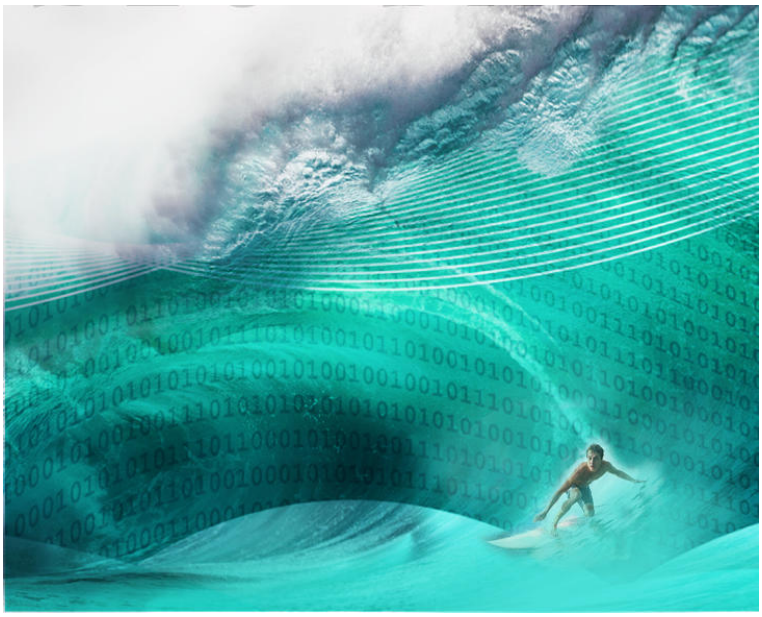
Certificat Science des Données – Module de Sensibilisation

Explorer des Données Multidimensionnelles

Philippe Besse & Béatrice Laurent

Université de Toulouse - INSA



1 Introduction

Cette partie est consacrée à la préparation des données ou *data munging* et à leur exploration. Comme il est souligné en introduction, une bonne préparation des données est essentielle pour espérer atteindre les objectifs de prévision, comme par exemple la reconnaissance de l'activité d'un porteur de smartphone.

Il se trouve que, comme pour beaucoup de données publiques, celles-ci sont déjà prêtes à l'emploi. C'est donc leur exploration qui va principalement nous occuper plus que leur "nettoyage". L'objectif est de se familiariser avec leurs propriétés, leur structure afin de guider certains choix de modélisation ou d'algorithme d'apprentissage. Néanmoins, les mêmes outils exploratoires auraient pu, le cas échéant, mettre en évidence des défaillances ou la présence d'incohérences dans les données.

La deuxième section développe une simple visualisation des signaux bruts afin de révéler les difficultés que soulèvent ce type de données. Difficultés qui ont conduit Anguita et collaborateurs (2013) à utiliser toutes les ressources du traitement du signal afin de construire de nouvelles variables, caractéristiques ou *features*, à partir des signaux bruts. Face aux dimensions des données générées et caractérisées par 561 variables, des outils d'exploration multidimensionnelle sont alors indispensables.

La section 3 propose donc une introduction à l'analyse en composantes principales (ACP), méthode conduisant à des représentations graphiques de dimension réduite, puis à l'analyse factorielle discriminante (AFD) mieux adaptée à prendre en compte la variable cible visée. Les deux méthodes sont illustrées sur des exemples de données jouet avant d'être appliquées (section 4) à l'exploration des données obtenues par transformation des signaux d'un smartphone avec pour objectif de répondre à la question : ces nouvelles variables permettront-elles de correctement distinguer ou discriminer, les classes d'activité. Un autre exemple de données complexe illustre l'ACP. Il prépare à l'objectif de prévision de la concentration en ozone.

Enfin, la section 5 introduit un algorithme élémentaire (*k-means*) de classificaiton non supervisée ou *clustering* en anglais pour l'appliquer à un problème de segmentation d'image (segmentation des pixels de l'image) afin de construire une carte géologique de Mars.

2 Prendre en Charge les Données

2.1 Data munging

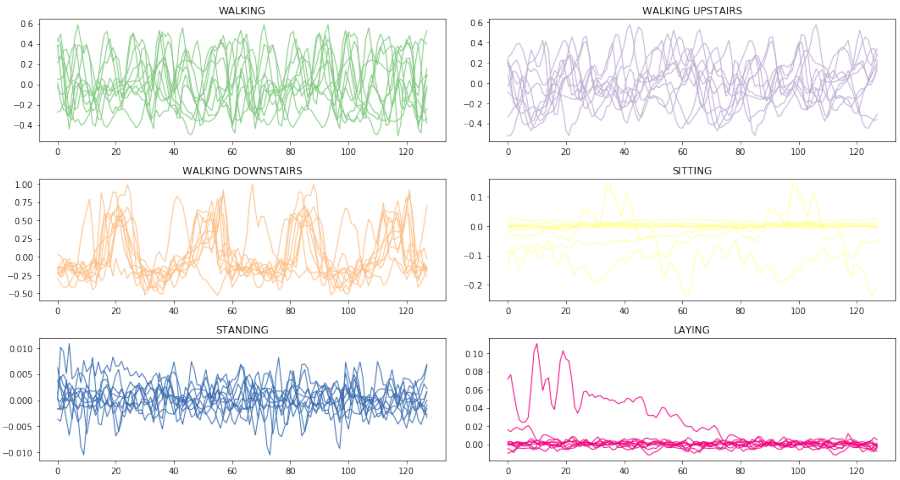
L'introduction insiste particulièrement sur le rôle crucial de la préparation des données ; détection d'erreurs ou de valeurs atypiques, analyse des distributions et transformation des variables, imputations de données manquantes... Les données fournies à la suite de l'expérimentation (Anguita et al. 2013) des différentes activités enregistrées par un smartphone sont "propres" ou alors elles ont été déjà nettoyées des scories, erreur de manipulations, pannes et autres

sources de défaillance. IL en est de même pour les données fournies par Météo France (prévision de la concentration en ozone) ou de la photo de mars ; prendre conscience que, dans beaucoup de situations réelles, ce n'est pas le cas et qu'il faut, en général, passer beaucoup de temps pour l'obtention d'une base de données qui pourra être analysée avec pertinence.

2.2 Visualiser des signaux bruts

La première opération à réaliser consiste à visualiser les données. La représentation, simultanée de tous les signaux, vue dans la première partie est bien trop confuse pour appréhender leur structure et les questions qu'elle soulève.

Aussi, la représentation ci-dessous ne concerne qu'un seul type de signal : accélération en x, pour chacune des activités.



Intuitivement il est assez clair que certaines activités : *laying*, produisent des signaux très spécifiques, donc très différents de ceux des autres activités. En revanche les différents types de marches produisent des signaux similaires avec une forte composante périodique.

La principale question concerne la façon de mesurer une *distance* entre deux signaux temporels, c'est-à-dire deux courbes ou fonctions. En effet, la distance usuelle entre deux fonctions, déduite de la norme L^2 des fonctions de carré intégrable, est l'intégrale des carrés des écarts entre 2 courbes. Comme les courbes sont discrétisées, il s'agit tout simplement de la somme des carrés des écarts ou plutôt de la racine carrée de cette quantité pour en faire une distance.

Le principal problème observé sur ces courbes concerne leur absence de synchronisation ou, c'est équivalent, leur déphasage ou décalage temporel. Au sens de la distance L_2, deux signaux correspondant à la même activité peuvent être très proches ou très éloignés par le simple fait du déphasage.

C'est la principale raison pour laquelle, vouloir analyser directement ces

signaux temporels bruts avec des distances euclidiennes classiques est voué à l'échec.

Notons par ailleurs des enregistrements atypiques, par exemple dans l'activité "couché" ; certains signaux laissent penser que le porteur était en train de se coucher, même chose pour l'activité "assis". Ces activités spécifiques seront évidemment difficiles à identifier correctement.

2.3 Transformer les signaux

Pour dépasser les questions d'absence de synchronisation des signaux, des compétences en traitement du signal sont mises à profit pour calculer toute une batterie de nouvelles variables ou caractéristiques sur ces signaux. Les détails de cette étape sont décrits par Anguita et al. (2013). Voici une liste des principales fonctions calculées sur chaque signal ou paire de signaux de chaque activité : valeur moyenne, écart-type, valeur absolue médiane, plus grande valeur, plus petite valeur, zone d'amplitude du signal, somme des carrés moyens, inter quartile, entropie, coefficient d'auto-régression, coefficient de corrélation, composante de plus grande fréquence, moyenne pondérée des fréquences, coefficient d'asymétrie des fréquences, kurtosis des fréquences, énergie dans une bande de fréquences, angle entre deux vecteurs...

Certaines transformations sont calculées sur une base de décompositions de Fourier comme l'intensité du signal dans certaines bandes de fréquences ; elles fournissent justement des quantités qui ne dépendent pas des décalages temporels des expérimentations.

Finalement, calculées sur les 9 types de signaux et leurs combinaisons deux à deux (corrélations), ce sont $p = 561$ variables qui sont considérées par la suite.

Aborder des données d'une telle complexité multidimensionnelle nécessite des moyens appropriés. C'est le domaine d'application privilégié de l'analyse en composantes principales (ACP).

3 Comprendre l'Analyse en Composantes Principales

3.1 Objectifs

Le meilleur moyen d'explorer des données complexes consiste à construire des représentations graphiques appropriées. Il est élémentaire et très utile de visualiser la nature de la liaison entre deux variables quantitatives avec un nuage de n points ; c'est rappelé dans le tutoriel de statistique descriptive (`ML4IoT-Tutorial-Ozone`). Lorsque ce sont p variables avec $p > 3$ qui sont observées, il reste possible de construire une matrice de nuages de points croisant toutes les variables 2 à 2 si p n'est pas trop grand mais l'*analyse en composantes principales* (ACP) apporte une solution plus satisfaisante surtout avec p grand. Une vignette Wikistat décrit plus en détail les fondements mathématiques de cette méthode exploratoire multidimensionnelle très utilisée.

Considérons p variables quantitatives X^j observées sur n individus ou unités statistiques. L'objectif de l'ACP est de construire une double représentation graphique.

— Représentation plane, ou de petite dimension, du nuage de points des n individus en respectant au mieux leurs positions respectives, leurs distances deux à deux,

— représentation des p variables illustrant la structure des corrélations linéaires entre celles-ci.

D'un point de vue statistique, l'ACP est la recherche de nouvelles variables, combinaisons linéaires des variables initiales et orthogonales deux à deux, de sorte que la variance de la première combinaison soit la plus grande puis celle de la deuxième, orthogonale à la précédente, soit à nouveau de plus grande variance etc...

Il est aussi possible de proposer une analogie physique. Considérons l'ensemble des n points ou individus dans l'espace $\mathbb{R}^p$. Ils constituent un solide dont les axes d'inertie sont déterminés par les vecteurs propres de la matrice d'inertie de ce solide. Ces vecteurs propres définissent les axes de plus grande dispersion des points du nuage ou solide.

Mathématiquement, l'ACP est un simple *changement de base* : passer d'une représentation dans la base canonique des variables initiales à une représentation dans la base des *facteurs* définis par les vecteurs propres de la matrice des variances-covariances (inerties) ou de celle des corrélations.

3.2 Exemple jouet

Les données

Une présentation très élémentaire de cette démarche est proposée sur un exemple jouet de données. Considérons les notes (de 0 à 20) obtenues par 9 élèves dans 4 disciplines (mathématiques, physique, français, anglais) :

	MATH	PHYS	FRAN	ANGL
jean	6.00	6.00	5.00	5.50
alan	8.00	8.00	8.00	8.00
anni	6.00	7.00	11.00	9.50
moni	14.50	14.50	15.50	15.00
didi	14.00	14.00	12.00	12.50
andr	11.00	10.00	5.50	7.00
pier	5.50	7.00	14.00	11.50
brig	13.00	12.50	8.50	9.50
evel	9.00	9.50	12.50	12.00

Il est classique d'analyser séparément chacune de ces 4 variables, soit en faisant un *graphique*, soit en calculant des *résumés numériques*. Les *liaisons entre 2 variables* (par exemple mathématiques et français), sont illustrées en faisant un graphique du type nuage de points et évaluées en calculant leur *coefficient de corrélation linéaire*.

Mais comment faire une étude simultanée des 4 variables, ne serait-ce qu'en réalisant un graphique ? La difficulté vient de ce que les individus (les élèves) ne sont plus représentés dans un plan, espace de dimension 2, mais dans un espace de dimension 4 ; chacun étant caractérisé par les 4 notes qu'il a obtenues.

L'objectif de l'analyse en composantes principales est de projeter les points sur un espace de dimension réduite (par exemple, ici, 2) en déformant le moins possible la réalité, c'est-à-dire les positions respectives des élèves entre eux. Il s'agit donc d'obtenir le *résumé le plus pertinent* des données initiales.

Descriptions uni et bivariée

Tout logiciel statistique fournit la moyenne, l'écart-type, le minimum et le maximum de chaque variable. Il s'agit donc, pour l'instant, d'études univariées.

Statistiques élémentaires

Variable	Moyenne	Ecart-type	Minimum	Maximum
MATH	9.67	3.37	5.50	14.50
PHYS	9.83	2.99	6.00	14.50
FRAN	10.22	3.47	5.00	15.50
ANGL	10.06	2.81	5.50	15.00

Notons au passage la grande homogénéité des 4 variables considérées : même ordre de grandeur pour les moyennes, les écarts-types, les minima et les maxima.

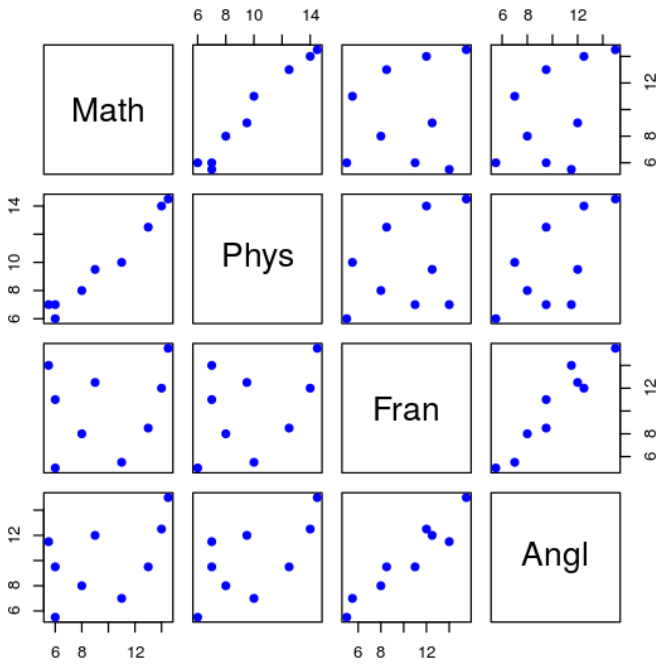
Le tableau suivant est la *matrice des corrélations*. Elle donne les coefficients de corrélation linéaire des variables prises deux à deux. C'est une succession d'analyses bivariées, constituant un premier pas vers l'*analyse multivariée*.

Coefficients de corrélation

	MATH	PHYS	FRAN	ANGL
MATH	1.00	0.98	0.23	0.51
PHYS	0.98	1.00	0.40	0.65
FRAN	0.23	0.40	1.00	0.95
ANGL	0.51	0.65	0.95	1.00

Remarquons que toutes les corrélations linéaires sont positives, ce qui signifie que toutes les variables varient, en moyenne, dans le même sens, certaines étant très fortes (0.98 et 0.95), d'autres moyennes (0.65 et 0.51), d'autres enfin plutôt

faibles (0.40 et 0.23).



La figure ci-dessus fournit la matrice des nuages de points en considérant toutes les variables deux à deux. Le principal objectif de cette représentation est de s'assurer qu'il n'existe pas de liaison *non linéaire* entre les variables. En effet une telle liaison serait négligée par des indicateurs (corrélation) ou toute analyse linéaire comme l'ACP ou la régression. Dans le cas contraire, une transformation de certaines variables (fonction puissance, log...) suffit généralement à linéariser les relations. C'est élémentaire mais indispensable à la prise en compte de liaisons non linéaires.

Décomposition spectrale de la matrice des covariances

Résultats numériques Continuons l'analyse par l'étude de la *matrice des variances-covariances*. La diagonale de cette matrice fournit les variances des 4 variables considérées.

```
          Matrice des variances-covariances

           MATH    PHYS     FRAN    ANGL
MATH      11.39    9.92     2.66    4.82
```

PHYS	9.92	8.94	4.12	5.48
FRAN	2.66	4.12	12.06	9.29
ANGL	4.82	5.48	9.29	7.91

Le tableau ci-dessous fournit les *valeurs propres* de la matrice des variances-covariances. Notez que la somme des valeurs propres est aussi la somme des variances ou trace de la matrice des covariances. D'un point de vue théorique, la trace d'un endomorphisme ne dépend pas de la base de représentation. Elle est identique dans la base canonique : matrice des covariances, ou celle des vecteurs propres : matrice diagonale des valeurs propres.

Valeurs propres ; variances expliquées

FACTEUR	VAL. PR.	PCT. VAR.	PCT. CUM.
1	28.23	0.70	0.70
2	12.03	0.30	1.00
3	0.03	0.00	1.00
4	0.01	0.00	1.00
	-----	----	
	40.30	1.00	

Interprétation statistique Chaque ligne du tableau ci-dessus correspond à une variable virtuelle (un *facteur* ou variable principale) dont la colonne VAL. PR. (valeur propre) fournit la variance. Un facteur ou variable principale est une combinaison linéaire des variables initiales dans laquelle les coefficients sont donnés par les coordonnées des vecteurs propres (changement de base).

Rappelons que l'ACP peut être définie comme la recherche des *combinaisons linéaires de plus grande variance, des variables initiales* (les valeurs propres).

La colonne PCT. VAR, ou pourcentage de variance, correspond au pourcentage de variance de chaque ligne par rapport au total. La colonne PCT. CUM. représente le cumul de ces pourcentages en dimension 1, 2... Additionnons maintenant les variances des 4 variables initiales (diagonale de la matrice des variances-covariances) : $11.39 + 8.94 + 12.06 + 7.91 = 40.30$. La dispersion totale des individus considérés, en dimension 4, est ainsi égale à 40.30.

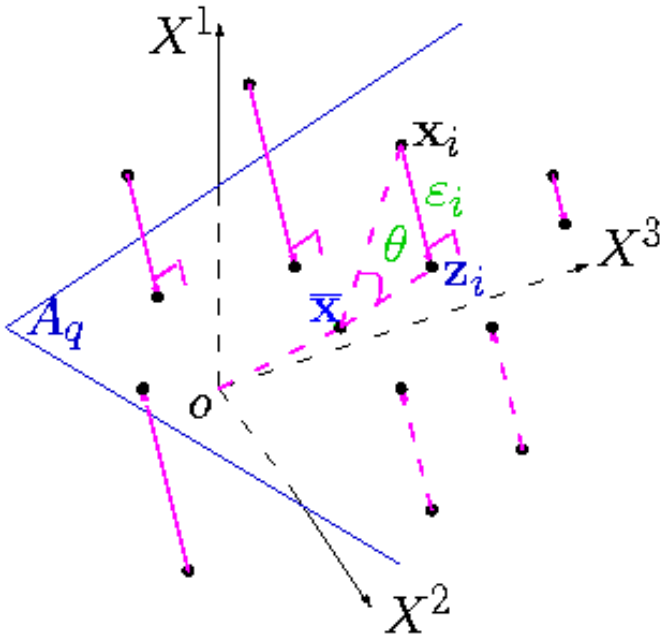
Additionnons par ailleurs les 4 valeurs propres obtenues : $28.23 + 12.03 + 0.03 + 0.01 = 40.30$. Le nuage de points en dimension 4 est toujours le même et sa dispersion globale n'a pas changé. Il s'agit d'un simple changement de base dans un espace vectoriel.

C'est la répartition de cette dispersion, selon les nouvelles variables de plus grande dispersion qui sont les facteurs ou encore *composantes principales*. Observer que les 2 premiers facteurs restituent à eux seuls la quasi-totalité de la dispersion du nuage, ce qui permet de négliger les 2 autres.

Par conséquent, les graphiques en dimension 2 présentés ci-dessous résument presque parfaitement la configuration réelle des données qui se trouvent en dimension 4. L'objectif de résumé pertinent des données en plus petite dimension est donc atteint.

Interprétation géométrique

Une autre interprétation est d'ordre géométrique. Chaque individu x_i (resp. variable x^j) est considéré comme un vecteur à p (resp. n) composantes dans un espace vectoriel. L'ACP est la recherche du meilleur plan (ou sous-espace affine) de projection A_q : le plus proche au sens des moindres carrés, pour obtenir la représentation la plus fidèle, ou la moins déformée, des individus (resp. des variables) dans un sous-espace de dimension réduite.



Sur ce graphique, z_i est la projection orthogonale de x_i, défini par le vecteurs des valeurs $X^j(x_i)$, sur le plan A_q qui passe par le barycentre $\bar{x}$ du nuage des points.

ACP réduite ou non

Les sections précédentes mettent l'accent sur la matrice de variance et sa décomposition en éléments propres. Les variables de ces données fictives présentent de bonnes propriétés, elles sont de même unité, toutes des notes, et de variances homogènes. Dans le cas contraire, hétérogénéité des unités ou des variances, il est important d'apporter une forme de normalisation. En effet, comme la variance dépend de l'unité choisie ou même si une ou des variables ont de très grandes variances par rapport aux autres variables, cela peut avoir un effet délétère sur l'intérêt de l'ACP. Une seule variable, celle évidemment de

grande variance, peut à elle seule accaparer le premier axe au détriment de la compréhension globale des relations entre les variables.

C'est la raison pour laquelle, si les variables ne partagent pas la même unité ou si de toute façon les variances sont hétérogènes, il est vivement conseillé de réduire (standardiser) les variables en les divisant par leur écart-type. Toutes les variables sont alors sans unité et de variance 1, elles jouent le même rôle et leur structure de corrélation est mise en exergue.

Comme la matrice des variances de variables réduites est la matrice des corrélation, c'est elle qui est diagonalisée pour fournir les vecteurs propres et valeurs propres de somme p, le nombre de variables, car la diagonale de la matrice des corrélations est composées de 1s.

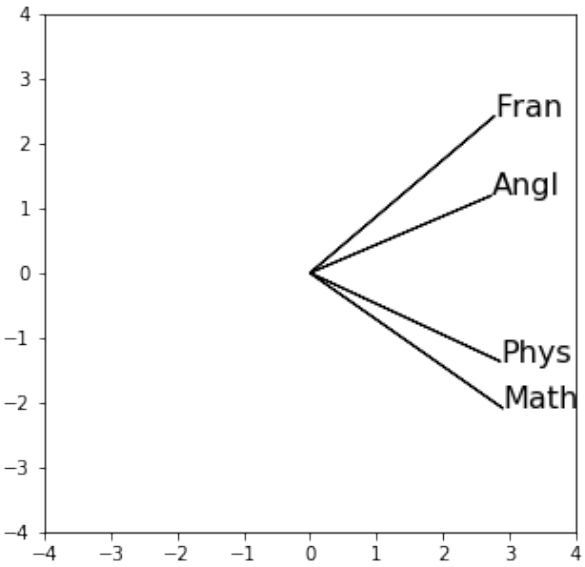
Étude des variables

Résultats numériques Un résultat important pour aider à l'interprétation est fourni par le tableau des *corrélations variables-facteurs*. Il s'agit des coefficients de corrélation linéaire entre les variables initiales et les nouvelles variables dites principales ou facteurs. Ce sont ces corrélations qui vont permettre de donner une signification aux facteurs ou variables principales pour les interpréter.

```
            Corrélations  variables-facteurs
FACTEURS  -->  F1       F2       F3       F4
    MATH      0.81    -0.58     0.01    -0.02
    PHYS      0.90    -0.43    -0.03     0.02
    FRAN      0.75     0.66    -0.02    -0.01
    ANGL      0.91     0.40     0.05     0.01
```

Les deux premières colonnes de ce tableau permettent, tout d'abord, de

réaliser le *graphique des variables* ci-dessous.



Noter que les deux dernières colonnes ne seront pas utilisées puisque que seulement deux dimensions sont nécessaires pour représenter les données.

Interprétation Par construction, le cosinus de l'angle de deux vecteurs variables approche le coefficient de corrélation entre ces variables. Ainsi, on lit sur le graphique ci-dessus que le premier facteur est corrélé positivement, et assez fortement, avec chacune des 4 variables initiales : plus un élève obtient de bonnes notes dans chacune des 4 disciplines, plus il a un score élevé sur l'axe 1 ; réciproquement, plus ses notes sont mauvaises, plus son score est négatif.

Le premier facteur, combinaison des notes avec approximativement les mêmes coefficients positifs, représente la note moyenne (centrée sur la moyenne de la classe) de chaque élève.

En ce qui concerne l'axe 2, il oppose d'une part, le français et l'anglais (corrélations positives) et, d'autre part, les mathématiques et la physique (corrélations négatives).

Le facteur 2 est approximativement la moyenne des notes littéraires déduite de la moyenne des notes scientifiques.

Cette interprétation aide à comprendre la représentation des individus.

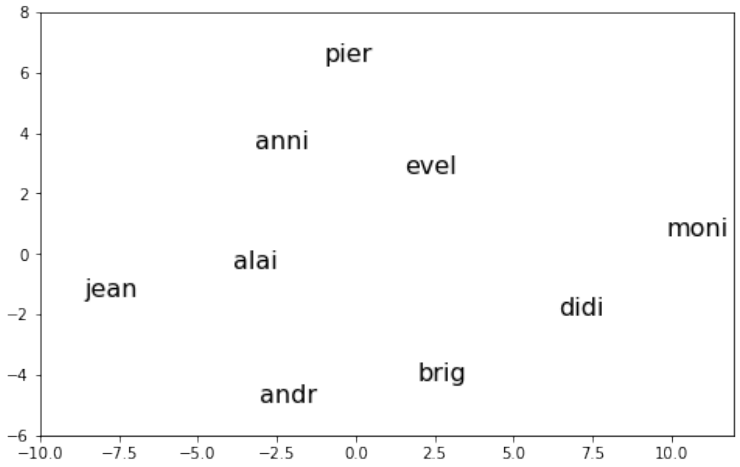
Étude des individus

Résultats numériques Le tableau ci-dessous contient tous les résultats importants sur les individus.

	Coordonnées des individus et cosinus carrés				
	POIDS	PC1	PC2	COSCA1	COSCA2
jean	0.11	-8.61	1.41	0.97	0.03
alan	0.11	-3.88	0.50	0.98	0.02
anni	0.11	-3.21	-3.47	0.46	0.54
moni	0.11	9.85	-0.60	1.00	0.00
didi	0.11	6.41	2.05	0.91	0.09
andr	0.11	-3.03	4.92	0.28	0.72
pier	0.11	-1.03	-6.38	0.03	0.97
brig	0.11	1.95	4.20	0.18	0.82
evel	0.11	1.55	-2.63	0.25	0.73

On notera que chaque individu représente 1 élément sur 9, d'où un poids (une pondération) de $1/9 = 0.11$, ce qui est fourni par la première colonne du tableau ci-dessus.

Les 2 colonnes suivantes fournissent les coordonnées des individus (les élèves) sur les deux premiers axes (les facteurs) et ont donc permis de réaliser le **graphique des individus** ci-dessous. Elles sont appelées *composantes principales* ou valeurs prises par les individus sur les variables principales, combinaisons linéaires des variables initiales.



Interprétation L'axe 1 représente le résultat d'ensemble des élèves : leur score – ou coordonnée – sur l'axe 1, fournit le même classement que leur moyenne générale. Par ailleurs, l'élève "le plus haut" sur le graphique, celui qui a la coordonnée la plus élevée sur l'axe 2, est Pierre dont les résultats sont les plus contrastés en faveur des disciplines littéraires (14 et 11.5 contre 7 et 5.5). C'est exactement le contraire pour André qui obtient la moyenne dans les disciplines scientifiques (11 et 10) mais des résultats très faibles dans les disciplines littéraires (7 et 5.5). Monique et Alain ont un score voisin de 0 sur l'axe 2 car ils ont des résultats très homogènes dans les 4 disciplines, mais à des niveaux très distincts, ce qu'a déjà révélé l'axe 1.

Compléments à l'interprétation

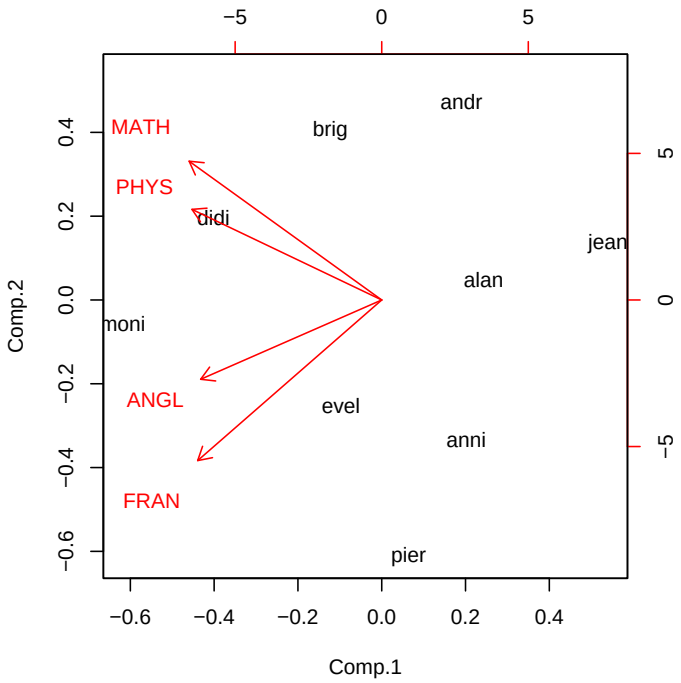
Les 2 dernières colonnes du tableau sont des cosinus carrés qui fournissent la *qualité de la représentation* de chaque individu sur chaque axe. Ces quantités s'additionnent axe par axe, de sorte que, en dimension 2, Évelyne est représentée à 98 % (0.25 + 0.73), tandis que les 8 autres individus le sont à 100 %.

Avec les données initiales, chaque individu (chaque élève) est représenté par un vecteur dans un espace de dimension 4 (les éléments – ou coordonnées – de ce vecteur sont les notes obtenues dans les 4 disciplines). Résumé en dimension 2, et donc représenté dans un plan, chaque individu est alors représenté par la projection du vecteur initial sur le plan. Le cosinus carré relativement aux deux premières dimensions (par exemple, pour Évelyne, 0.98 ou 98 %) est celui de l'angle formé par le vecteur initial et sa projection dans le plan. Plus le vecteur initial est proche du plan, plus l'angle en question est petit et plus le cosinus, et son carré, sont proches de 1 (ou de 100 %) : la représentation est alors très bonne. Au contraire, plus le vecteur initial est loin du plan, plus l'angle en question est grand (proche de 90 degrés) et plus le cosinus, et son carré, sont proches de 0 : la représentation est alors très mauvaise.

Représentation simultanée

Un troisième type de représentation graphique associant individus et variables (*biplot*) est détaillé dans le document décrivant plus précisément l'analyse en composantes principales. Ce graphe, représentant des vecteurs individus et variables appartenant à des espaces vectoriels différents nécessite un développe-

ment plus détaillé pour en justifier la construction et l'interprétation.



Ce développement est basé sur la *décomposition en valeurs singulières* de la matrice X des variables centrées. Géométriquement, le produit scalaire entre un vecteur variable et un vecteur individu fournit une approximation de la valeur de cette variable sur l'individu. Cela permet de comparer schématiquement la valeur prise par une variable sur un individu par rapport à la moyenne, le barycentre. Ainsi, Brigitte à une moyenne en maths supérieure à la moyenne de la classe mais inférieure à la moyenne de la classe en français. Monique a des notes plus grandes que la moyenne de la classe dans toutes les matières.

Cette représentation n'est possible que pour des jeux de données restreints, lorsque les dimensions sont trop importantes il devient illisible. Il a été produit par la commande `biplot` de R plutôt qu'avec Python comme pour les autres graphiques. Noter que les signes des axes ont changé avec le changement de logiciel. Plus précisément, le signe d'un vecteur propre n'est pas une information pertinente et peut changer d'un logiciel à l'autre. Ce qui et important, c'est la direction portée par le vecteur propre et pas sa direction. L'interprétation est d'ailleurs identique.

3.3 L'analyse factorielle discriminante, cas particuler d'ACP

Considérons la même situation que celle de l'ACP : p variables quantitatives X^j observées sur n individus ou unités statistiques. Ajoutons l'observation d'une variable qualitative à q classes pour définir l'analyse factorielle discriminante (AFD). Comme en ACP, l'objectif est de fournir une meilleure représentation graphique en petite dimension mais en tenant compte de l'observation de la variable qualitative, donc des classes.

L'objectif devient alors la recherche de la représentation privilégiant le plus possible les différences entre les classes d'individus au détriment des différences au sein des classes.

Principes de l'AFD

Comme vu précédemment, l'ACP est basée sur la recherche de combinaisons linéaires de plus grande variance. L'AFD, se focalise sur la variance inter-classe, pour mettre en évidence les différences des classes tout en cherchant à minimiser les effets de la variance intra-classe qui disperse les individus au sein de chaque classe.

La justification de l'AFD est détaillée dans une vignette de Wikistat. Il y est montré que l'AFD est un cas particulier optimal d'ACP permettant d'atteindre l'objectif de meilleure représentation des classes. L'AFD est finalement l'ACP des barycentres des classes afin de mettre en exergue la variance inter-classe mais en affectant l'espace des individus d'une métrique particulière, dite de *Mahalanobis*. Pour minimiser l'effet de la variance intra-classes considérée comme parasite, du bruit, la métrique de Mahalanobis est définie par la matrice inverse de la variance intra-classe, matrice carrée symétrique définie positive, .

Deux techniques cohabitent sous la même appellation d'analyse discriminante ; celle :

descriptive ou exploratoire qui vise la meilleure représentation graphique des classes,

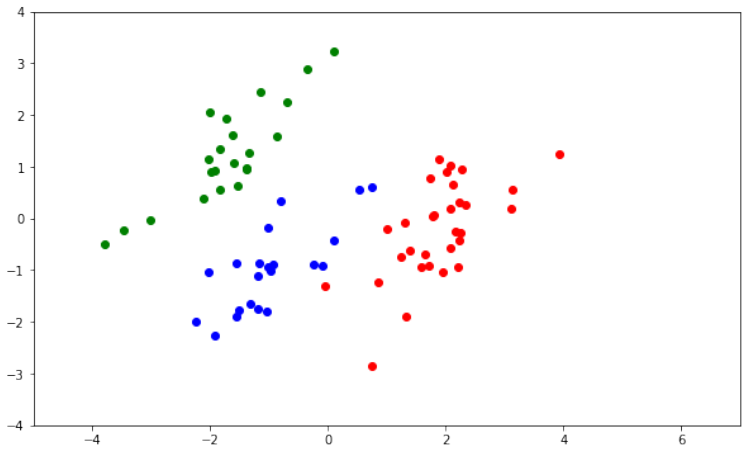
décisionnelle qui, connaissant pour un individu donné les valeurs des X^j, cherche à prédire la classe inconnue de la variable qualitative.

Il faut lire l'AFD comme un préalable à la construction d'une analyse discriminante décisionnelle ou à tout autre méthode ou algorithme de classification supervisée. L'objectif principal est de représenter les qualités discriminatoires des variables quantitatives pour séparer correctement les classes de la variables qualitative. Le graphique des individus permet d'apprécier la qualité de discrimination tandis que, si le nombre de variables n'est pas trop grand, une représentation des variables permet, comme en ACP, d'expliquer la discrimination des classes par les variables quantitatives.

Exemple jouet : les insectes de Lubitsch

Cette méthode est illustrée par une comparaison des sorties graphiques issues d'une ACP et d'une AFD. Les données décrivent trois classes d'insectes sur

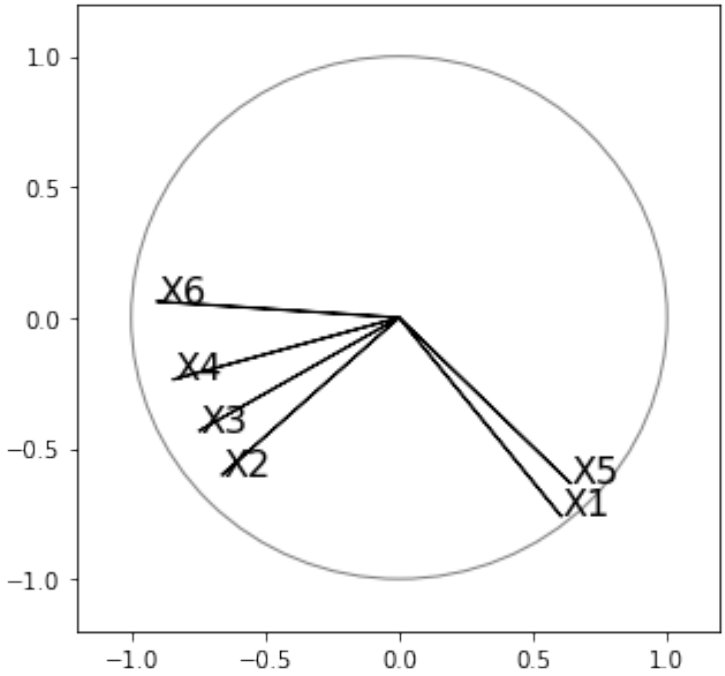
lesquels ont été réalisées 6 mesures anatomiques sur les ailes, élytres, antennes, pattes des insectes. On cherche à savoir si ces mesures permettent de retrouver la typologie de ces insectes. Ce jeu de données académique conduit à une discrimination assez évidente. La comparaison entre l'ACP et l'AFD met clairement en évidence le rôle de la distance de Mahalanobis sur la forme des nuages de chaque classe en analyse discriminante.



Les données académiques sont faciles à étudier, l'ACP ci-dessus montre déjà que les trois nuages d'insectes se distinguent assez bien dans le premier plan principal. Observer la forme identiques des trois nuages ; les trois classes partagent approximativement la même matrice de variance qui est aussi la variance intra-classe. Attention, bien évidemment, si les dispersions à l'intérieur de chaque classe ne sont pas homogènes (hétéroscédasticité), l'effet de la métrique de Mahalanobis ne sera pas aussi flagrant.

Le graphique ci-dessous représente les variables et plus précisément leur

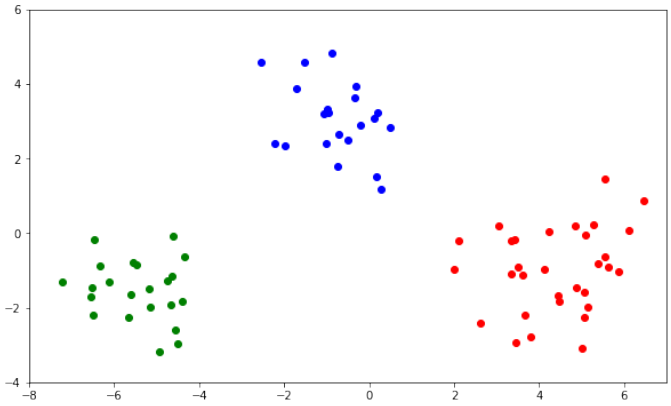
structure de corrélation.



Le cercle présent sur le graphique est appelé *cercle des corrélations*. Il est défini par l'intersection de la boule unité avec le plan de projection des vecteurs variables. Les variables sont centrées et surtout réduites, divisées par leur écart type. Dans ce cas ce sont des vecteurs de longueur 1 donc placés sur la boule unité. Ils se projettent nécessairement tous à l'intérieur du cercle. Le cercle des corrélations est une aide graphique pour apprécier la qualité de représentation de chaque variable. Plus une variable est proche du cercle, meilleure est sa représentation. Dans le cas contraire, le vecteur variable présente un angle grand, proche de 90°, avec le plan de projection et sa représentation est mauvaise ; ne pas en tenir compte dans l'interprétation.

L'AFD produit le graphique ci-dessous. Cette ACP des barycentres des classes sépare encore mieux les classes d'insectes et surtout, le changement de métrique rend sphérique la forme des classes qui sont donc mieux regroupées

autour de leur barycentre.



4 Explorer des données complexes

4.1 ACP de données météo

Besse et al. (2007) considèrent un jeu de données observées dans le contexte de la prévision, pour le lendemain, de la concentration en ozone dans différentes agglomérations afin d'évaluer les risques de dépassement du seuil légal. Le problème peut être considéré comme un cas de régression : la variable à prévoir est une concentration en ozone, mais également comme une discrimination binaire : dépassement ou non du seuil légal. Il n'y a que 8 variables explicatives dont une (`MOCAGE`) est déjà une prévision de concentration d'ozone mais obtenue par un modèle déterministe de mécanique des fluides qui résout les équations de Navier et Stockes. Il s'agit d'un exemple d'*adaptation statistique*. La prévision déterministe sur la base d'un maillage global (30 km) est améliorée localement, à l'échelle d'une ville, par un modèle statistique incluant cette prévision ainsi que des informations connues sur la base d'une grille locale, spatiale et temporelle plus fine :

`JOUR` Le type de jour, férié ou non ;

`O3obs` La concentration d'ozone effectivement observée le lendemain à 17h locales correspondant souvent au maximum de pollution observée ;

`MOCAGE` Prévision de cette pollution obtenue par un modèle déterministe de mécanique des fluides (équation de Navier et Stockes) ;

`TEMPE` Température prévue par MétéoFrance pour le lendemain 17h ;

`RMH2O` Rapport d'humidité ;

`NO2` Concentration en dioxyde d'azote ;

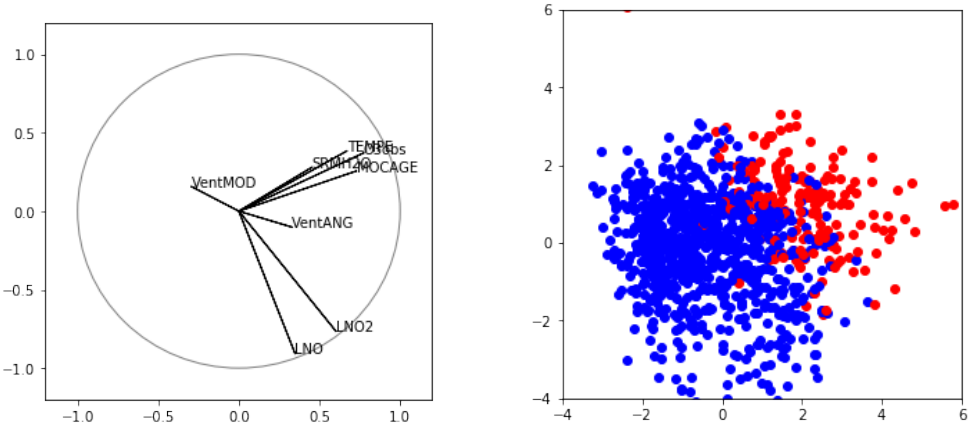
`NO` Concentration en monoxyde d'azote ;

`STATION` Lieu de l'observation : Aix-en-Provence, Rambouillet, Munchhausen, Cadarache et Plan de Cuques ;

`VentMOD` Force du vent ;

`VentANG` Orientation du vent.

L'étude préliminaire rudimentaire a conduit à la transformation (log) de certaines variables de concentration (cf. `ML4IoT-Tutorial-Ozone`). Les données sont résumées par leur représentation dans le premier plan de l'analyse en composantes principales réduite.



Ce graphique résume la structure de corrélation assez intuitive des variables. Un premier groupe (température, mocage, rapport d'humidité) est corrélé et concoure à la concentration en ozone. Deux autres variables liées à la concentration en oxyde d'azote sont également corrélées entre elles et avec le premier axe mais décorrélées avec le premier groupe. Enfin, les variables décrivant le vent sont mal représentées dans le premier plan et liées avec le 3-ème axe. Le graphique des individus à droite met en évidence les difficultés à venir pour discriminer les deux classes respectivement colorées en rouge et bleu : présence ou non d'un pic de concentration avec dépassement du seuil légal. En présence d'une variable qualitative à deux classes, l'AFD qui se réduit à un graphe de dimension 1, défini par l'axe reliant les deux barycentres, n'a que peu d'intérêt.

4.2 ACP et AFD des transformations des signaux

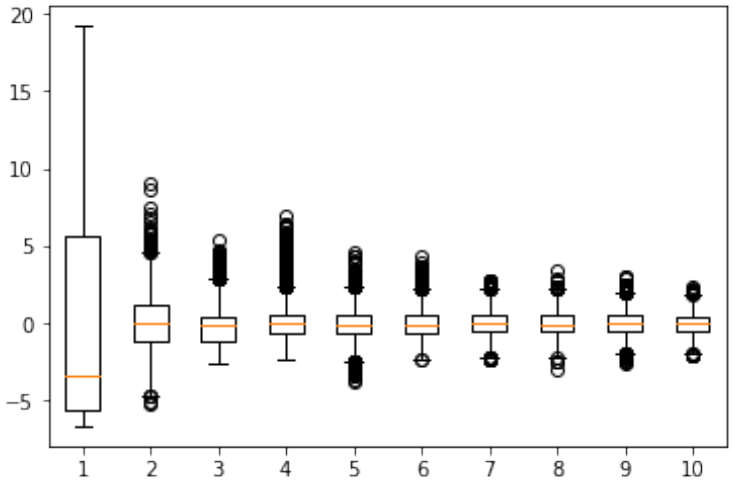
Appliquons les deux méthodes d'exploration multidimensionnelle aux données obtenues après transformation des signaux temporels issus d'un smartphone. D'abord avec l'ACP en considérant seulement les données issues des signaux puis avec l'AFD en ajoutant la variable qualitative correspondant au type d'activité (debout, assis, couché...).

L'ACP prend tout sont sens lorsqu'elle est appliquée à des données de très grande dimension ; $p = 561$ variables ou caractéristiques sont observées ou plutôt calculées à partir des signaux bruts sur $n = 10299$ individus ou plutôt expérimentations d'une activité. Le tutoriel `ML4IoT-UseCase-Har` réalise tous les

calculs en python pour représenter l'analyse en composantes principales réduites puis l'AFD de ces données.

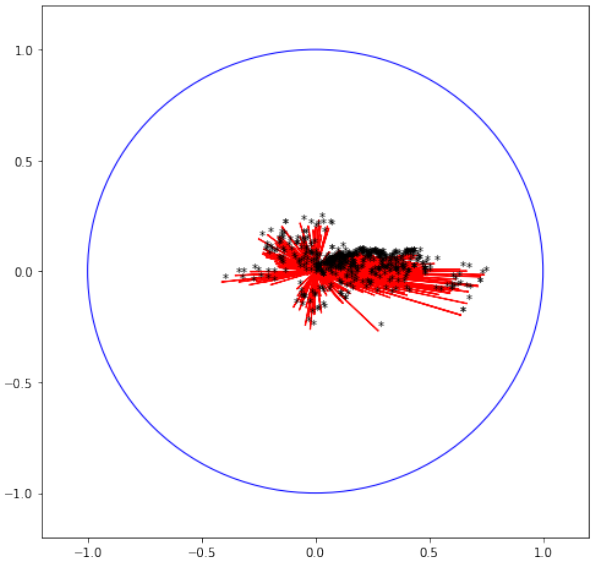
Décroissance des valeurs propres

Une première information est fournie par la décroissance des valeurs propres c'est-à-dire par la décroissance des variances de chaque variable principale. Un graphique spécifique précise les choses en opérant des diagrammes boîtes parallèles des composantes principales.



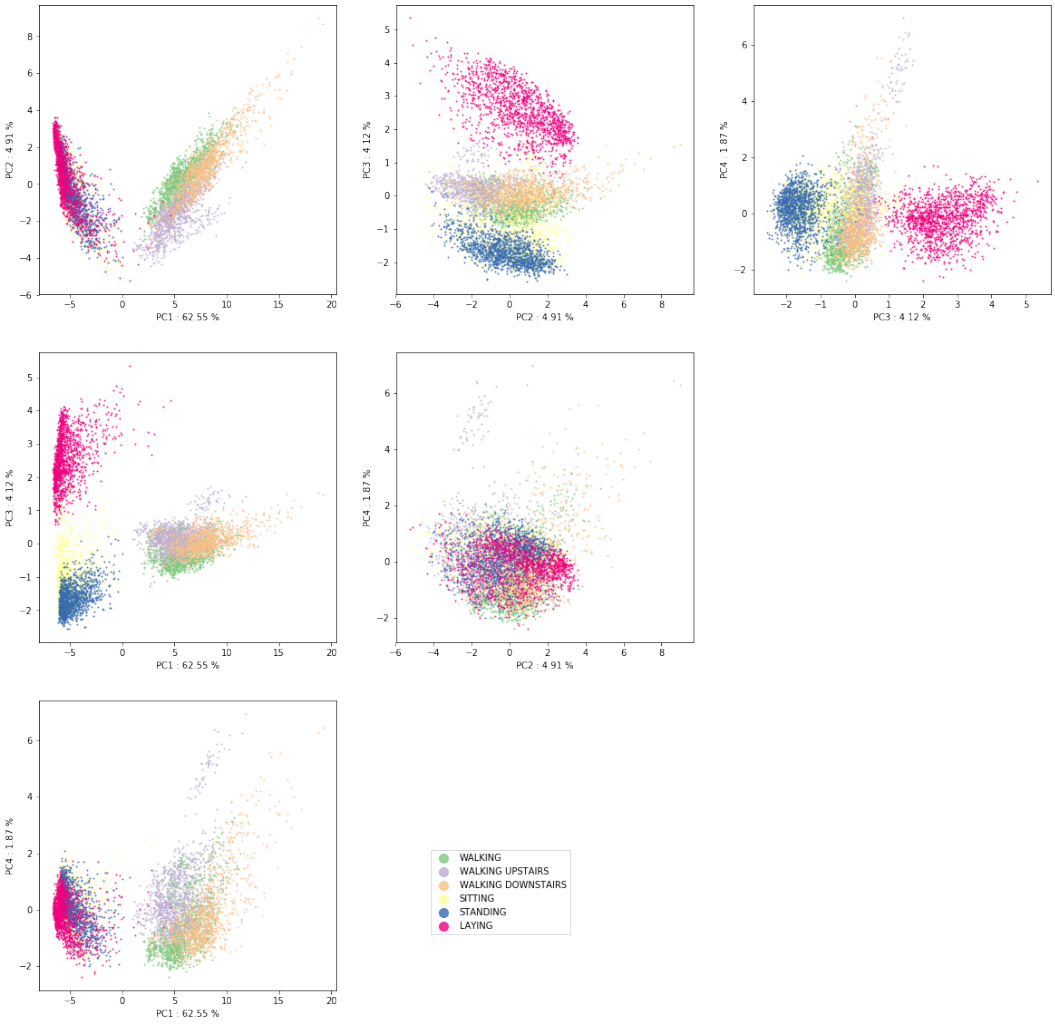
De gauche à droite, chaque diagramme boite représente la distribution de chaque composante principale, ici, les 10 premières des 561 composantes. La première composante de plus grande variance est associée à une grande boîte accompagnée d'une grande moustache puis, la variance décroissant, les boîtes deviennent de plus en plus petites. Noter la présence de nombreuses valeurs atypiques. Nous nous limiterons aux trois premières composantes, considérant ensuite que les composantes de boîte trop petite ne sont associées qu'à du bruit sans information pour distinguer les activités.

Représentation des variables



Lorsque le nombre p de variables est raisonnable, la représentation de celles-ci dans le premier plan factoriel aide à comprendre la structure de corrélation. Avec p trop grand cette représentation est inexploitable.

Représentation des individus



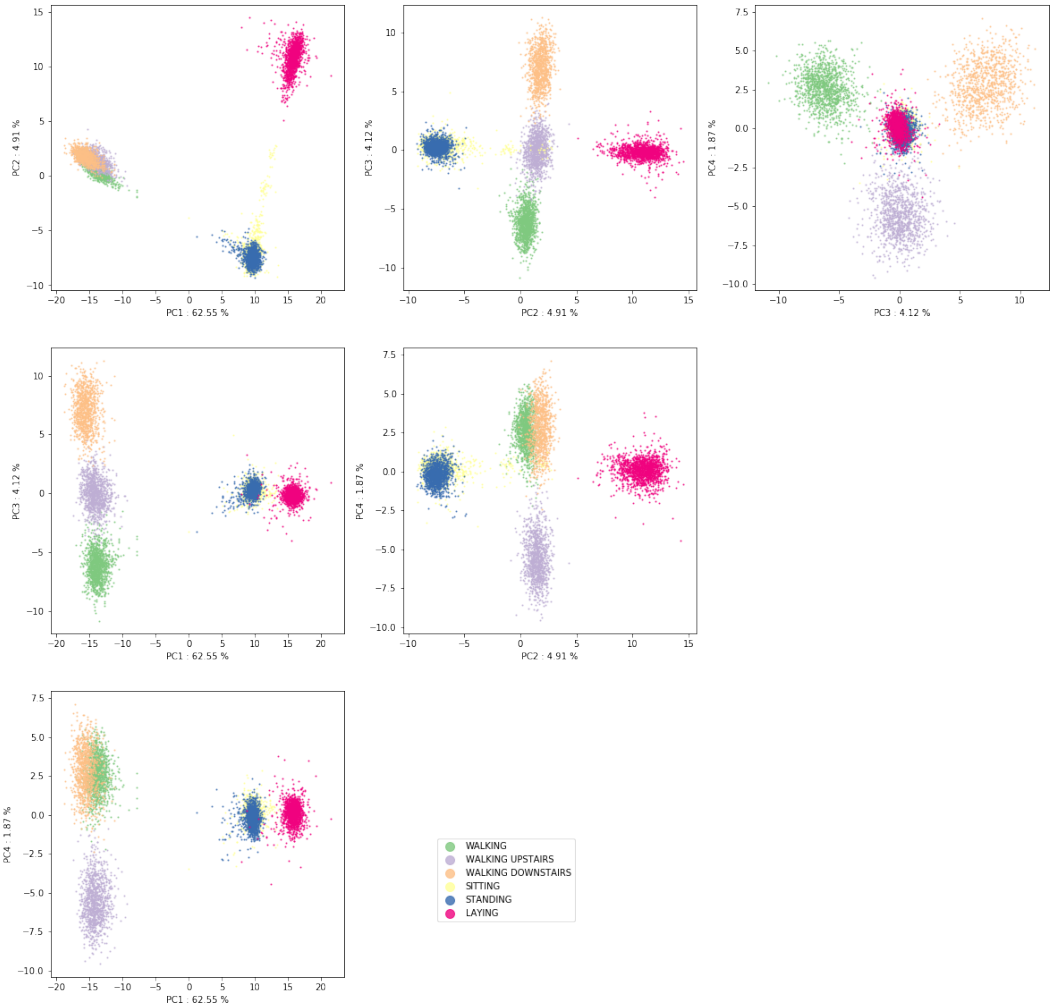
En revanche il est très informatif de représenter chacune des 6 activités dans les trois premiers axes factoriels en associant une couleur à chaque activité. Il apparaît que, même sans savoir qu'il y a 6 types de données, les activités se distinguent assez naturellement. Plus précisément, deux groupes d'activités : dynamique ou statique, sont clairement distingués expliquant ainsi la grande variance du premier axe.

Néanmoins, il apparaît comme difficile de séparer, discriminer nettement certains couples d'activités.

Attention, l'ACP se comporte comme si les activités ne sont pas connues,

c'est une approche non supervisée. Comme les activités sont *a priori* connues, il est possible de prendre en compte cette connaissance dans une approche plus adaptée que l'ACP. Il s'agit de l'analyse factorielle discriminante (AFD) sommairement introduite dans la section précédente.

4.3 AFD des transformations des signaux



Représentation de chacun des 6 types d'activités dans les premiers plans factoriels de l'AFD. La couleur d'un point est fonction de l'activité connue. L'ACP des 6 barycentres des classes en utilisant la métrique de Mahalanobis (matrice inverse de la variance intra-classe) a pour effet d'encore mieux séparer

les 6 classes. En considérant les trois premiers axes ou plutôt les projections sur les plans définis par ces trois premiers axes, il apparaît que les classes d'activités sont bien séparées dans cet espace à 3 dimensions. À l'exception des deux classes : assis, couché, plus difficiles à séparer, il existe un plan de projection dans lequel une classe se distingue des autres. Autrement dit il existe des hyperplans ou séparations linéaires des classes.

La construction de ces hyperplans, frontières entre les classes, est l'objectif implicite de la prochaine partie. Leur connaissance conduit à des règles de décision pour la prévision d'une activité dont les signaux, ou plutôt les caractéristiques, sont connues mais pas la classe.

Que faut-il retenir ?

Les signaux temporels bruts sont très confus notamment à cause de leurs décalages temporels ou déphasages. Ils sont difficiles à caractériser en l'état et sont pour le moment laissés de côté. En revanche les transformations de ces données par des techniques de traitement du signal dont leur décomposition de Fourier présentent des caractéristiques plus encourageantes.

Ce sont 561 variables qui sont calculées sur les 9 courbes enregistrées par chaque activité des porteurs d'un smartphone. Nous sommes donc toujours dans une situation de grande dimension dont l'exploration nécessite des méthodes appropriées. L'analyse en composantes principales permet de représenter le nuage des activités en seulement 3 dimensions avec des séparations encourageantes des classes d'activités sans les connaître a priori ; 561 variables ou caractéristiques sont donc résumées de façon satisfaisante par l'ACP avec seulement trois nouvelles variables ou variables principales, combinaisons linéaires de celles initiales.

Néanmoins, le nombre trop important de ces variables ne permet pas d'obtenir simplement et directement une interprétation utile des variables principales.

Comme nous sommes dans un contexte supervisé car les classes d'activités sont connues, c'est l'analyse factorielle discriminante qui est appropriée. Cette ACP particulière calculée sur les 6 barycentres des classes en utilisant comme métrique la matrice inverse de la variance intra-classe conduit, toujours en dimension 3 à de très bonne séparations des classes. A l'exception des deux activités : assis et couché, il semble bien que toutes les autres peuvent être séparées par des hyperplans. En d'autres termes, une méthode ou un algorithme linéaire de classification supervisée devrait permettre d'atteindre l'objectif de reconnaissance de l'activité.

5 Classification non supervisée ou *clustering*

5.1 Objectif

Ce cours met particulièrement l'accent sur les méthodes d'apprentissage supervisée mais, à titre d'exemple, voici un type d'algorithme d'apprentissage ou classification non supervisée dont l'objectif déborde du cadre strictement exploratoire. Ces algorithmes visent la recherche d'une *typologie*, ou *segmentation*,

c'est-à-dire d'une partition, ou répartition des individus en *classes* homogènes, ou catégories. Ceci est fait en optimisant un *critère* visant à regrouper les individus dans des classes, chacune le plus homogène possible et, entre elles, les plus distinctes possible. Cet objectif est à distinguer des procédures de discrimination, ou encore de classement (en anglais *classification*) pour lesquelles une typologie est *a priori* connue, au moins pour un échantillon d'apprentissage. Nous sommes dans une situation d'apprentissage *non-supervisé*, ou en anglais de *clustering*[1].

5.2 Les méthodes

Un calcul de combinatoire montre que le nombre de partitions possibles d'un ensemble de n éléments croît exponentiellement avec n ; le nombre de partitions de n éléments en k classes est le nombre de Stirling, le nombre total de partitions est celui de Bell. Pour $n = 20$ il est de l'ordre de 10^13. Il n'est donc pas question de chercher à optimiser le critère sur toutes les partitions possibles. Les méthodes se limitent à l'exécution d'un *algorithme itératif* convergeant vers une bonne partition et correspondant en général à un optimum local.

Plusieurs choix sont laissés à l'initiative de l'utilisateur :

- une mesure d'éloignement (dissemblance, dissimilarité ou distance) entre individus ;
- le critère d'homogénéité des classes à optimiser : il est, dans le cas de variables quantitatives, généralement défini à partir de la trace d'une matrice de variances-covariances ; soit les variances et covariances interclasses (la trace correspond alors à l'inertie de la partition), soit les variances et covariances intraclasses ;
- la méthode : classification ascendante hiérarchique, ré-allocation dynamique, mélanges gaussiens et DBSCAN sont les plus utilisées, seules ou combinées ;
- le nombre de classes qui est un point délicat.

Enfin, différents outils tentent d'évaluer la qualité des classes obtenues ou recherchent une interprétation, ou des caractérisations, des classes.

Seul un algorithme de type ré-allocation dynamique ou encore par agrégation autour de centres mobiles : *k-means* est décrit ci-après.

5.3 Agrégation autour de centres mobiles

Principes

Différents types d'algorithmes ont été définis autour du même principe de *ré-allocation dynamique* des individus à des centres de classes, eux-mêmes recalculés à chaque itération. Ces algorithmes sont facilement définis lorsque une représentation vectorielle des individus est disponible dans $\mathbb{R}^p$ muni d'une métrique, généralement euclidienne. Il est important de noter que, contrairement

1. Faire attention aux faux amis français / anglais : discrimination / *classification* (supervisée) et classification / *clustering* (non-supervisée)

à la méthode dite *hiérarchique ascendante*, le nombre de classes k doit être déterminé *a priori*. En revanche un très gros avantage des ces algorithmes est leur faible niveau de complexité, leur permettant de traiter des volumes de données important et même de passer à l'échelle lorsque les données massives sont distribuées sur plusieurs ordinateurs.

Ces méthodes sont itératives : après une initialisation des centres consistant, par exemple, à tirer aléatoirement k individus, l'algorithme répète deux opérations jusqu'à la convergence d'un critère :

1. Chaque individu est affecté à la *classe* dont le centre est le plus proche au sens d'une métrique.
2. Calcul des k *centres* des classes ainsi constituées.

Principale méthode

Il s'agit de la version proposée par Forgy (1965) des algorithmes de type *k-means*.

Algorithm 1 Algorithme de Forgy

Initialisation Tirer au hasard, ou sélectionner pour des raisons extérieures à la méthode, k points dans l'espace des individus, en général k individus de l'ensemble, appelés centres ou noyaux.

repeat

Allouer chaque individu au centre (c'est-à-dire à la classe) le plus proche au sens de la métrique euclidienne choisie ; on obtient ainsi, à chaque étape, une classification en k classes, ou moins si, finalement, une des classes devient vide.

Calculer le centre de gravité de chaque classe : il devient le nouveau noyau ; si une classe s'est vidée, on peut éventuellement retirer aléatoirement un noyau complémentaire.

until Le critère de variance interclasses ne croisse plus de manière significative, c'est-à-dire jusqu'à la stabilisation des classes.

Propriétés

Convergence Le critère (la variance interclasses) est majoré par la variance totale. Il est simple de montrer qu'il ne peut que croître à chaque étape de l'algorithme, ce qui en assure la convergence. Il est équivalent de maximiser la variance interclasses ou de minimiser la variance intraclasse. Cette dernière est alors décroissante et minorée par 0. Concrètement, une dizaine d'itérations suffit généralement pour atteindre la convergence.

Optimum local La solution obtenue est un optimum local, c'est-à-dire que la répartition en classes dépend du choix initial des noyaux. Plusieurs exécutions de l'algorithme permettent de s'assurer de la présence de *formes fortes*, c'est-à-dire de classes, ou partie de classes, présentes de manière stable dans la majorité des partitions obtenues.

Toujours sous la même appellation (une option de la commande `kmeans` de R) Mac Queen (1967) a proposé une modification de l'algorithme précédent. Les noyaux des classes, ici les barycentres des classes concernées, sont recalculés à chaque allocation d'un individu à une classe. L'algorithme est ainsi plus efficace, mais la solution dépend de l'ordre des individus dans le fichier.

5.4 Exemple : segmentation d'une image

Données et objectif

Le calepin `CSdD-Cluster-Mars-Python` déroule l'analyse d'une image en utilisant les capacités des librairies Python. Les données sont constituées d'une image hyperspectrale de la surface de Mars. L'imagerie visible et en proche infrarouge est une technique clef de télédétection pour étudier le système planétaire à l'aide de spectromètres embarqués sur des satellites. En mars 20914, l'équipement OMEGA (Mars Express, ESA, Bibring et. al. 2005) a collecté 310 GO d'images brutes. Il a cartographié la surface de Mars avec une résolution variant entre 300 et 3000 m fonction de l'altitude du véhicule spatial. Il a acquis, pour chaque pixel, la réponse spectrale présente entre 0.36 et 5.2 μm et échantillonnées dans 255 canaux. L'objectif est de caractériser la composition matérielle de la surface et en particulier de distinguer différentes classes de silicates, minéraux, oxydes, carbonates et parties gelées.

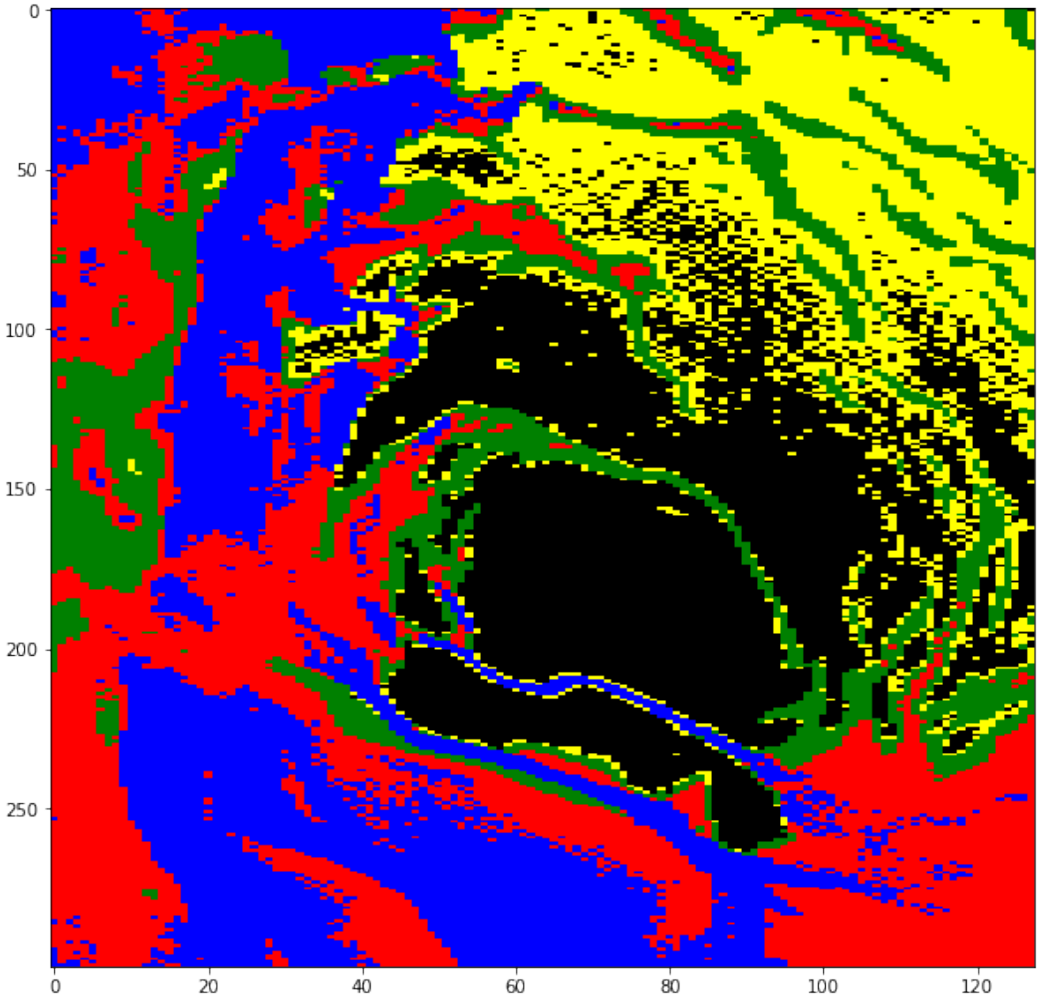
Ceci est illustré par l'analyse d'une image 300×128 image. A chacun de ces 38 400 pixels, individus, instances ou unités statistiques est associée un vecteur de 255 valeurs spectrales ou variables, caractéristiques ou *features*.

Selon les experts, il y a $K = 5$ classes minéralogiques à identifier sur la carte. L'objectif est donc d'opérer une classification non supervisée des pixels conduisant à une segmentation de l'image en 5 types ou (fausses) couleurs des pixels. Avant d'opérer la segmentation, une approche exploratoire permet de se familiariser avec ce type particulier de données.

Résultats

Le calepin `CSdD-Cluster-Mars-Python` compare plusieurs stratégies après ACP ou non et algorithmes sur ces données mais La simple utilisation de l'algorithme *k-means* conduit à des résultats déjà satisfaisants. Ceci est illustré par l'image ci-dessous. Les fausses couleurs représentent les classes 5 obtenues, classes qu'il faudrait ensuite associé à des composants chimiques en analysant

les spectres moyens de chaque classe.



Références

Anguita D., Ghio A., Oneto L., Parra X., Reyes-Ortiz J.L. (2013). A Public Domain Dataset for Human Activity Recognition Using Smartphones, *ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, 437-442.

Besse P., Milhem H., Mestre O., Dufour A., Peuch V.-H. (2007). Comparaison de techniques de Data Mining pour l'adaptation statistique des prévi-

sions d'ozone du modèle de chimie-transport MOCAGE, *Pollution Atmosphérique*, 195, 285-292.
Bibring J.P. et al.(2005). Mars Surface Diversity as Revealed by the OMEGA/Mars Express Observations, *Science*, Vol. 307, Issue 5715, 1576-1581.
Forgy R. (1965). Cluster Analysis of Multivariate Data : Efficiency versus Interpretability of Classification, *Biometrics* , 21, 768 ?769
Macqueen J.(1967). Some methods for classification and analysis of multivariate observations, In 5-th Berkeley Symposium on Mathematical Statistics and Probability, 281 ?297.