique, la distinction entre une variable continue et discrète n'est pas toujours simple. Par exemple, l'âge en années correspond aussi au nombre d'années révolues depuis la naissance et de ce point de vue peut être vu comme une variable discrète. Le même commentaire pourrait tenir pour la variable *Duration* : en considérant qu'elle porte sur la durée du crédit on peut la considérer comme une variable continue, tandis qu'elle sera discrète si elle est vue comme le nombre de mois sur lesquels le crédit porte. In fine, c'est surtout le nombre de valeurs prises par la variable quantitative qui sera l'argument pour déterminer sa représentation comme variable discrète ou continue.

Afin de décrire la présentation des variables continues, nous rappelons quelques définitions classiques.

Definition 3.8 (Classe) Soit a_0 et a_k deux réels tels que $a_0 \leq x_{(1)}$ et $a_k \geq x_{(n)}$ et soit une partition de l'intervalle $[a_0, a_k]$ en k intervalles. On appelle *classe* tout intervalle de la forme $[a_q - 1, a_q]$ ($0 \leq q \leq k$)

Definition 3.9 (Amplitude) On appelle *amplitude* (ou longueur) de la classe $[a_{q-1}, a_q]$ la différence $a_q - a_{q-1}$

Definition 3.10 (Effectif d'une classe) On appelle *effectif* de la classe $[a_{q-1}, a_q]$ le nombre n_q de valeurs de la série contenues dans cette classe

$$n_q = \sum_{i=1}^n \mathbf{1}_{[a_{q-1}; a_q]}(x_i)$$

Definition 3.11 (Fréquence d'une classe) On appelle *fréquence* de la classe $[a_{q-1}, a_q]$ la proportion $f_q = n_q/n$ de valeurs de la série contenues dans cette classe

La présentation d'une variable continue sous forme de tableau s'effectuera de façon similaire à une variable discrète en présentant les classes dans l'ordre croissant. Une présentation de la variable *Age* en 12 classes est proposée en Table 3.5.

Table 3.5: Résumé de la variable Âge du jeu de données German Credit

	[15;20[[20;25[[25;30[[30;35[[35;40[[40;45[[45;50[[50;55[[55;60[[60;65[[65;70[[70;75[
n_q	16	174	221	177	138	88	73	42	26	27	12	6
f_q	0.016	0.174	0.221	0.177	0.138	0.088	0.073	0.042	0.026	0.027	0.012	0.006
$N_{q'}$	16	190	411	588	726	814	887	929	955	982	994	1000
$F_{q'}$	0.016	0.19	0.411	0.588	0.726	0.814	0.887	0.929	0.955	0.982	0.994	1

Il existe différentes règles pour choisir l'amplitude des classes ainsi que leur nombre. De façon générale, des classes d'amplitude variable mais d'effectifs constant fourniront une description plus précise des données, tandis que des classes de longueurs fixes fourniront un résumé plus facile à lire. Le nombre de classes pourra par exemple être choisi selon la règle de Sturges préconisant de choisir $k = 1 + \ln(n)/\ln(2)$ classes.

Graphiquement, les représentations d'une variable continue sont d'une part le diagramme cumulatif, représentant la fonction de répartition empirique de façon similaire à une variable discrète, et l'*histogramme* d'autre part.

Definition 3.12 (Histogramme) On appelle *histogramme* la figure constituée des rectangles dont les bases sont les classes et dont les aires sont égales aux fréquences de ces classes.

Un histogramme dont les amplitudes des classes sont fixes sera appelé *histogramme à pas fixe*, dans le cas contraire on parlera d'*histogramme à pas variable*. Le choix des classes (nombre et amplitude) influence l'allure de l'histogramme. On représente par exemple en Figure 3.4 deux histogrammes (à pas fixe et pas variable) de la variable *Age*.

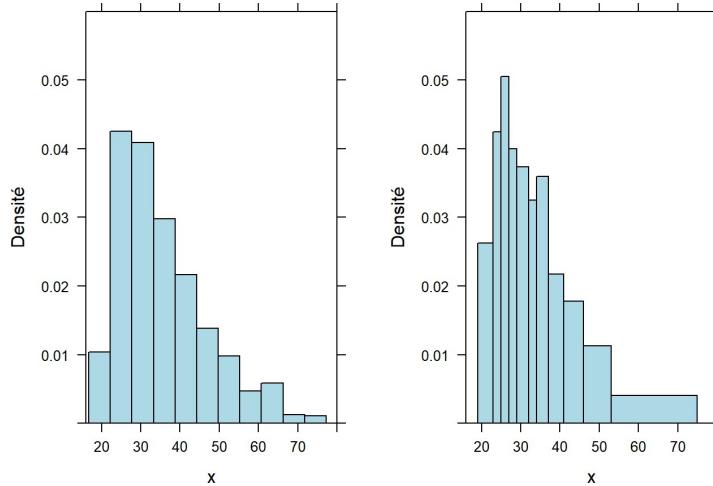


Figure 3.4: Histogramme de la variable Age à pas fixe (à gauche), à pas variable (à droite)

Ces histogrammes révèlent une asymétrie de la variable *Age* avec une queue à droite. On pourra parfois être amené à transformer cette variable de façon à se ramener à une distribution approximativement normale (e.g. si on souhaite par la suite effectuer une analyse discriminante linéaire qui repose sur la multinormalité des variables). Les transformations les plus courantes sont la transformation logarithmique ou racine carrée pour les variables positives (en ajoutant éventuellement la valeur 1 pour gérer les zéros) asymétriques à droite, tandis qu'on utilisera plutôt une transformation puissance (2 ou 3) pour les asymétries à gauche. Dans le cas de proportion, on pourra utiliser la transformation arcsinus.

Remarque : un histogramme peut être vu comme une estimation de la fonction de densité d'une variable continue par une fonction constante par morceaux (correspondant aux lignes horizontales de chaque barre de l'histogramme). Comme la fonction de densité d'une variable continue est continue, il fait sens d'utiliser une représentation plus lissée de l'histogramme qu'une fonction constante par morceaux. Ce lissage est obtenu en utilisant un *noyau* (voir par exemple Wikistat (2016b), pp. 7-8). On obtient ainsi une *courbe de densité* (un exemple est donné en Figure 3.5).

L'inspection de la distribution d'une variable quantitative peut être utile pour une discréétisation des données. En effet, certaines méthodes de data-mining nécessitent par exemple d'être appliquées sur des données de même nature (voir exemple Section 5.2.2 pour l'ACM). La discréétisation permettra de transformer toutes les variables quantitatives en variables qualitatives (ordonnées). L'inspection de la distribution permettra notamment de choisir un découpage ``naturel'' de la variable, i.e. la déformant le moins possible. Par exemple, sur la variable *Age*, un découpage naturel en 4 classes pourrait être (0,30],(30,40],(40,53],(53,75]. Notons que d'autres découpages, plus systématiques, pourraient être considérés, par exemple en formant des classes de mêmes largeurs, ou de mêmes effectifs.

3.2 Indicateurs numériques

Pour une variables quantitative, on complète la description de la série par des résumés numériques. Si le nombre de variables est grand, il sera en effet plus rapide de regarder ces indicateurs en premier lieu plutôt que d'inspecter tous les graphiques un à un. On distingue :

- les indicateurs de *tendance centrale* qui renseignent sur l'ordre de grandeur de la variable en donnant une valeur autour de laquelle les valeurs de la série se répartissent
- les indicateurs de *dispersion* qui renseignent sur la variabilité de la série, généralement autour d'un indicateur de tendance centrale
- les indicateurs de *forme* qui permettent notamment d'apprécier le caractère Gaussien des données

3.2.1 Indicateurs de tendance centrale

Les trois indicateurs de tendance centrale les plus communs sont la moyenne empirique, la médiane et le mode.

3.2.1.1 Moyenne empirique

Definition 3.13 (Moyenne empirique) On définit la moyenne empirique d'une série statistique (x_1, \dots, x_n) par $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$

Géométriquement, la moyenne empirique est le centre de gravité des n points définis par la série, affecté du même poids $1/n$. C'est un indicateur sensible aux valeurs aberrantes.

3.2.1.2 Médiane empirique

Definition 3.14 (Médiane empirique) On définit la médiane empirique d'une série statistique (x_1, \dots, x_n) par

$$\tilde{x} = \begin{cases} x_{(\lceil n/2 \rceil)} & \text{Si } n \text{ est impair} \\ \frac{x_{(\lceil n/2 \rceil)} + x_{(\lceil n/2 \rceil + 1)}}{2} & \text{sinon} \end{cases}$$

où $\lceil a \rceil$, avec a un réel, désigne la partie entière supérieure de a , i.e. le plus petit entier supérieur ou

égal à α

La médiane empirique correspond au nombre réel partageant l'échantillon en deux groupes de même effectif. C'est un indicateur robuste aux valeurs aberrantes, c'est-à-dire qu'une modification des valeurs aberrantes de la série modifie relativement peu la valeur de l'indicateur.

3.2.1.3 Mode

Definition 3.15 (Mode) Pour une variable discrète, le *mode* est défini comme la valeur la plus fréquente de la série. Pour une variable continue, il correspond à la classe la plus fréquente de l'histogramme.

Pour une variable continue, le mode dépend donc du choix des classes de l'histogramme. Notons que si la distribution de la variable est parfaitement symétrique, *i.e.* que le graphe de la densité est symétrique, on a égalité entre la moyenne, la médiane et le mode. En revanche, si elle est asymétrique à gauche (resp. à droite), la médiane est supérieure (resp. inférieure) à la moyenne.

Par exemple, pour la variable *Duration*, la moyenne vaut $\bar{x} = 20.90$, la médiane vaut $\tilde{x} = 18.00$, et le mode vaut 24.00. La dernière valeur de la série ordonnée (72) peut ici être considérée comme une valeur aberrante car assez différente des autres valeurs. En la supprimant, la moyenne vaut $\bar{x} = 20.85$, tandis que la médiane (et le mode) reste inchangée.

Parfois, la série statistique considérée est issue d'un mélange de plusieurs sous-populations (données dites *multi-groupes*). Ainsi, on peut observer plusieurs modes, un par sous-population. La courbe de densité sera très utile pour détecter le nombre de modes (cf Figure 3.5). Si des sous-populations sont répercutées, on pourra alors vouloir segmenter la population, de façon à considérer une modélisation particulière pour chaque sous-population, soit utiliser des modélisations tenant compte de cette hétérogénéité (*e.g.* via les modèles de mélanges).

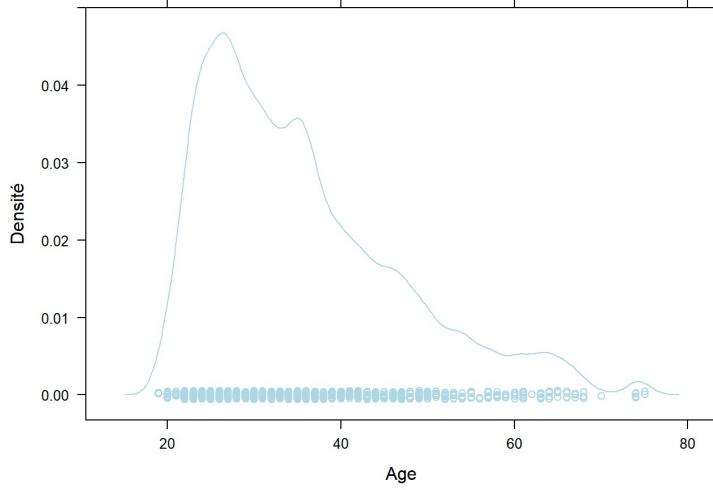


Figure 3.5: Courbe de densité de la variable Age

Le caractère multimodal de la variable *Age* n'est pas évident. On identifie clairement un mode autour de 25 ans, et potentiellement un autre autour de 36 ans.

3.2.2 Indicateurs de dispersion

3.2.2.1 Variance empirique

Definition 3.16 (Variance empirique) On définit la variance empirique d'une série statistique (x_1, \dots, x_n) par

$$s_X^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

$s_X = \sqrt{s_X^2}$ est appelé *écart-type*. L'intérêt de l'écart-type par rapport à la variance est qu'il s'exprime dans la même unité que la variable, tandis que la variance s'exprime en unité au carré. Il reste néanmoins difficile à interpréter.

Une limite de cet indicateur est que la dispersion doit toujours se comparer à la valeur moyenne. Par exemple, les variables *Duration* et *Age* ont des écarts-types proches (12.1 et 11.4 respectivement). Pour autant, la moyenne de la durée de crédit vaut 20.9 mois tandis que la moyenne de l'âge est de 35.5 ans. Ainsi, on peut considérer que, bien que les écart-types soient proches, relativement à leurs moyennes, la durée du crédit varie nettement plus d'un individu à l'autre que l'âge du client.

3.2.2.2 Coefficient de variation empirique

Definition 3.17 (Coefficient de variation empirique) On définit le *coefficient de variation empirique* d'une série statistique (x_1, \dots, x_n) par

$$CV = \frac{s_X}{\bar{x}}$$

Il s'agit d'un indicateur sans dimensions qui contrairement au précédent relativise la dispersion à la moyenne. En pratique, on utilise le seuil de

0.15 pour établir une faible/forte variabilité. Sur l'exemple précédent, le coefficient de variation pour la variable *Duration* vaut 0.57 contre 0.32 pour la variable *Age*.

3.2.2.3 Etendue

Definition 3.18 (Etendue) On définit l'*étendue* d'une série statistique (x_1, \dots, x_n) par

$$x_{(n)} - x_{(1)}$$

L'*étendue* est la différence du maximum et du minimum de la série, c'est un indicateur évidemment très sensible aux valeurs aberrantes. Quand on dispose de mesures d'un même caractère à des différents temps, la comparaison des étendues pour les différents temps permettra de détecter des valeurs aberrantes. Par exemple, si on mesure le poids de boîtes de conserve sur une chaîne de production à différentes heures, une différence entre la boîte la plus lourde et la plus légère "nettement" supérieure à celles relevées dans le passé indiquera qu'au moins une boîte est non-conforme, et donc qu'il y a potentiellement un problème sur la chaîne de production.

3.2.2.4 Quantiles

Les quantiles empiriques sont des valeurs qui partagent la série ordonnée en un certain nombre de parties de même effectif

Definition 3.19 (Quantile empirique) Pour une série ordonnée de valeurs $(x_{(1)}, x_{(2)}, \dots, x_{(n)})$, le quantile empirique d'ordre α ($0 < \alpha < 1$), peut être défini par

$$Q(\alpha) = \begin{cases} \frac{x_{(n\alpha)} + x_{(n\alpha+1)}}{2} & \text{si } n\alpha \text{ est un entier} \\ x_{(\lceil nq \rceil)} & \text{sinon} \end{cases}$$

D'autres définitions sont parfois utilisées dans les logiciels.

Par exemple, sur la série définie par la variable *Duration*, on a $Q(1/4) = 12$, $Q(1/2) = 18$, $Q(3/4) = 24$, ce qui indique que 25% des durées des crédits sont inférieures à 12 mois, 50% sont inférieures à 18 mois et 75% sont supérieures à 24 mois.

3.2.2.5 Distance inter-quartiles

Definition 3.20 (Distance inter-quartiles) On définit la distance inter-quartiles d'une série statistique (x_1, \dots, x_n) par

$$IQ = Q(3/4) - Q(1/4)$$

Comme tout indicateur basé sur les quantiles, la distance inter-quartiles est robuste vis-a-vis des valeurs aberrantes. On considère généralement une valeur au-delà de $Q(3/4) + 1.5 \times IQ$ (ou inférieure à $Q(1/4) - 1.5 \times IQ$) comme aberrante.

3.2.2.6 Représentation

On peut représenter une partie de ces indicateurs sous la forme d'une boîte à moustaches (aussi appelée *diagramme en boîte*, ou encore *boxplot*). Les représentations de la boîte à moustaches varient selon les logiciels. La Figure 3.6 est une des représentations possibles correspondant à la variable *Duration*.

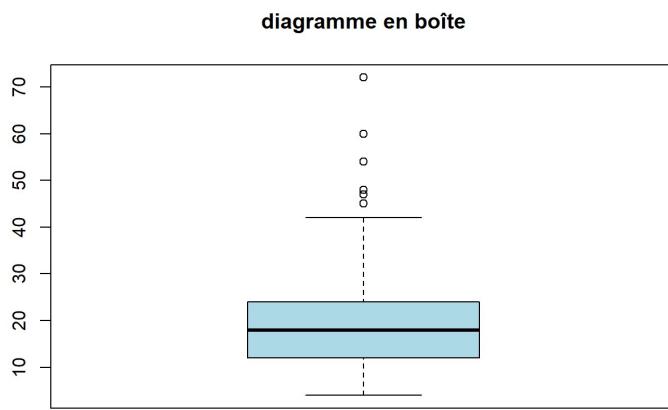


Figure 3.6: Diagramme en boîtes de la variable *Duration*

La boîte est définie par les 1er et 3eme quartiles, entre les deux est indiquée en gras la médiane. De part et d'autre de cette boîte se répartissent deux moustaches. Leur longueur correspond à 1.5 fois la distance inter-quartiles. Toutefois, si aucune valeur de la série n'excède la limite de la moustache, alors on choisit de définir l'extrémité comme la valeur la plus grande de la série pour la moustache supérieure ou comme la plus petite pour la moustache inférieure. Les valeurs au delà des moustaches sont ici représentées par des points. Cette représentation permet d'identifier rapidement des valeurs aberrantes sur une variable. Elle permet aussi d'apprécier le caractère symétrique de

celle-ci. En l'occurrence, la Figure 3.6 indique à nouveau que la variable *Duration* est asymétrique à droite et que certaines valeurs peuvent être considérées comme aberrantes (ou au moins remarquables).

La boîte à moustaches sera également utile pour comparer rapidement deux distributions. Nous y reviendrons en Section 4.2.

3.2.3 Indicateurs de forme

3.2.3.1 Coefficient d'asymétrie

Definition 3.21 (Coefficient d'asymétrie de Fisher) On définit le coefficient d'asymétrie de Fisher d'une série statistique (x_1, \dots, x_n) par

$$\gamma_1 = \frac{1}{n} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_X} \right)^3$$

Le *coefficient d'asymétrie* de Fisher vaut 0 pour une série dont la répartition est symétrique (e.g. loi normale), est positif si la queue de la distribution est à droite (e.g. loi exponentielle), et négatif si la queue de la distribution est à gauche. Il est sensible aux variations des valeurs aberrantes, au même titre que la moyenne ou la variance.

3.2.3.2 Coefficient d'aplatissement

Definition 3.22 (Coefficient d'aplatissement de Fisher) On définit le *coefficient d'aplatissement de Fisher* d'une série statistique (x_1, \dots, x_n) par

$$\gamma_2 = \frac{1}{n} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_X} \right)^4 - 3$$

Le *coefficient d'aplatissement de Fisher* sert à comparer la concentration des valeurs à celle d'une loi normale centrée pour laquelle ce coefficient vaut 3. Un coefficient d'aplatissement positif indique une densité est plus concentrée que celle d'une distribution Gaussienne, sinon il indique une densité moins concentrée. Cet indicateur, basé sur la moyenne empirique, est sensible aux valeurs aberrantes.

Ces deux coefficients sont reportés pour la variable *Age* sur la Figure 3.7. Le coefficient d'asymétrie positif indique une asymétrie à droite, et le coefficient d'aplatissement une concentration plus forte des valeurs autour de la moyenne que pour une loi normale. Le caractère Gaussian de cette variable peut donc être remis en cause, ce qui est également confirmé par le décalage entre l'histogramme (en bleu clair) et la densité attendue si les données étaient Gaussiennes (en bleu foncé). Il est aussi possible de tester l'hypothèse Gaussienne à l'aide d'un test de Kolmogorov-Smirnov ou de Shapiro-Wilk. Cependant, la distribution Gaussienne n'étant quasiment jamais observée sur les données, en travaillant sur des échantillons de grandes tailles, on tend alors à rejeter cette hypothèse de façon systématique.

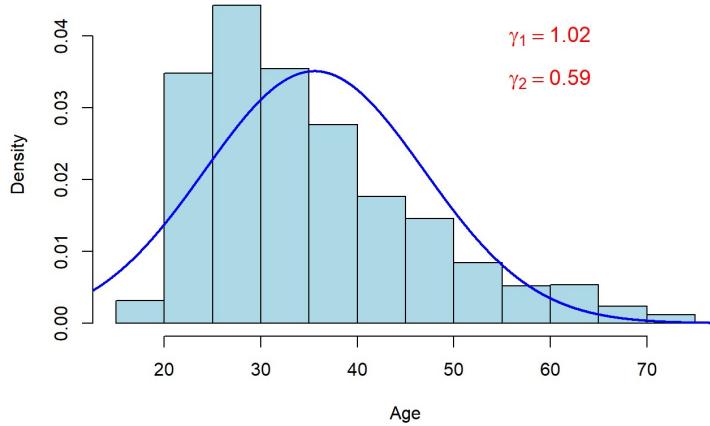


Figure 3.7: Distribution de la variable Age. En bleu clair l'histogramme, en bleu la densité de la loi normale centrée en la moyenne empirique et ayant pour variance la variance empirique.

4 Analyse bivariée

L'analyse univariée est essentielle, mais reste limitée dans le sens où elle ne renseigne pas sur les relations entre les variables. On la complète par l'étude des relations entre les variables deux à deux, appelée analyse bivariée. Cette analyse sera particulièrement utile en data-mining afin d'identifier des liaisons entre les variables à expliquer et explicatives d'une part, entre variables explicatives d'autre part. En effet, les modèles statistiques peuvent être assez influencés par ces liaisons. Par exemple, un modèle linéaire sera mis en défaut en présence de liaisons trop fortes entre variables explicatives, tandis qu'un modèle de régression logistique sera mis en défaut en présence d'une liaison trop forte entre variable explicative et la variable à expliquer. Aussi, la nature du lien entre variables explicatives et variable à expliquer amènera à choisir des

modèles adaptés ou à effectuer des transformations pour se ramener, par exemple, à un lien linéaire.

Par ailleurs, l'analyse bivariée pourra permettre de détecter des anomalies non détectables par l'analyse univariée. En effet, certains couples de valeurs peuvent être irréalisables alors que chaque valeur prise de façon isolée ne l'est pas (et ne seraient donc pas détectés par des analyses univariées). Par exemple, il est impossible qu'un demandeur de crédit à la consommation ait 18 ans et qu'il ait un emploi depuis plus de 7 ans. En revanche il n'est pas impossible qu'un demandeur de crédit ait 18 ans ou qu'un demandeur de crédit ait un emploi depuis plus de 7 ans.

L'analyse bivariée s'effectue à nouveau par l'intermédiaire de graphiques et également par la lecture d'indicateurs quantifiant la liaison entre les variables. Ceux-ci dépendent à nouveau de la nature des deux variables étudiées.

4.1 Lien entre variables quantitatives

4.1.1 Représentation graphique

L'étude d'une liaison entre deux variables quantitatives commence par sa représentation graphique via un nuage de points (Figure 4.1). Cette représentation est sensible au choix des échelles. Si les variables sont *homogènes*, i.e. qu'elles représentent la même grandeur dans une même unité, on pourra considérer les mêmes échelles sur les deux axes. Sinon, on pourra représenter les données *centrées-réduites* ou choisir des échelles telles que ce soit sensiblement ces variables là que l'on représente. Le *centrage-réduction* d'une variable consiste à lui retrancher sa moyenne (centrage), puis à diviser cette valeur par l'écart-type (réduction). La Table 4.1 fournit un extrait des données centrées-réduites pour les variables *Credit.Amount* et *Duration*.

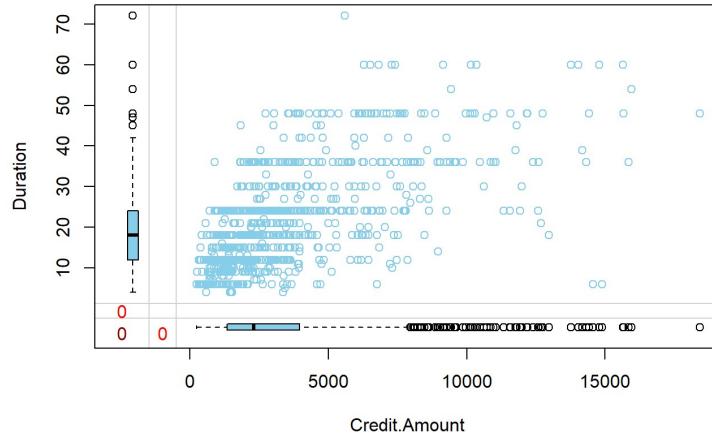


Figure 4.1: Distribution de la variable Duration en fonction de la variable Credit.Amount.

Table 4.1: Données brutes et données centrées-réduites pour les variables Credit.Amount et Duration du jeu de données German Credit

	Credit.Amount	Duration	Credit.Amount_CR	Duration_CR
1	1169.00	6.00	-0.74	-1.24
2	5951.00	48.00	0.95	2.25
3	2096.00	12.00	-0.42	-0.74
4	7882.00	42.00	1.63	1.75
5	4870.00	24.00	0.57	0.26
6	9055.00	36.00	2.05	1.25
Moyenne	3271.26	20.90	0.00	0.00
Ecart-type	2822.74	12.06	1.00	1.00

Le nuage de points permet d'identifier rapidement des aberrantes et d'apprécier la nature du lien entre les deux variables. La Figure 4.1 montre par exemple que la durée du crédit croît avec le montant de celui-ci. L'approximation linéaire est acceptable.

4.1.2 Corrélation

Le lien entre deux variables quantitatives se mesure classiquement à l'aide du coefficient de corrélation linéaire.

Definition 4.1 (Coefficient de corrélation linéaire) Soit X et Y deux variables quantitatives et $(x_1, \dots, x_n), (y_1, \dots, y_n)$ les séries statistiques correspondantes. Le coefficient de corrélation linéaire est donné par

$$\begin{aligned}\rho(X, Y) &= E\left[\frac{X - E[X]}{\sqrt{Var[X]}} \times \frac{Y - E[Y]}{\sqrt{Var[Y]}}\right] \\ &= \frac{Cov(X, Y)}{\sqrt{Var(X) Var(Y)}}\end{aligned}$$

où $E[X]$ désigne l'espérance de X , $Var[X]$ sa variance et $Cov(X, Y)$ la covariance entre X et Y . Il s'estime selon le coefficient de corrélation linéaire empirique

$$\begin{aligned}r &= \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \\ &= \frac{s_{xy}}{s_x s_y}\end{aligned}$$

Le coefficient de corrélation linéaire (ou son estimation) est compris entre -1 et 1. Une valeur égale à 0 indique une absence de lien linéaire, tandis qu'un coefficient de 1 indique un lien parfaitement linéaire. Attention, bien que l'indépendance entre les variables implique que $\rho = 0$, la nullité du coefficient n'implique pas l'indépendance entre les variables, mais simplement une absence de lien linéaire.

Quand le lien entre les variables n'est plus linéaire, on privilégiera un coefficient de corrélation basé sur les rangs (*cf Table 4.2*) comme le coefficient de *Spearman* (noté r_s) ou coefficient de *Kendall* (noté τ). Ces coefficients sont robustes aux valeurs aberrantes, mais ne permettent de détecter que des relations monotones entre les variables (comme le coefficient de corrélation linéaire).

Table 4.2: 10 premières valeurs et rangs associés des variables Duration et Credit Amount

	i=1	i=2	i=3	i=4	i=5	i=6	i=7	i=8	i=9	i=10
Duration	6	48	12	42	24	36	24	36	12	30
Rang Duration	8	937	181	920	587	831	588	832	182	788
Credit Amount	1169	5951	2096	7882	4870	9055	2835	6948	3059	5234
Rang Credit Amount	155	847	449	928	811	949	599	893	630	824

Le coefficient de Spearman évalue la corrélation linéaire sur les rangs des observations, plutôt que sur les observations elles-mêmes. Il est également compris entre -1 et 1, valant 1 si les valeurs des variables sont exactement dans le même ordre, -1 si elles sont opposées et 0 si il y a indépendance entre les deux variables. On le calcule en déterminant les rangs de chaque valeur dans les deux séries (en déterminant un rang moyen en cas d'égalité) puis en évaluant le coefficient de corrélation linéaire sur ces rangs.

Le coefficient de Kendall, lui se détermine à partir de la proportion de paires de valeurs concordantes.

Definition 4.2 (Coefficient de Kendall) Soit X et Y deux variables quantitatives et $(x_1, \dots, x_n), (y_1, \dots, y_n)$ les séries statistiques correspondantes. Le coefficient de corrélation des rangs de Kendall est donné par

$$\tau = P((X_i - X_{i'})(Y_i - Y_{i'}) > 0) - P((X_i - X_{i'})(Y_i - Y_{i'}) < 0)$$

Pour l'estimer, on note, pour tous les couples (i, i') d'individus ($i \neq i'$) : 1 si deux individus i et i' sont dans le même ordre pour les deux variables (i.e. $x_i < x_{i'}$ et $y_i < y_{i'}$) et -1 si les deux classements discordent (i.e. $x_i < x_{i'}$ et $y_i > y_{i'}$). On somme les valeurs obtenues que l'on divise par le nombre total de paires distinctes.

Le coefficient de Kendall est également compris entre -1 et 1. Il vaut 1 si les paires sont toutes concordantes, et -1 si elles sont toutes discordantes. L'indépendance entre les variables implique la nullité du coefficient, mais la réciproque est à nouveau fausse.

En général, lorsque les coefficients de corrélation de rangs sont ``nettement'' supérieurs au coefficient de corrélation linéaire, des transformations monotones non-linéaires (logarithme, racine carrée, etc) sur certaines variables peuvent se révéler utiles pour se ramener à une situation de linéarité (si cela est nécessaire pour les méthodes employées).

Les distributions des estimateurs des coefficients de corrélation, Spearman et Kendall sous l'hypothèse d'indépendance sont connues, ce qui permet de construire des tests statistiques de nullité des coefficients théoriques respectifs (voir par exemple Saporta (2006), pp.131-139).

Pour les deux variables *Duration* et *Credit Amount* on observe : $r = 0.62$, $r_s = 0.62$, $\tau = 0.47$, avec à chaque fois des p-valeurs proches de 0 traduisant un lien clair entre les deux variables. Par ailleurs, la positivité des coefficients indiquent que les deux variables évoluent dans le

même sens. On remarque que ces informations sont cohérentes avec le nuage de points en Figure 4.1.

4.2 Lien entre variables quantitatives et qualitatives

4.2.1 Représentation graphique

Pour visualiser la liaison entre deux variables quantitative et qualitative, on représente généralement la distribution de la variable quantitative en fonction des modalités de la variable qualitative (en utilisant une représentation graphique adaptée, cf Section 3.1). Si une différence entre les deux distributions (dites alors *conditionnelles*) est observée, alors on pourra conjecturer que les deux variables sont liées. En Figure 4.2 est représentée la distribution de la durée des crédit en fonction de la capacité à rembourser du client. On s'aperçoit que les mauvais payeurs ont tendance à avoir des durées de crédit plus longues que les bons payeurs (ce qui n'est pas surprenant).

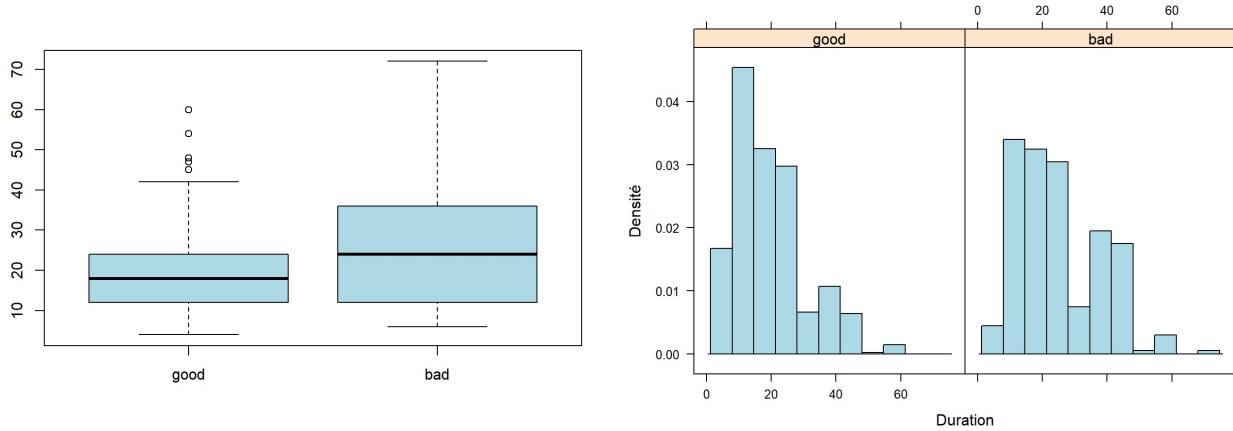


Figure 4.2: Distribution de la variable Duration en fonction de la variable Creditability : boxplots parallèles et histogrammes.

4.2.2 Rapport de corrélation

L'indicateur numérique pour apprécier la liaison entre une variable quantitative et qualitative est le *rapport de corrélation*.

Definition 4.3 (Rapport de corrélation) Soit X une variable qualitative à m_k modalités, Y une variable quantitative et $(x_1, \dots, x_n), (y_1, \dots, y_n)$ les séries statistiques correspondantes. On note Ω l'ensemble des individus de l'échantillon et Ω_q ($1 \leq q \leq k$) le sous-ensemble des individus pour lesquels $X = m_q$ et $\bar{y}_q = \frac{1}{n_q} \sum_{i \in \Omega_q} y_i$ la moyenne au sein de chaque sous-ensemble Ω_q .

Le rapport de corrélation η est défini par

$$\sqrt{\frac{\text{Var}[E[Y|X]]}{\text{Var}[Y]}}$$

où $E[Y|X]$ désigne l'espérance de Y pour une valeur de X donnée, appelée *espérance conditionnelle de Y sachant X* .

Le rapport de corrélation s'estime par le rapport de corrélation empirique

$$\sqrt{\frac{\frac{1}{n} \sum_{q=1}^k n_q (\bar{y}_q - \bar{y})^2}{\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2}}$$

Grossièrement, ce coefficient évalue (à la fonction racine carré près) la part de variance de la variable quantitative qui peut être expliquée par la variable qualitative. La variance expliquée correspond à la dispersion des valeurs moyennes au sein de chaque sous-ensemble définis par la variable qualitative. Plus la variance expliquée est proche de la variabilité de Y , plus le rapport de corrélation est proche de 1. Dans le cas où les moyennes au sein de chaque sous ensemble sont identiques, la variance expliquée est nulle et le rapport de corrélation vaut 0.

Il est à nouveau possible de tester la nullité de ce coefficient, via un test de Fisher à condition que la variance conditionnelle de Y soit la même quelque soit la valeur de X . Dans le cas où cette hypothèse n'est pas vérifiée, on pourra utiliser un test de Welch si la variable qualitative est binaire.

Le rapport de corrélation entre les variables *Duration* et *Creditability* vaut 0.21. Au vu des diagrammes en boîte parallèles (Figure 4.2), l'hypothèse d'égalité des variances ne paraît pas raisonnable. Le test de Welch nous indique un lien clair entre les deux variables, avec une p-valeur inférieure à 10^{-16} .

4.3 Lien entre variables qualitatives

4.3.1 Table de contingence

On suppose ici que X et Y sont deux variables à k et k' modalités respectivement. On présente généralement ces données sous la forme d'une *table de contingence* reportant les effectifs de chaque couple de modalités $(m_q, m_{q'})$, appelés *effectifs conjoints* et notés $n_{qq'}$ (voir Table 4.3). Les sommes en lignes et en colonnes des effectifs conjoints s'appellent respectivement *marges colonnes* et *marges lignes*. On présente en Table 4.4 la table de contingence entre la capacité à rembourser du client la présence de garant (ou co-souscripteur).

Table 4.3: Table de contingence

	m_1	m_2	...	$m_{q'}$...	$m_{k'}$	somme
m_1	n_{11}	n_{12}	...	$n_{1q'}$...	$n_{1k'}$	$n_{1.}$
m_2	n_{21}	n_{22}	...	$n_{2q'}$...	$n_{2k'}$	$n_{2.}$
...	:	:	:	:	:	:	:
m_q	n_{q1}	n_{q2}	...	$n_{qq'}$...	$n_{qk'}$	$n_{q.}$
...	:	:	:	:	:	:	:
m_k	n_{k1}	n_{k2}	...	$n_{kq'}$...	$n_{kk'}$	$n_{k.}$
somme	$n_{.1}$	$n_{.2}$...	$n_{.q'}$...	$n_{.k'}$	n

Afin d'apprécier le lien entre deux variables qualitatives, on compare les distributions conditionnelles d'une variable en fonction des niveaux de l'autre (c'est ce qui a été également effectué dans le cas d'une variable continue et qualitative). Par exemple, on comparera la capacité à rembourser en fonction de la présence de garants, plus la proportion de bons payeurs diffère selon la présence ou non de garants, plus le lien entre deux variables sera jugé fort. Pour comparer les distributions conditionnelles, on construit à partir de la table de contingence la table des *profils-lignes* (resp. *profils-colonnes*) en divisant les effectifs conjoints par les marges colonnes (resp. marges lignes). Les profils-lignes (ou profils-colonnes) sont en effet des estimations des distributions conditionnelles. Les profils-lignes et profils-colonnes de la table de contingence entre les variables sont reportées en Table 4.6 et 4.5.

La table des profils-lignes nous indique par exemple que la proportion de bon payeurs est bien plus grande en présence de garant (81 %) que quand le crédit est partagé entre deux souscripteurs (56 %), ce qui suggère un lien entre les deux variables. De la même façon, la table des profils-colonnes indique que la proportion de client à avoir des garants est deux fois plus grande les bons payeurs (6 %) que chez les mauvais (3 %)

Table 4.4: Table de contingence entre les variables Creditability et Guarantors

	good	bad	Somme
None	635	272	907
CoApplicant	23	18	41
Guarantor	42	10	52
Somme	700	300	1000
	~~		
	~~		

Table 4.6: Table des profils-lignes entre les variables Creditability et Guarantors

	good	bad	Somme
None	0.70	0.30	1
CoApplicant	0.56	0.44	1
Guarantor	0.81	0.19	1

Table 4.5: Table des profils-colonnes entre les variables Creditability et Guarantors

	good	bad
None	0.91	0.91
CoApplicant	0.03	0.06
Guarantor	0.06	0.03
Somme	1.00	1.00

4.3.2 Représentation graphique

Etant donné que les profils-lignes et profils-colonnes se déduisent directement de la table de contingence, il est possible de visualiser les différences entre les profils directement à partir de la représentation graphique de la table de contingence, appelée graphique en mosaïque (*mosaic plot*). Cette représentation consiste à représenter chaque effectif conjoint par un rectangle dont l'aire est proportionnelle à l'effectif associé. Le mosaic plot de la Table 4.3 est représenté en Figure 4.3 (graphique de gauche). Pour construire ce graphique, on a fixé la dimension du côté gauche de chaque rectangle de façon à ce qu'elle soit proportionnelle à la distribution marginale de la variable représentée à gauche (ici Guarantor). L'autre côté du rectangle est alors tel que l'aire du rectangle soit proportionnelle à l'effectif conjoint qu'il représente. Dans une situation d'indépendance, on s'attend à ce que ce second côté soit de même longueur quelque soit la modalité de la variable représentée en colonne (ici Creditability) car la fréquence conjointe est alors le produit des fréquences marginales. Cette situation est représentée dans le graphique de droite de la Figure 4.3.

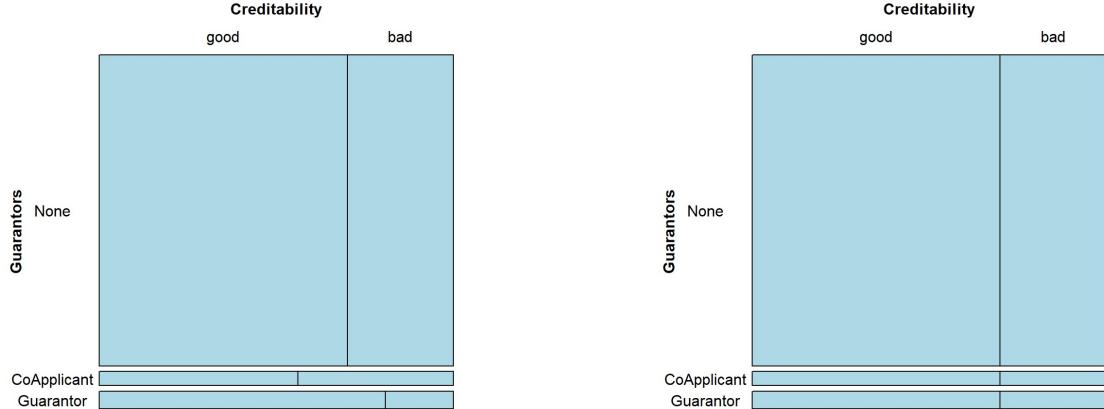


Figure 4.3: Mosaic plot de la table de contingence croisant les variables Creditability et Guarantors : table observée (à gauche) et table attendue en cas d'indépendance (à droite)

Par exemple, on voit que les rectangles correspondants aux effectifs conjoints $n_{None,good}$, $n_{CoApplicant,good}$ et $n_{Guarantor,good}$ sont de longueurs différentes dans le graphique de gauche. Ceci permet de dire que la proportion de bons payeurs est différente en fonction de la valeur de la variable *Guarantors*. Il en est naturellement de même pour les rectangles correspondants aux effectifs conjoints portant sur la modalité *mauvais payeur* ($n_{None,bad}$, $n_{CoApplicant,bad}$ et $n_{Guarantor,bad}$). Ainsi, on identifie des différences entre les profils-lignes traduisant un lien entre les deux variables.

On visualise de façon similaire les différences entre les profils-colonnes en permutant le rôle des variables.

4.3.3 Khi deux

La mesure de liaison classiquement utilisée entre deux variables qualitatives est le *khi-deux*. Elle quantifie l'écart entre les effectifs conjoints observés et ceux attendus s'il y avait indépendance entre les deux variables.

Definition 4.4 (Khi-deux)

$$\chi^2 = \sum_{q=1}^k \sum_{q'=1}^{k'} \frac{(n_{qq'} - \frac{n_{q \cdot} n_{\cdot q'}}{n})^2}{\frac{n_{q \cdot} n_{\cdot q'}}{n}}$$

En effet, sous l'hypothèse d'indépendance, on a $P(X = m_q, Y = m_{q'}) = P(X = m_q) \times P(Y = m_{q'})$, donc l'effectif conjoint attendu sous cette hypothèse est $n \times P(X = m_q) \times P(Y = m_{q'})$ et s'estime par $\frac{n_{q \cdot} n_{\cdot q'}}{n}$.

Le khi-deux a l'inconvénient de dépendre du nombre de modalités des deux variables, ainsi que du nombre d'individus. Par conséquent, on ne peut généralement pas comparer des valeurs de khi-deux entre elles afin de hiérarchiser les variables en fonction de leurs liaisons avec une autre. On lui préférera le coefficient T de Tschuprow, le coefficient V de Cramer, ou encore la p-valeur du test du chi-deux d'indépendance.

Definition 4.5 (Coefficient T de Tschuprow)

$$T = \sqrt{\frac{\chi^2/n}{\sqrt{(k-1)(k'-1)}}$$

Definition 4.6 (Coefficient V de Cramer)

$$V = \sqrt{\frac{\chi^2/n}{\min(k, k') - 1}}$$

Les coefficients T de Tschuprow et V de Cramer sont compris entre 0 et 1, une valeur élevée traduisant une liaison forte, et une valeur proche de 0 une liaison faible.

Par exemple, pour les deux variables *Creditability* et *Guarantors*, on observe une p-valeur du test du chi-deux à 0.04, un coefficient T de 0.08 et un coefficient V de 0.07. La variable la plus liée au sens de ces 3 critères avec la variable *Creditability* est la variable *Status* correspondant au statut du compte courant du demandeur de crédit (p-valeur= 10^{-26} , coefficient $T = 0.35$, coefficient $V = 0.27$), une personne avec une quantité importante d'argent sur son compte courant étant plutôt un bon payeur.

5 Analyse multivariée

Les approches descriptives univariées et bivariées permettent respectivement d'analyser les distributions de chacune des variables et leurs

relations deux à deux. Ces approches sont très utiles pour identifier les méthodes d'analyse qui pourront être appliquées par la suite ou pour identifier les transformations qui devront être apportées pour pouvoir appliquer ces méthodes sur les données pré-traitées. Néanmoins, quand le nombre de variables devient élevé, il n'est plus envisageable de regarder tous les couples de variables deux à deux, car leur nombre devient rapidement trop grand, on préférera alors utiliser les approches multivariées.

Pour effectuer des analyses descriptives multivariées, on utilise souvent les méthodes d'*analyse factorielle* et les méthodes de *classification*. Les méthodes d'analyse factorielle sont utilisées pour identifier les relations entre variables ainsi que les ressemblances entre individus. Ces méthodes reposent sur des approches géométriques dont le principe est de réduire la dimension du jeu de données en effectuant des projections des données sur des espaces de dimension inférieure. Les méthodes de classification (non-supervisées) permettent de répondre aux mêmes objectifs, en construisant une partition des individus, ou des variables, à partir de l'ensemble des données.

L'objet de cette partie est de présenter comment les différentes méthodes d'analyse multivariées permettent de comprendre et d'évaluer la structure des données en vue de les analyser. Nous nous intéressons ici à deux d'entre elles : l'analyse en composantes principales (ACP) et l'analyse des correspondances multiples (ACM). Chacune de ces méthodes est adaptée à un type de données particulier : quantitatives, qualitatives respectivement. Nous préciserons ultérieurement comment les méthodes de classification peuvent compléter ces analyses.

5.1 Données quantitatives

5.1.1 Analyse en composantes principales

5.1.1.1 Principe de la méthode

Les n individus et les p variables (quantitatives) sont décrits par une matrice $\mathbf{X}_{n \times p}$ que l'on suppose centrée. Cette matrice peut être représentée comme un nuage de n points dans \mathbb{R}^p appelé *espace des individus*, ou de façon équivalente comme un nuage de p points dans l'*espace des variables* \mathbb{R}^n (voir Figure 5.1). Par la suite, on présentera l'analyse du nuage des n individus, noté \mathcal{N} , mais une analyse similaire peut être effectuée pour le nuage des variables.

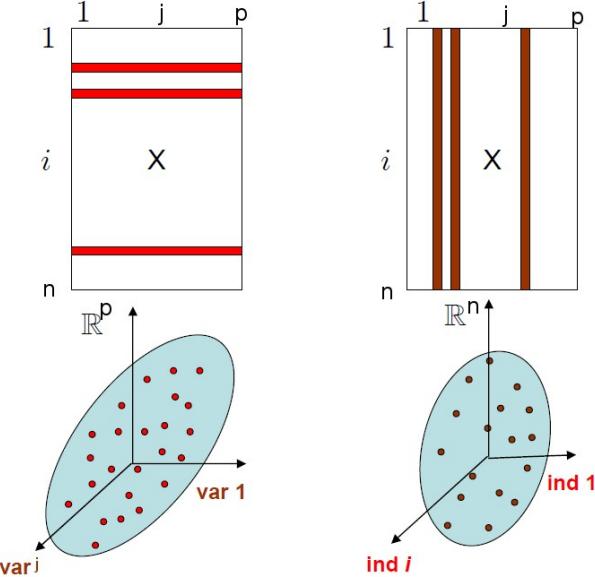


Figure 5.1: Représentation schématique du nuage des individus dans l'espace des individus (à gauche) et du nuage des variables dans l'espace des variables (à droite). Source : Husson (2016a)

Des individus avec des profils différents ont donc des coordonnées différentes dans l'espace des individus. Pour analyser la diversité des profils, l'ACP consiste à rechercher un sous-espace (par exemple un plan) dans le nuage des individus tel que la projection du nuage sur ce sous-espace résume au ``mieux'' le nuage initial. Le critère utilisé pour mesurer la qualité de la projection est l'*inertie*. L'inertie est une mesure de dispersion d'un nuage de points, elle peut être vue comme une généralisation de la notion de variance. Plus les individus sont éloignés les uns des autres, plus l'inertie du nuage est importante. Au contraire, si tous les individus sont identiques, alors l'inertie est nulle. L'objectif de l'ACP consiste alors à rechercher un sous-espace tel que l'inertie du nuage projeté sur ce sous-espace soit la plus grande possible, de façon à ce que la diversité des positions des individus projetés reflète au mieux la diversité des profils avant projection.

Afin de définir l'inertie, il faut d'abord définir une distance entre les individus. Ainsi, en ACP la distance entre deux individus i et i' est définie selon :

$$dist_{i,i'}^2 = \sum_{j=1}^p \left(\frac{x_{ij} - x_{i'j}}{s_{X_j}} \right)^2$$

Ce choix de distance correspond à celui d'une ACP dite *normée*. Il donne à chaque variable la même importance a priori quelque soit sa variance. En effet, la différence $x_{ij} - x_{i'j}$ est en générale grande si la variable j a une grande variance, et petite dans le cas contraire. Par conséquent, si

une variable avait une variance beaucoup plus élevée que toutes les autres, alors cette variable aurait tendance à être la seule à jouer un rôle dans la définition de la distance entre individus, ce qui n'est pas souhaitable car on s'intéresse à un ensemble de variables et non à une seule d'entre elles. En normalisant par l'écart-type des variables (s_X) on diminue l'importance (i.e. le poids) de cette variable de telle sorte que la distance entre deux individus n'est plus influencée par les variances.

Ce choix de distance se justifie en particulier quand les variables n'ont pas les mêmes unités, la réduction rendant les variables comparables, sans unité. En revanche, quand les variables ont les mêmes unités et que les variances ne sont pas trop différentes, il est possible d'utiliser la distance euclidienne classique (i.e. sans normalisation par l'écart-type) qui correspond à une ACP dite *non-normée*.

On voit que le choix de la distance entre les individus est directement relié à une pondération des variables. Ainsi, l'ACP est une méthode d'analyse factorielle caractérisée par une pondération particulière des variables. On affecte également un poids aux individus, généralement choisi comme uniforme. Ces pondérations sont alors résumées sous la forme de matrices notées respectivement $\mathbf{D}_{n \times n}$ et $\mathbf{M}_{p \times p}$:

$$\mathbf{D} = \begin{pmatrix} 1/n & \dots & 0 & \dots & 0 \\ 0 & \ddots & 1/n & \dots & 0 \\ \vdots & \dots & 0 & \ddots & \vdots \\ 0 & \dots & 0 & \dots & 1/n \end{pmatrix} \quad \mathbf{M} = \begin{pmatrix} 1/s_{X_1}^2 & \dots & 0 & \dots & 0 \\ 0 & \ddots & 1/s_{X_j}^2 & \dots & 0 \\ \vdots & \dots & 0 & \ddots & \vdots \\ 0 & \dots & 0 & \dots & 1/s_{X_p}^2 \end{pmatrix}$$

Dans le cas particulier de l'ACP non-normée, la matrice \mathbf{M} correspond à la matrice identité, i.e. la matrice composée de 1 sur la diagonale.

Disposant d'une distance entre les individus, et d'une pondération pour chacun d'entre eux, on peut définir l'inertie d'un nuage de points

$$\mathcal{I}(\mathcal{N}) = \sum_{i=1}^n \mathbf{D}_{ii} dist_{i,G}^2$$

avec \mathbf{D}_{ii} ($1 \leq i, i' \leq n$) l'élément (i, i') de la matrice \mathbf{D} et $G = (\bar{x}_1, \dots, \bar{x}_p)$ le centre de gravité du nuage avec $\bar{x}_j = \sum_{i=1}^n \mathbf{D}_{ii} x_{ij}$ ($1 \leq j \leq p$), le vecteur dont les coordonnées sont les moyennes des variables.

Remarquons que les données étant supposées centrées, le centre de gravité du nuage initial des individus correspond ici au vecteur nul, on peut montrer qu'il en est de même pour le nuage projeté. Par ailleurs, notons que l'expression de l'inertie coïncide avec celle de la variance pour $p = 1$. L'inertie correspond simplement à la somme des variances de chacune des variables. En particulier, pour une ACP normée, chaque variable a une variance égale à 1 et donc l'inertie du nuage est égale au nombre de variables.

L'inertie du nuage projeté se calcule de la même façon que celle du nuage initial, en considérant les coordonnées des points après projection.

Le sous-espace qui maximise l'inertie des points projetés, autrement dit, tel que la dispersion du nuage projeté soit la plus grande possible au sens des pondérations précédentes, peut se construire de manière itérative de la façon suivante : on commence par rechercher un axe (sous-espace de dimension 1) qui maximise l'inertie des points projetés. Une fois cet axe obtenu, on recherche l'axe, orthogonal au premier qui maximise l'inertie des points projetés. L'espace engendré par ces deux premiers axes forme alors un sous-espace de dimension deux (appelé plan principal). On cherche ensuite un troisième axe, orthogonal aux deux premiers, etc. Les p vecteurs qui engendrent chacun des axes forment alors un nouveau repère de l'espace des individus. On appelle *vecteurs propres*, les vecteurs normés qui définissent le nouveau repère. Les nouvelles coordonnées des individus dans ce repère définissent alors de nouvelles variables. Celles-ci sont des combinaisons linéaires des variables d'origine et sont appelées *composantes principales*. Les coefficients de cette combinaison sont donnés par les coordonnées des vecteurs propres dans le repère initial. Enfin, on appelle *valeurs propres*, l'inertie du nuage projeté sur chacun des axes, autrement dit, la variance de chaque nouvelle variable. Ainsi, l'ACP peut être vue comme la recherche des combinaisons linéaires des variables initiales de variance maximale.

L'analyse du nuage projeté sur ces différents axes nous permettra de résumer les profils des individus. Par construction, l'inertie de projection portée par le premier axe est supérieure à l'inertie portée par le second, etc. Ainsi, le premier axe résume mieux le jeu de données que le second, qui lui-même le résume mieux que le troisième, etc. On mesurera le caractère exhaustif du résumé fourni par chacun des axes par l'*inertie relative* définie par

$$\frac{\lambda_s}{\sum_{\ell=1}^S \lambda_\ell}$$

avec λ_s la valeur propre associée à l'axe s ($1 \leq s \leq S$).

Remarque : la somme des valeurs propres correspond à l'inertie totale du nuage de points avant projection. Dans le cas d'une ACP normée, elle correspond au nombre de variables.

En projetant le nuage des individus, on peut identifier les ressemblances et différences entre individus selon le principe suivant : si deux individus sont proches dans le nuage projeté, alors, dans la mesure où ils sont bien représentés, ils sont également proches dans le nuage initial, c'est-à-dire qu'ils ont des profils semblables. La qualité de représentation d'un individu sur le plan projeté se mesure par l'angle formé par : la position de l'individu dans le nuage initial, le centre de gravité du nuage et l'individu projeté. Plus cet angle est faible, meilleure est la qualité de représentation (si il est nul, alors cela signifie que l'individu et sa projection sont confondus). Plutôt que de regarder directement ces angles, on regarde plutôt leur cosinus élevé au carré. Un individu sera alors bien représenté sur un axe si son cosinus carré est proche de 1. La qualité de

représentation sur un plan est obtenue en sommant les cosinus carrés pour les deux axes qui le composent.

De la même façon que l'on résume le nuage des individus en regardant sa projection sur un sous-espace, on peut résumer le nuages des variables. Ceci permet d'identifier les relations entre variables et notamment d'identifier des groupes de variables corrélées entre elles (ce qui est plus difficile en analyse bivariée). En effet, une particularité importante est que l'angle entre deux variables correspond à la corrélation entre ces deux variables. Ainsi, si les variables sont bien représentées sur le plan, alors celui-ci nous fournit une visualisation des corrélations entre toutes les variables deux à deux. Par ailleurs, la coordonnée d'une variable sur un axe correspond à la corrélation entre cette variable et la composante principale définie par le nuage des individus. Ceci permet d'interpréter un axe en regardant les variables qui lui sont le plus corrélées.

5.1.1.2 Individus et variables supplémentaires

Les individus aberrants ont tendance à jouer un rôle important dans la construction des axes, en les écartant on peut obtenir des axes, et donc des représentations graphiques différentes. L'influence d'un individu sur la construction d'un axe se mesure en calculant sa *contribution*, définie comme le rapport entre sa coordonnée sur cet axe, élevée au carré, et la somme des coordonnées élevées au carré des individus sur ce même axe. On calcule la contribution d'un individu à la construction d'un plan en sommant les contributions à la construction des différents axes qui le composent. La simple lecture des contributions permet ainsi d'identifier les potentielles valeurs aberrantes. Plutôt que de supprimer purement et simplement les individus aberrants (i.e. aux grandes contributions), on peut les considérer comme des *individus supplémentaires* (ou *individus illustratifs*). Il s'agit de projeter ces individus sur le sous-espace sans qu'ils interviennent dans la construction des axes. Pour cela, on leur affecte un poids nul, ce qui implique que les axes seront construits sans ces individus, et on les projette ensuite a posteriori sur le sous-espace une fois les axes construits, ce qui permet néanmoins de les positionner par rapport aux autres. Par opposition, les individus qui contribuent à la construction des axes sont appelés des *individus actifs*.

Il est également possible de considérer des *variables supplémentaires*. Comme les individus supplémentaires, on pourra les représenter sur les graphiques sans qu'elles aient servi à construire les axes. Le choix des variables supplémentaires est essentiellement lié à la problématique à laquelle on veut répondre. Par exemple, dans le jeu de données German Credit, si on souhaite comprendre ce qui différencie les individus uniquement par rapport à leur profil financier on pourra mettre en supplémentaire les variables portant sur le profil personnel et le type de crédit demandé. Ainsi, les variables ayant trait au profil personnel et au type de crédit ne serviront pas à construire les axes. De la même manière, si on souhaite analyser les différents profils personnels des clients, on considérera les variables ayant trait au profil financier et au type de crédit comme variables supplémentaires. Par ailleurs, pour aider à l'interprétation des axes, on pourra aussi ajouter d'autres variables supplémentaires jugées pertinentes et voir si elles sont corrélées ou non aux différentes composantes. Par exemple, un score calculé à partir des variables actives n'est pas pertinent en tant que variable active car il n'apporte aucune information nouvelle par rapport aux autres variables, en revanche, on sait lui donner un sens, ce qui peut être pratique pour interpréter les axes, on pourra donc l'inclure en tant que variable supplémentaire.

5.1.2 Application au jeu German Credit

On effectue ici l'ACP sur l'ensemble des 7 variables quantitatives du jeu de données German Credit. Cette ACP va permettre de résumer à la fois les différences entre profils d'individus, et ainsi repérer de potentiels profils aberrants, et de résumer les relations entre variables quantitatives.

Les projections du nuage des variables et des individus sont reportées en Figure 5.2.

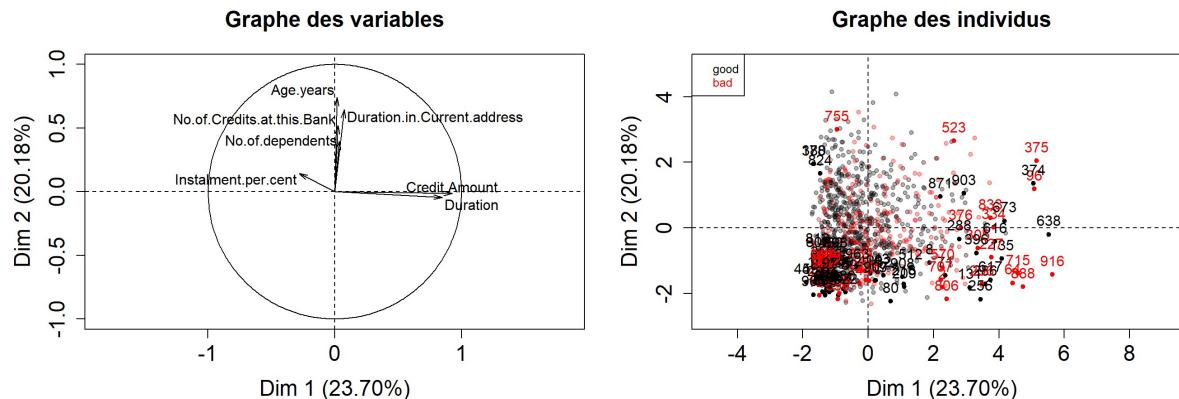


Figure 5.2: Graphes des variables (à gauche) et des individus (à droite) du jeu German Credit pour le plan principal. Sur le graphe des individus, seuls les individus les mieux représentés (cos carrés les plus élevés) ont été numérotés, les autres apparaissent en transparent.

Le graphe des variables (appelé cercle des corrélations) met en évidence deux groupes de variables positivement corrélées entre elles. Le premier groupe correspond aux variables *Credit Amount* et *Duration* qui caractérisent le type de crédit demandé. Ces variables sont les seules fortement corrélées à la première composante principale. Le second groupe rassemble les variables *Age*, *Duration in current address*, *Number of credits at this bank* et *Number of dependents*. Ces variables quant à elles caractérisent le profil financier et personnel du client. Ce sont les seules fortement corrélées à la seconde dimension. Etant donné que les composantes principales sont orthogonales, ces groupes de variables apparaissent comme plutôt indépendants entre eux, i.e. que le type de crédit demandé n'est pas relié au profil financier et personnel du client.

Enfin la variable *installment per cent* est très peu corrélée aux deux premières dimensions, elle apparaît comme relativement indépendante des autres variables.

La première dimension oppose donc à droite les individus demandant des crédits élevés sur des durées longues à droite, à des individus demandant des petits crédits sur des petites périodes à gauche. La seconde dimension quant à elle oppose en haut les individus les plus âgés, vivant depuis relativement longtemps dans la même résidence, possédant un nombre relativement élevé de résidences et ayant déjà souscrit un nombre relativement élevé de crédits, à ceux, en bas, plus jeunes, vivant depuis relativement peu de temps dans leur résidence, possédant un nombre relativement faible de résidences et ayant peu souscrit de crédits par le passé.

Ainsi, l'ACP permet de résumer les relations entre les variables quantitatives du jeu de données German Credit et d'interpréter les principales dimensions de variabilité du nuage des individus. Pour aller plus loin on ira analyser les axes suivants (en commençant par regarder le plan formé par les axes 3 et 4).

Par ailleurs, l'ACP peut permettre d'identifier d'éventuelles valeurs aberrantes via la visualisation du graphe des individus (graphique de droite en Figure 5.2), ou via la lecture des contributions des individus à la construction des axes. Ici, ce graphe ne met pas en évidence de réelles valeurs aberrantes. On note toutefois que les individus 375, 374, 96, 638, 916, se détachent un peu sur la droite du plan. Ce sont donc plutôt des individus ayant demandés des crédits très longs et très importants (relativement aux autres individus). On peut voir également que la majorité des individus se situent en bas à gauche du plan, signifiant que la majorité des individus demandent des crédits plutôt petits et courts et sont plutôt des personnes jeunes, vivant depuis peu dans leur résidence, possédant un faible nombre de résidences et ayant souscrit peu de crédits par le passé.

Enfin, bien que seules les variables quantitatives puissent être utilisées pour construire les axes en ACP, il est possible d'analyser les relations entre les variables qualitatives et les composantes principales (résumant les relations entre variables quantitatives). Ceci permet de faire le lien entre les coordonnées des individus sur un axe et ces variables. Pour cela on peut regarder directement le rapport de corrélation entre ces variables et les différentes composantes (cf Table 5.1). Par exemple, la variable *Purpose* est la plus liée à la première dimension. En effet, le montant et la durée du crédit, représentant le mieux la composante, sont directement liés à l'objet du crédit. Aussi, la variable *Creditability*, est principalement liée à la première dimension ($\eta = 0.18$ contre $0.07, 0.11, 0.06, 0.07$ pour les 4 suivantes). Ainsi, parmi les variables quantitatives, la durée et le montant du crédit sont les critères qui sont les plus liés au statut bon/mauvais payeur. En effet, on peut penser d'une part qu'il est plus facile pour le client de rembourser un crédit quand celui-ci est peu élevé et d'autre part qu'il est peu risqué pour la banque de classer un individu demandant un petit crédit comme bon payeur, et donc de lui accorder le crédit, que quand celui-ci est élevé.

Table 5.1: Valeurs des rapports de corrélation entre les variables qualitatives et les 5 premières composantes principales

	comp1	comp2	comp3	comp4	comp5
Status	0.14	0.12	0.06	0.05	0.08
History	0.22	0.36	0.07	0.32	0.38
Purpose	0.36	0.20	0.15	0.16	0.10
Savings account/bonds	0.13	0.09	0.04	0.10	0.06
Length.of.current.employment	0.11	0.45	0.13	0.13	0.06
Sex.Marital.Status	0.18	0.28	0.05	0.18	0.18
Guarantors	0.08	0.03	0.04	0.02	0.04
Property	0.33	0.21	0.14	0.09	0.11
Other.installment.plans	0.06	0.07	0.05	0.09	0.03
Housing	0.22	0.27	0.08	0.18	0.13
Job	0.28	0.11	0.18	0.03	0.11
Telephone	0.24	0.13	0.07	0.02	0.04
Foreign.Worker	0.08	0.01	0.15	0.02	0.02
Creditability	0.18	0.07	0.11	0.06	0.07

On peut également regarder où se positionnent les modalités des variables qualitatives par rapport aux différents axes (Figure 5.2). La position d'une modalité est définie par le barycentre des individus prenant cette modalité. Par exemple, on voit que la modalité *Other* de la variable *Purpose* a une coordonnée très élevée sur la première dimension, ce qui indique d'une part que les individus prenant cette modalité sont plutôt

situés sur la droite, et d'autre part que quand un individu ne précise pas la nature de l'objet de son crédit (i.e. quand il prend la modalité *Other* pour la variable *Purpose*), celui-ci correspond généralement à des crédits parmi les plus élevés et les plus longs. On voit également que les bons payeurs sont plutôt situés sur la gauche, tandis que les mauvais payeurs sont plutôt situés sur la droite, ce qui confirme l'interprétation du rapport de corrélation entre la première composante et la variable *Creditability*.

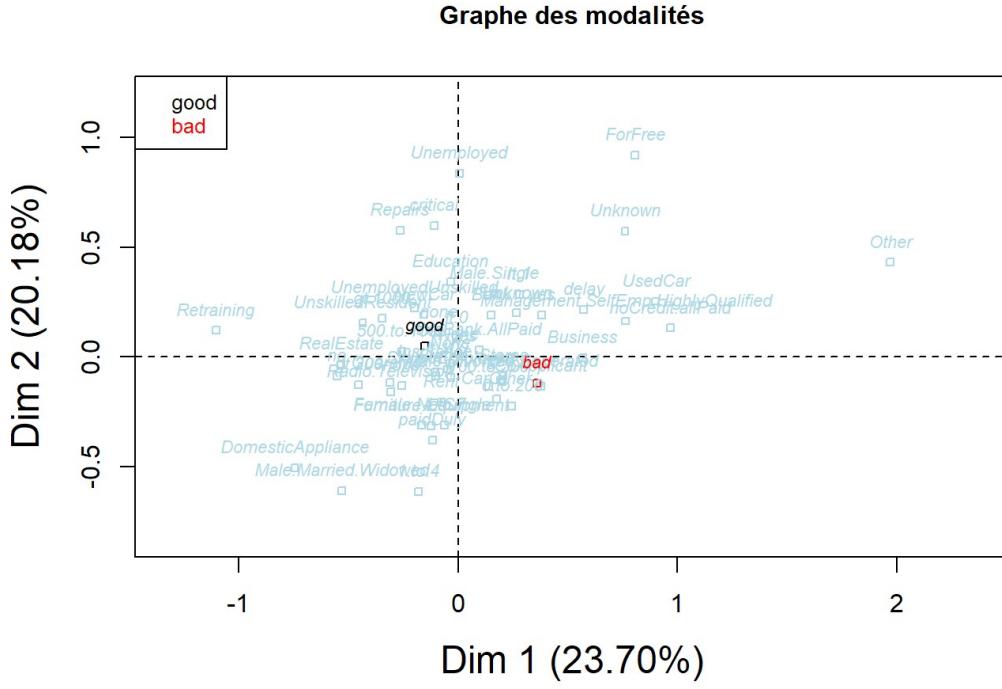


Figure 5.3: Graphes des modalités pour le plan principal.

Néanmoins, ces liaisons apparentes ne sont pas nécessairement statistiquement significatives. Naturellement, on utilisera des tests statistiques pour évaluer ce caractère significatif (voir Lebart, Morineau, and Piron (2006)). Il est alors commode de trier les modalités selon leur valeur test afin d'identifier rapidement les modalités dont les coordonnées sont significativement non nulles sur un axe (cf Table 5.2).

Table 5.2: Modalités aux plus grandes valeurs test pour la première composante

	comp1
RealEstate	-8.21
Unknown	7.97
Management.SelfEmp.HighlyQualified	7.82
none	-7.67
yes	7.67
UsedCar	7.50
ForFree	6.89
good	-5.75
bad	5.75
UnskilledResident	-5.34
Other	5.33
noCredit.allPaid	4.85
Radio.Television	-4.73

	comp1
Business	4.57
Male.Single	4.44

5.1.2.1 Choix du nombre de dimensions

Dans l'exemple précédent, nous avons regardé uniquement le plan principal, c'est-à-dire celui construit à partir des deux premiers axes. Le premier plan résume en effet 44% de l'inertie du nuage de points, ce qui est important relativement au nombre de variables, mais pas très grand dans l'absolu (56% de la variabilité des profils n'a pas été analysée). En général, il est également nécessaire d'aller analyser les plans suivants (formés par les axes 3-4, 5-6, etc). Néanmoins, l'ACP reste une méthode qui doit résumer les données, on ne peut donc pas analyser tous les plans. Il existe différentes façons de déterminer le nombre de dimensions à analyser. Nous donnons ici les plus classiques.

5.1.2.1.1 Règle de Kaiser

Le critère de Kaiser est le plus connu. Il consiste à identifier, dans le cadre d'une ACP normée, le nombre de valeurs propres supérieures à 1. En effet, dans le cas contraire, il y a moins de variabilité sur la composante que sur les variables elles-mêmes, il n'y a donc plus d'intérêt à aller analyser ces composantes. Plus généralement, cette règle consiste à analyser les axes dont les valeurs propres sont supérieures à la moyenne de toutes les valeurs propres. La Table 5.3 donne les différentes valeurs propres pour l'exemple précédent. Selon cette règle, on retient donc les 3 premiers axes.

Table 5.3: Valeurs propres associées à chaque dimension

	comp 1	comp 2	comp 3	comp 4	comp 5	comp 6	comp 7
λ_s	1.66	1.41	1.12	0.94	0.87	0.72	0.28

5.1.2.1.2 Règle du coude

Un autre critère couramment utilisé est la *règle du coude*. Cette règle consiste à regarder la décroissance des valeurs propres et d'y repérer une cassure (ou coude). Il s'agit de déterminer l'indice de l'axe pour lequel l'inertie portée est "nettement" inférieure à celle des axes précédents. On représente en Figure 5.4 les différents valeurs propres par un diagramme en barres appelé parfois *éboulis des valeurs propres*. Une cassure nette est visible entre les axes 6 et 7, ce qui suggère d'analyser 6 dimensions, mais il n'est pas pertinent d'analyser 6 axes en présence de 7 variables car on perd alors l'intérêt synthétique de la méthode. On observe également un saut moins net entre les axes 2 et 3 ce qui pourrait suggérer de regarder simplement les 2 premiers axes selon cette règle.

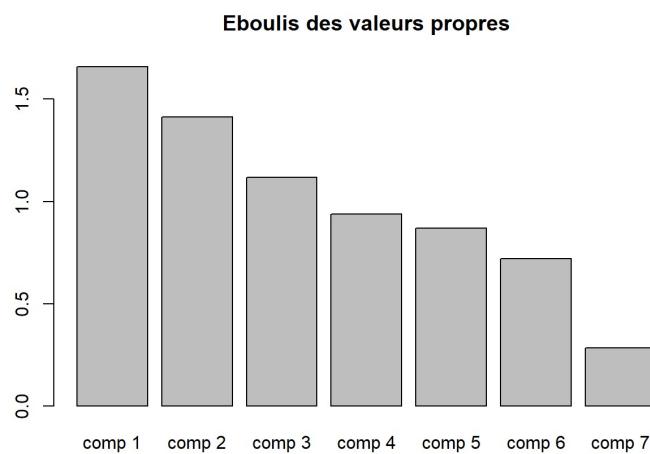


Figure 5.4: Graphique des valeurs propres

5.1.2.1.3 Autres

Les règles de Kaiser et du coude sont très classiques, mais restent indicatives, il est possible d'effectuer le choix du nombre de dimensions de façons différentes. Une possibilité consiste à simplement essayer d'interpréter chacun des plans et de s'arrêter quand il semble que la répartition des individus est aléatoire.

Plus généralement, le choix du nombre de dimensions est un sujet qui fait l'objet de nombreux travaux. Des méthodes plus avancées comme celles basées sur des approches dites de *validation croisée* peuvent notamment être utilisées. Les approches de validation croisée sont très classiques pour la validation de modèle, elles seront vues plus précisément par la suite de ce cours. Leur principe général dans le cadre de l'ACP est d'affecter des poids nuls à une partie des données (par exemple un individu, en le considérant comme supplémentaire) et de comparer les

projections de ces données, sur un sous-espace de dimension s fixé, à leurs valeurs d'origine. On répète l'opération pour différentes données (par exemple différents individus) et on moyenne les écarts obtenus. Si l'écart est grand, c'est que la projection n'est pas pertinente pour le nombre de dimensions s choisi. On fait ceci pour différentes valeurs de s et on compare les différents écarts moyens entre eux. On retient alors le nombre de dimensions qui minimise ces écarts. On représente en Figure 5.5 ces écarts en fonction du nombre d'axes retenus pour le jeu German Credit.

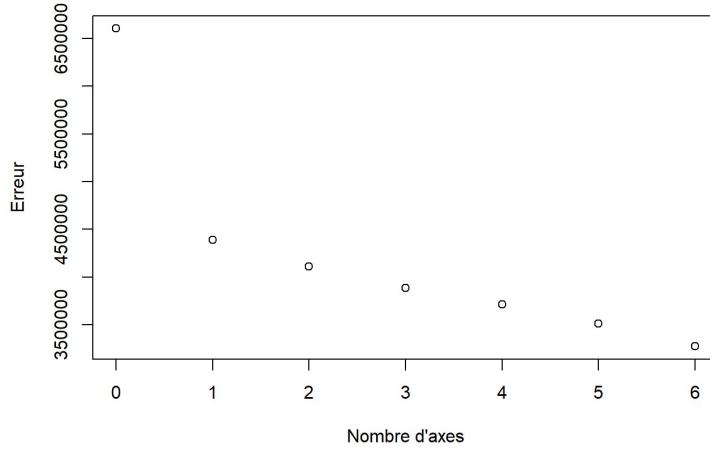


Figure 5.5: Choix du nombre de dimensions par validation croisée : erreur en fonction du nombre d'axes considérés. Le cas $s = 0$ correspond à une projection de tous les individus au centre de gravité du nuage

L'erreur est minimale pour 6 dimensions, mais à nouveau, ce choix n'est pas pertinent ici. On préférera retenir un seul axe selon ce critère car l'erreur de validation croisée ne diminue plus beaucoup au delà.

Remarque : il est possible d'affecter un poids nul à des cellules du jeu de données, plutôt qu'à des individus tout entiers. Les approches de validation croisée de ce type sont plus performantes (Bro et al. (2008), Josse and Husson (2016)).

En pratique, ces différentes manières de déterminer le nombre d'axes à analyser sont utilisées conjointement.

Au besoin, on pourra compléter cette présentation générale de l'ACP en consultant la présentation suivante et la vidéo associée : présentation (http://math.agrocampus-ouest.fr/infoglueDeliverLive/digitalAssets/100454_AnaDo_ACP_cours_slides.pdf) - vidéo (<https://www.youtube.com/watch?v=8qw0bNfK4H0>)

5.2 Données qualitatives

L'Analyse des Correspondances Multiples (ACM) peut être vue comme une extension de l'ACP utilisant d'autres pondérations des variables. La présence de variables qualitatives impose un recodage de ces variables car l'ACP s'applique sur une matrice constituée de données quantitatives uniquement.

5.2.1 Analyse des correspondances multiples

Dans le cadre de l'ACM, qui est la méthode d'analyse factorielle adaptée à des variables uniquement qualitatives, l'ensemble des variables est recodé sous la forme d'un tableau disjonctif complet : chaque variable qualitative est ainsi remplacée par autant d'indicatrices que le nombre de modalités qu'elle possède. La Figure 5.4 illustre un tel recodage.

Table 5.4: Recodage d'un jeu de données qualitatifs sous la forme d'un tableau disjonctif complet. A gauche le jeu qualitatif, à droite le tableau disjonctif complet correspondant.

Données brutes			Tableau recodé					
X_1	\dots	X_p	X_1^A	X_1^B	\dots	X_p^A	X_p^B	X_p^C
A	\dots	A	1	0	\dots	1	0	0
A	\dots	A	1	0	\dots	1	0	0
A	\dots	A	1	0	\dots	1	0	0
B	\dots	B	0	1	\dots	0	1	0

Données brutes			Tableau recodé					
X_1	\dots	X_p	X_1^A	X_1^B	\dots	X_p^A	X_p^B	X_p^C
B	\dots	C	0	1	\dots	0	0	1
B	\dots	B	0	1	\dots	0	1	0

Le tableau \mathbf{X} dans le cadre de l'ACM correspond au tableau disjonctif centré. Il possède n lignes et J colonnes où J correspond au nombre total de modalités de réponses. Ainsi, on s'intéresse ici à l'*espace des indicatrices ou des modalités* (plutôt qu'à l'*espace des variables*) et à celui des individus. La pondération adoptée sur les indicatrices est $\mathbf{M} = \frac{1}{J}\mathbf{D}_\Sigma$ avec $\mathbf{D}_\Sigma = \text{diag}(n/n_1, \dots, n/n_j, \dots, n/n_J)$, la matrice diagonale avec les inverses des proportions des individus par modalité pour éléments diagonaux.

La distance entre deux individus est alors donnée par :

$$d_{i,i'}^2 = \frac{1}{J} \sum_{j=1}^J \left(\frac{x_{ij} - x_{i'j}}{\sqrt{n_j}} \right)^2$$

Cette pondération implique notamment que deux individus ne prenant pas la même modalité de réponse sont davantage éloignés si l'une des modalités est rare. Ceci permet d'identifier facilement les individus atypiques car ceux-ci se retrouvent éloignés des autres. La pondération des individus est inchangée ($\mathbf{D} = \frac{1}{n}\mathbb{I}_n$). On obtient alors les axes de façon similaire à une ACP, i.e. en recherchant de façon itérative les axes orthogonaux qui maximisent l'inertie des points projetés.

On interprétera la dimension de l'ACM essentiellement à partir des modalités des variables qualitatives. On peut positionner celles-ci sur les axes en considérant les barycentres des individus qui prennent ces modalités (comme pour les modalités des variables illustratives en ACP). En fait, leurs positions peuvent s'obtenir directement en projetant le nuage des indicatrices (via les relations dites *de transition* permettant d'exprimer les coordonnées des individus projetés en fonction de celles des variables projetées, et réciproquement). Les modalités qui contribuent le plus à la construction des axes sont celles qui permettront de donner un sens à celui-ci. La contribution d'une modalité est donnée par le rapport

$$\frac{\frac{n_j}{np} F_j}{\lambda}$$

avec F_j la coordonnée de la modalité m_j sur l'axe considéré, λ la valeur propre associée. On remarque que dans cette expression, la contribution d'une modalité est pondérée par sa fréquence ce qui fait que la contribution d'une modalité éloignée du centre de gravité ne sera pas nécessairement très grande si elle contient peu d'individus.

On pourra également regarder les valeurs tests associées (voir Lebart, Morineau, and Piron (2006)), mais celles-ci ne doivent être considérées qu'à titre indicatif ici car les variables ont été utilisées pour la construction des axes ! Elles sont donc nécessairement liées à ces derniers.

Quand le nombre de modalités est élevé, il pourra être utile de s'appuyer sur le graphe des variables, obtenu en positionnant chaque variable qualitative en fonction de son rapport de corrélation au carré avec la composante. Il donne une idée générale des variables utiles pour interpréter les dimensions de variabilité.

Pour plus de précisions sur l'ACM, on pourra consulter le cours (http://maths.cnam.fr/IMG/pdf/ANALYSE_DES_CORRESPONDANCES_MULTIPLES-2012-2_cle838d4f.pdf) de P-L Gonzalez.

5.2.2 Application au jeu German Credit

Le jeu de données German Credit comportant des variables à la fois quantitatives et qualitatives, il n'est pas possible d'y appliquer directement une ACM (des méthodes adaptées seront présentées ultérieurement). On peut néanmoins discréteriser les variables quantitatives afin de se ramener à un jeu de variables qualitatives. Nous effectuons ici un découpage en 4 classes selon la méthodes des quantiles (conduisant à des classes de même effectifs) pour les variables continues, ou discrètes avec un nombre élevé de valeurs. Les autres sont simplement converties en variables qualitatives en créant une modalité pour chaque valeur de la variable. Notons que les modalités rares sont ici gérées par *ventilation*. La ventilation consiste à réaffecter chaque modalité rare à une modalité plus fréquentes de façon aléatoire. La variable réponse est utilisée en tant que variable supplémentaire en affectant un poids nul aux indicatrices codant pour cette variable. De cette façon, on pourra potentiellement mettre en évidence un lien entre cette variable et l'ensemble des autres variables, par l'intermédiaire des composantes.

On reporte en Figure 5.6 le graphe des individus, celui des modalités et celui des variables pour les axes 1 et 2. Seules les 10 variables les plus liées ont été reportées. Les variables *Property*, *Housing* et *Credit Amount* sont les plus liées au premier axe, tandis que les variables *Age*, *Duration*, *Credit Amount*, *Length of current employment* sont elles les plus liées à la seconde dimension. Dans les deux cas, ces relations semblent plutôt ténues.

Le graphe des modalités lui nous permet d'affiner le sens des liaisons pour les variables qualitatives. Il représente uniquement les modalités aux plus fortes contributions. On voit par exemple que les individus occupant un logement gratuitement sans en être propriétaire sont situés à droite de l'axe 1 (modalité *ForFree* de la variable *Housing*), tout comme ceux dont le statut de propriétaire est inconnu (modalité *PropUnknown* de la variable *Property*). On retrouve également de ce côté de l'axe des individus pour lesquels la durée de crédit et le montant sont élevés (modalités 4 des variables *CreditAmount* et *Duration*). A gauche de cet axe on retrouve les individus possédant des biens immobiliers (*RealEstate* de la variable *Property*) et ceux sans qualification mais propriétaire (modalité *UnskilledResident* de la variable *Job*). On retrouve également de ce côté des individus avec des crédits court de petits montants (modalités 1 des variables *CreditAmount* et *Duration*). Concernant l'axe 2, on voit par exemple que les individus empruntant peu sur des durées courtes (modalités 1 des variables *CreditAmount* et *Duration*), agés (modalités 4 de la variable *Age*), sans emploi (modalité *Unemployed* de la variable *Length.of.current.employment*) sont plutôt situés en haut de l'axe. En bas de cet axe on retrouvera des individus avec des longs crédits aux montants élevés (modalités 4 des variables *CreditAmount* et *Duration*) propriétaire d'un véhicule ou d'un autre bien (modalité *CarOther* de la variable *Property*).

Ces liaisons entre les modalités et les composantes sont confirmées par les valeurs tests (Table 5.5)

Table 5.5: Modalités aux plus grandes valeurs tests pour les deux premières composantes

comp 1		comp 2	
modalité	v.test	modalité	v.test
Property_Unknown	18.95	Duration_1	15.14
ForFree	18.90	Credit.Amount_1	14.50
Credit.Amount_4	17.19	Age.years_4	13.95
Telephone_none	-14.93	Unemployed	13.16
Telephone_yes	14.93	Property_CarOther	-12.57
Management.SelfEmp.HighlyQualified	14.50	Credit.Amount_4	-12.27
Duration_4	13.42	Duration_4	-11.67
Male.Single	13.32	No.of.dependents_1	-10.55
Property_RealEstate	-13.23	No.of.dependents_2	10.55
Age.years_1	-12.97	Age.years_1	-10.29
UsedCar	12.57	UnskilledResident	10.11
Duration_1	-12.06	Duration.in.Current.address_4	9.15
Unemployed	11.75	NewCar	9.09
Credit.Amount_1	-10.92	Property_RealEstate	9.01
Duration.in.Current.address_4	10.32	critical	8.95

Les modalités de la variable réponse sont principalement liées au second axe (valeurs tests = 4.15 en valeur absolue). Les clients classés comme mauvais payeurs étant plutôt situés en bas, et les bons payeurs en haut. Les variables liées à cette dimension apparaissent donc comme importantes pour déterminer si une personne est classée comme bon ou mauvais payeur.

Le graphe des individus ne révèle pas de valeurs clairement aberrantes.

Remarque : la présence de variables qualitatives rend plus difficile le choix du nombre d'axes à analyser. Les contraintes d'orthogonalité des indicatrices codant pour ces variables impose une répartition de l'inertie sur plusieurs dimensions, ce qui rend l'inertie projetée moins importante sur chaque axe par construction. L'application de la règle du coude (cf Section 5.1.2.1.3), basée sur la répartition des valeurs propres, ne sera donc pas très informative. On pourra cependant retenir les axes dont la variance est supérieure à $1/p$ (la moyenne des valeurs propres), ou interpréter les axes pour lesquels la position des individus projetés ne semble pas aléatoire, ou s'orienter vers des approches de validation-croisée (Josse et al. (2012), Josse and Husson (2016)).

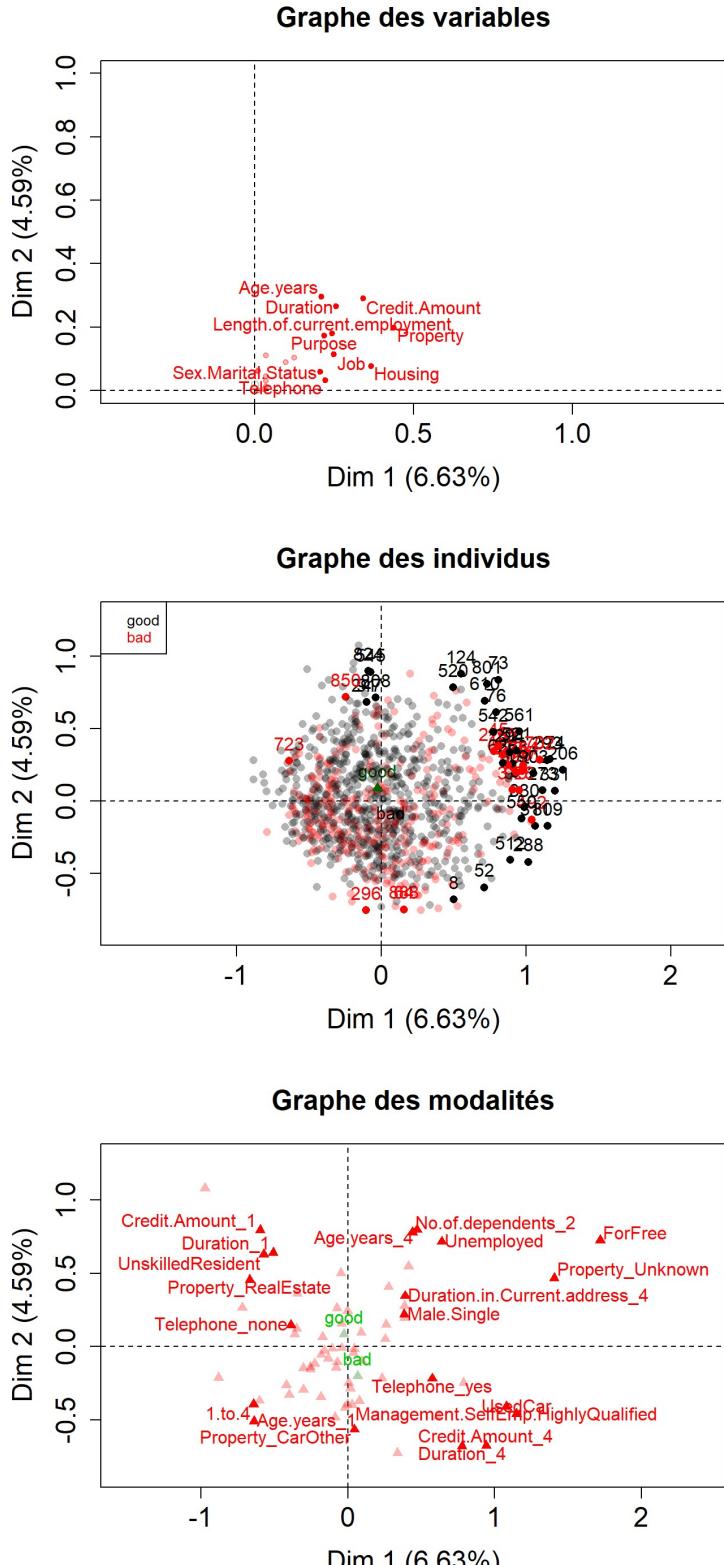


Figure 5.6: Graphe des variables, graphe des individus et graphe des modalités (plan principal). Sur le graphe des individus, seuls ceux les mieux représentés (cosinus carrés les plus élevés) ont été nommés. Sur le graphe des modalités, seules celles aux contributions les plus élevées ont été nommées.

Références

Bro, Rasmus, Karin Kjeldahl, AK Smilde, and HAL Kiers. 2008. "Cross-Validation of Component Models: A Critical Look at Current Methods." *Analytical and Bioanalytical Chemistry* 390 (5). Springer: 1241–51.

- Husson, F. 2016a. "Cours d'Analyse En Composantes Principales." [\(https://husson.github.io/img/AnaDo_ACP_cours_slides.pdf\).](https://husson.github.io/img/AnaDo_ACP_cours_slides.pdf)
- . 2016b. "Cours d'Analyse Factorielle Multiple (Cours Complet)." [\(https://www.youtube.com/watch?v=wCTaFaVKGAM\).](https://www.youtube.com/watch?v=wCTaFaVKGAM)
- Josse, J., and F. Husson. 2016. "MissMDA: A Package for Handling Missing Values in Multivariate Data Analysis." *Journal of Statistical Software, Articles* 70 (1): 1–31. doi:10.18637/jss.v070.i01 (<https://doi.org/10.18637/jss.v070.i01>).
- Josse, J., M. Chavent, B. Liquet, and F. Husson. 2012. "Handling Missing Values with Regularized Iterative Multiple Correspondence Analysis." *Journal of Classification* 29: 91–116.
- Lebart, L., A. Morineau, and M. Piron. 2006. *Statistique Exploratoire Multidimensionnelle: Visualisations et Inférences En Fouille de Données*. Sciences Sup. Mathématiques. Dunod.
- Saporta, G. 2006. *Probabilités, Analyse Des Données et Statistique*. Editions Technip.
- Tufféry, S. 2007. *Data Mining et Statistique décisionnelle: L'intelligence Des Données*. Editions Technip.
- . 2015. *Modélisation Prédictive et Apprentissage Statistique Avec R*: Éditions Technip.
- Wikistat. 2016a. "Statistique Descriptive Bimensionnelle — Wikistat." [\(http://wikistat.fr/pdf/st-l-des-bi.pdf\).](http://wikistat.fr/pdf/st-l-des-bi.pdf)
- . 2016b. "Statistique Descriptive Unidimensionnelle — Wikistat." [\(http://wikistat.fr/pdf/st-l-des-uni.pdf\).](http://wikistat.fr/pdf/st-l-des-uni.pdf)