Analyses statistiques exploratoires de données en éducation : principes, concepts et implémentation - Le cas de l'adaptation des primo-entrants en IUT après la réforme du baccalauréat et du BUT

Exploratory statistical analyses of educational data: principles, concepts and implementation – the case of first-year students' adaptation in IUT after the reform of the baccalauréat and the BUT

Análises estatísticas exploratórias de dados em educação: princípios, conceitos e implementação – o caso da adaptação dos ingressantes nos IUT após a reforma do baccalauréat e do BUT

Patrick Pamphile
ID ORCID : 0000-0002-4560-9069
Université Paris-Saclay

Isabelle Bournaud
ID ORCID : 0000-0002-9819-2789
Université Paris-Saclay

Mots clés : analyse factorielle exploratoire, classification non supervisée, primo-entrants, structures cachées, transition lycée-université

À l'ère de la massification des données, la disponibilité croissante d'informations numériques constitue une tendance majeure, y compris en sciences de l'éducation. Les technologies utilisées dans les enseignements en ligne permettent désormais d'enregistrer l'ensemble des activités des étudiants. Cette accumulation de données offre indéniablement une richesse d'informations, mais rend les données complexes et difficiles à analyser. Les méthodes statistiques exploratoires sont utiles pour identifier des liaisons latentes entre les variables ou entre les individus. Cependant, leur utilisation nécessite une compréhension des concepts statistiques mobilisés pour garantir une implémentation appropriée. Cet article propose de clarifier les concepts statistiques sous-jacents et de guider, étape par étape, les praticiens dans leurs analyses statistiques exploratoires. Nous montrons, à partir d'une étude menée auprès d'étudiants primo-entrants en IUT, de quelle manière ces méthodes peuvent contribuer à une compréhension holistique de leurs difficultés d'adaptation en IUT.

Key words: clustering, exploratory factorial analysis, first-year students, hidden structures, high school-university transition

In the era of big data, the increasing accessibility of digital information is a significant trend, including in the field of educational sciences. Technologies used in online teaching now enable all student activities to be recorded. While this accumulation of data undeniably provides a wealth of information, it also makes the data more complex and difficult to analyze. Exploratory statistical methods are useful for identifying latent relationships between variables or individuals. However, their use requires an understanding of the statistical concepts mobilized to ensure appropriate implementation. This article proposes to clarify the underlying statistical concepts and to guide practitioners step-by-step in their exploratory statistical analyses. Based on a study of first-time IUT students, we show how these methods can contribute to a holistic understanding of their difficulties in adapting to IUT.

Note des auteurs : les auteurs tiennent à remercier les examinateurs pour leur précieux commentaires et regards critiques qui ont grandement contribué à éclaircir et à enrichir ce travail. La correspondance liée à cet article peut être adressée à patrick.pamphile@universite-paris-saclay.fr.

Palavras chave: análise fatorial exploratória, classificação não supervisionada, estruturas ocultas, ingressantes, transição ensino médio-universidade

Na era da massificação dos dados, a crescente disponibilidade de informações digitais constitui uma tendência importante, inclusive no âmbito das ciências da educação. As tecnologias utilizadas no ensino online permitem agora registar todas as atividades dos estudantes. Esta acumulação de dados oferece, sem dúvida, uma riqueza de informações, mas também torna os dados complexos e difíceis de analisar. Os métodos estatísticos exploratórios são úteis para identificar conexões latentes entre variáveis ou entre pessoas. No entanto, a sua aplicação requer a compreensão dos conceitos estatísticos envolvidos, a fim de garantir uma implementação apropriada. Este artigo propõe esclarecer os conceitos estatísticos subjacentes e guiar, passo a passo, os profissionais nas suas análises estatísticas exploratórias. A partir de um estudo realizado com estudantes ingressantes num IUT, mostramos como estes métodos podem contribuir para uma compreensão holística das dificuldades de adaptação enfrentadas nesse contexto.

Introduction

À l'ère de la massification des données, la disponibilité croissante d'informations numériques constitue une tendance majeure, y compris en sciences de l'éducation. Les technologies mobilisées dans les enseignements en ligne permettent non seulement d'enregistrer l'ensemble des activités des étudiants, mais aussi de collecter de manière systématique des traces et des rétroactions exploitables à des fins d'analyse. Parallèlement, la transformation numérique permet de recueillir les données des évaluations en ligne des enseignements et des formations, mais également les données administratives des étudiants tout au long de leur cursus de formation. Cette accumulation de données, recueillies sur plusieurs cohortes, offre indéniablement une richesse d'informations pour les équipes pédagogiques et pour les chercheurs. Toutefois, la visualisation de ces données massives est difficile. De plus, il n'est pas aisé d'en saisir rapidement des tendances, des motifs ou des anomalies. Par ailleurs, plus les données comportent de variables et d'individus, plus elles sont susceptibles de renfermer des interactions complexes entre variables et une forte hétérogénéité parmi les individus. Autrement dit, la nature complexe des données massives constitue un obstacle majeur à leur analyse. Cette dernière est d'autant

plus ardue que les interactions entre les variables et l'hétérogénéité entre les individus sont souvent cachées, alors qu'elles peuvent jouer un rôle significatif dans l'étude d'un phénomène. Les méthodes statistiques exploratoires sont utiles pour analyser la structure latente de telles données, que ce soit en matière de liaisons entre les variables (analyse factorielle exploratoire, analyse en composantes principales, …) ou en matière de regroupements pertinents d'individus (classification ascendante hiérarchique, algorithme des k-means, …). Leur usage facilite la compréhension des phénomènes sous-jacents étudiés : par exemple, en sciences de l'éducation, elles peuvent être utilisées pour identifier des facteurs explicatifs de la réussite académique ou des profils caractéristiques d'apprenants.

Si les analyses statistiques exploratoires sont adaptées pour étudier la structure latente de données massives, leur utilisation nécessite toutefois une expertise afin d'en éviter les écueils et d'obtenir des analyses fiables et pertinentes. L'objectif de cet article est de clarifier les concepts statistiques sous-jacents aux analyses statistiques exploratoires et de présenter, étape par étape, les choix méthodologiques à effectuer lors de leur mise en œuvre. Nous illustrerons nos propos sur une étude menée sur l'adaptation des primo-entrants à l'université. L'identification des facteurs explicatifs de l'adaptation et des profils d'adaptation des primo-entrants offre une approche holistique des données, en analysant à la fois les variables et les individus, ce qui permet de mieux comprendre les divers aspects et la complexité des difficultés d'adaptation. Les équipes pédagogiques peuvent alors ajuster ou proposer des activités pédagogiques afin de répondre au mieux aux besoins spécifiques des différents profils d'apprenants.

En France, la réforme du Baccalauréat général en 2019 et celle du Bachelor Universitaire de Technologie (BUT)[1] en 2021 ont engendré des changements dans les profils des lycéens intégrant les instituts universitaires de technologie (IUT), mais également des modifications dans les contenus des enseignements et dans les modalités d'évaluation des apprentissages en BUT. Ces divers changements et leurs interactions ne sont pas sans conséquences sur les difficultés d'adaptation rencontrées par les primo-entrants. Dans cette situation, il est apparu crucial d'appréhender les interactions complexes entre les différentes variables jouant un rôle dans l'adaptation à l'IUT et d'identifier différents profils d'adaptation parmi les

1. Le BUT est le diplôme national professionnalisation de niveau Bac +3 préparé dans les IUT qui sont des composantes des universités françaises.

primo-entrants. L'objectif pour les formations est de mieux appréhender les facteurs clés qui facilitent leur adaptation et ainsi d'améliorer leur expérience étudiante. Nous avons utilisé des analyses statistiques exploratoires afin d'étudier le phénomène d'adaptation des primo-entrants en IUT.

Dans la section suivante, nous présentons des analyses factorielles adaptées à l'analyse exploratoire de la structure des variables observées, en rappelant leurs objectifs, les concepts associés et les différentes étapes de leur mise en œuvre. La troisième partie aborde les différentes méthodes de classification non supervisées des individus, et précise leurs avantages et inconvénients respectifs. Dans la quatrième partie, nous illustrons la mise en œuvre de ces méthodes sur des données issues d'une étude menée sur l'adaptation de primo-entrants en IUT. Nous terminons par une réflexion sur les apports potentiels des analyses exploratoires pour l'étude de données multivariées en éducation.

Les méthodes exploratoires de données multivariées

Le concept de variabilité est fondamental en statistiques, car il fournit des informations sur l'homogénéité ou l'hétérogénéité des données. Si, par exemple, on observe p items d'un questionnaire administré auprès de n étudiants, les variations des résultats peuvent être dues soit aux liaisons entre les réponses aux items (par exemple, une question sur l'adaptation à l'allongement du temps de transport pour se rendre sur le lieu d'étude peut être fortement liée à une question sur l'adaptation académique), soit aux différences entre individus (par exemple, deux étudiants peuvent, selon leur parcours et leurs pratiques d'étude, répondre différemment à une question sur l'adaptation académique). Les variations sont donc source d'information : elles mettent en évidence des liens entre les variables observées sur les individus ou des différences ou des similarités entre les individus. On distingue alors deux grandes méthodes d'analyse exploratoire des données : les méthodes factorielles, dont l'objectif est d'explorer les liaisons entre des variables, et les méthodes de classification non supervisée, dont l'objectif est d'explorer la similarité entre des individus (Escoffier & Pagès, 1998 ; Lebart et al., 1995).

Les méthodes factorielles exploratoires

L'objectif principal des méthodes factorielles exploratoires est de faciliter l'analyse des données en fournissant une représentation simplifiée et structurée des variables (Escoffier & Pagès, 1998 ; Fabrigar & Wegener,

2011 ; Lebart et al., 1995 ; Mulaik, 2009). Elles visent à résumer la structure sous-jacente des liaisons entre les variables à l'aide d'un nombre restreint de variables appelées facteurs, qui expliquent les liaisons entre les variables observées. Les méthodes d'analyse factorielle exploratoire se différencient selon la nature des variables concernées : quantitatives ou qualitatives (Lebart et al., 1995).

Nous avons opté pour la présentation de deux méthodes d'analyse : l'analyse en composantes principales (ACP) et l'analyse factorielle exploratoire (AFE). Ces méthodes reposent sur le concept de corrélation et sont naturellement adaptées aux variables quantitatives. En sciences de l'éducation, les échelles de mesure majoritairement utilisées sont de type échelle de Likert, qui permettent de capturer des nuances d'attitude, d'opinion et de perception auprès d'individus. Ces échelles génèrent des données qualitatives ordinales qui, certes, présentent moins de variabilité que les variables quantitatives, mais qui permettent néanmoins le calcul de corrélations de Spearman et donc, l'utilisation de l'ACP et de l'AFE (Carifio & Perla, 2007 ; Muliak, 2009 ; Saporta, 2006). Il convient cependant de tenir compte des spécificités des échelles de Likert lors de l'implémentation de l'AFE pour de telles variables.

Précisons que nous ne nous intéressons pas dans cet article au cas des variables qualitatives non ordinales, pour lesquelles l'analyse des correspondances multiples (ACM) est la méthode d'analyse exploratoire factorielle adaptée. Les personnes souhaitant en savoir plus sur l'ACM et sur sa mise en œuvre peuvent consulter les ouvrages de Lebart et al. (1995) et de Saporta (2006).

L'analyse en composantes principales (ACP)

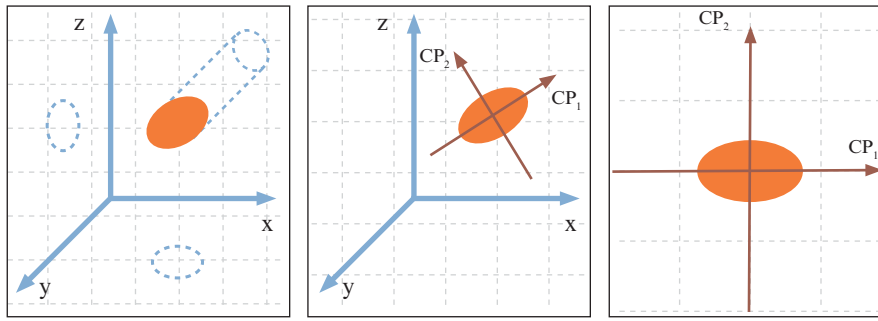
Le principal objectif de l'ACP est de réduire la dimensionnalité des données en conservant un maximum d'information (Lebart et al., 1995). Pour ce faire, l'ACP génère p variables (CP_i) non corrélées, chacune étant une combinaison linéaire des variables observées. Ces variables sont appelées composantes principales. Le modèle factoriel de l'ACP est le suivant :

$$\begin{aligned}
Y_1 &= a_{11}CP_1 + \cdots + a_{1p}CP_p \\
&\vdots \qquad\qquad \vdots \\
Y_i &= a_{i1}CP_1 + \cdots + a_{ip}CP_p \\
&\vdots \qquad\qquad \vdots \\
Y_p &= a_{p1}CP_1 + \cdots + a_{pp}CP_p
\end{aligned}$$

Les composantes principales sont construites de manière itérative en recherchant des composantes qui ne sont pas corrélées entre elles et qui maximisent la variance expliquée des données à chaque étape (voir figure 1, ci-après). En pratique, les composantes principales sont obtenues en diagonalisant la matrice de corrélations : les valeurs propres correspondent à la proportion de variance globale expliquée par les composantes principales, et les vecteurs propres, aux poids permettant de construire les composantes comme combinaisons linéaires des variables. Nous obtenons ainsi une décomposition de la variance globale des données en une somme de contributions attribuables à chaque composante principale, et ordonnées de façon décroissante. Cette décomposition présente plusieurs intérêts :

- **Réduire la complexité des données**. L'un des premiers objectifs de l'analyse exploratoire est de réduire la complexité des données multivariées. En projetant le nuage de n individus et de p variables sur les k premières composantes principales ($k<p$), nous passons d'un espace de données dans R^p à un espace de dimension réduite R^k, tout en conservant le plus de variance possible. Cela permet de simplifier l'analyse tout en conservant un maximum d'informations pertinentes. Plusieurs critères existent pour sélectionner un nombre optimal k de composantes principales (Lebart et al., 1995). Le principe général de ces critères est de limiter le nombre de composantes tout en conservant le maximum de variance expliquée. Nous illustrons la problématique de sélection du nombre de composantes sur l'exemple de l'adaptation des primo-entrants en IUT présenté dans la quatrième partie.
- **Filtrer le bruit d'échantillonnage**. Le bruit lié à l'échantillon correspond à des fluctuations purement aléatoires, telles les erreurs de mesure, et qui n'apportent pas d'information pertinente sur le phénomène étudié. Les composantes expliquant moins de 5% de la variance totale peuvent être associées au bruit lié à l'échantillon (Husson et al., 2011 ; Lebart et al., 1995). Se limiter aux premières composantes principales expliquant 95% de la variance totale rend ainsi les analyses plus robustes au bruit d'échantillonnage.
- **Faciliter la visualisation des données**. Le nuage des n individus est dans R^p. L'ACP permet de visualiser ce nuage dans un plan défini par deux composantes, tout en minimisant les distorsions, comme le montre la figure 1. Cette visualisation est très utile pour repérer d'éventuelles données aberrantes ou des groupes d'individus qui présentent des similarités latentes.

Figure 1
Principe de l'ACP : réduction de dimension par projection sur les composantes principales



a) Nuage de points de $\mathbb{R}^3$

b) Construction des deux premières composantes, indépendantes et maximisant la variance expliquée

c) Projection du nuage sur les deux premières composantes principales

L'analyse factorielle multiple (AFM)

Si l'ACP permet de réduire la dimensionnalité des données en analysant les corrélations entre les variables et de visualiser le nuage des individus dans des plans factoriels, dans le cas de l'existence *a priori* de groupes de variables, il peut être intéressant d'analyser les corrélations intra-groupes et les corrélations inter-groupes, ou encore d'étudier les individus à travers les groupes de variables. Par exemple, si dans un questionnaire plusieurs questions portent sur le même thème, alors les questions d'un thème constituent un groupe de variables. Si un questionnaire identique est administré à des groupes différents d'individus, alors les groupes de variables correspondent aux réponses apportées par chaque groupe d'individus.

Pour cela, on utilise une analyse factorielle multiple (AFM). L'AFM procède en trois étapes. Dans un premier temps, des ACP partielles sont réalisées par groupe de variables. Dans un second temps, les variables de chaque groupe sont normées par la plus grande valeur propre de l'ACP partielle du groupe. Ainsi, chaque groupe de variables possède le même poids dans l'analyse de la variance. En effet, si un groupe est composé de dix questions et un autre de deux questions, alors le groupe de dix questions aura, de manière artificielle, une contribution à la construction

des composantes supérieure à celle du groupe de deux variables. La normalisation élimine ce problème. Enfin, dans un troisième temps, une ACP globale est effectuée sur l'ensemble des variables normées.

L'AFM permet ainsi d'analyser les corrélations intra-groupes et les corrélations intergroupes. À l'instar de l'ACP, l'AFM permet de réduire la dimension du phénomène étudié. Enfin, elle permet de visualiser le nuage des individus tout en étudiant leurs similarités et leurs différences à travers les groupes de variables.

L'analyse factorielle exploratoire (AFE)

Pour le praticien menant une analyse de données multivariées, l'un des défis est de pouvoir distinguer l'information propre à une variable de l'information partagée avec d'autres variables. Pour cela, il est possible de décomposer la variance d'une variable en une part de variance partagée avec d'autres variables et une part unique (Fabrigar & Wegener, 2011) :

variance observée = variance commune + variance unique

où la variance commune est la variance partagée entre un groupe de variables observées que l'on cherche à expliquer par des facteurs communs, et la variance unique est spécifique à la variable observée, c'est-à-dire non partagée avec d'autres. À l'instar de l'ACP, l'AFE facilite donc la compréhension de la structure intrinsèque des données en réduisant la complexité des informations à un nombre réduit de variables, les facteurs. Toutefois, le modèle factoriel est plus complexe que celui de l'ACP (équation 2) :

$$\begin{aligned} Y_1 &= a_{11}F_1 + \cdots + a_{1k}F_k + U_1 \\ &\vdots \qquad\qquad \vdots \qquad\qquad \vdots \\ Y_i &= a_{i1}F_1 + \cdots + a_{ik}F_k + U_i \\ &\vdots \qquad\qquad \vdots \qquad\qquad \vdots \\ Y_p &= a_{p1}F_1 + \cdots + a_{pk}F_k + U_p \end{aligned}$$

Dans le modèle factoriel de l'AFE, la variance commune d'une variable est égale à la somme des carrés des coefficients de saturation a_{ij}. Un coefficient de saturation élevé (en valeur absolue) signifie que le facteur associé contribue significativement à la variance commune. Les facteurs correspondent alors à des dimensions sous-jacentes au phénomène étudié qui expliquent pourquoi certaines variables sont corrélées entre elles. Cette approche facilite la compréhension de la structure intrinsèque des

données en réduisant la complexité des informations à un nombre limité de facteurs, à condition que le praticien puisse interpréter ces derniers dans le contexte de l'étude.

L'analyse de la structure des variables : l'AFE

Les principaux objectifs de l'analyse exploratoire sont de résumer de façon concise les relations entre les variables et de faciliter l'interprétation des données en révélant les axes principaux qui structurent le nuage de points des individus. L'ACP et l'AFE sont ainsi toutes les deux des méthodes exploratoires permettant de réduire la dimensionnalité des données à l'aide des composantes principales pour l'ACP et des facteurs communs pour l'AFE. Toutefois, en raison de leurs objectifs distincts, ces méthodes présentent des différences notables :

- Il est possible que les facteurs de l'AFE soient corrélés entre eux, contrairement aux composantes principales de l'ACP qui sont conçues pour être mutuellement indépendantes. Or, l'hypothèse d'indépendance totale entre les dimensions sous-jacentes des données n'est pas toujours réaliste.
- L'AFE permet de distinguer la variance commune (partagée entre les variables via les facteurs) et la variance spécifique (unique à chaque variable) et ainsi, de mieux comprendre quelle part de la performance ou des attitudes mesurées peut être attribuée à des facteurs communs sous-jacents par rapport à des influences uniques.

Pour le praticien souhaitant entreprendre une analyse exploratoire de ses données, l'apport conceptuel de l'AFE est indéniable. Nous avons donc choisi de fournir une explication détaillée de la mise en œuvre de l'AFE. Cette mise en œuvre se déroule en plusieurs étapes, nécessitant chacune diverses méthodologies statistiques (Howard, 2016 ; Mulaik, 2009). Les étapes clés sont décrites ci-après.

Étape 1 : Vérifier l'adéquation des données

Le modèle factoriel de l'AFE suppose que les corrélations entre les variables sont expliquées par un petit nombre de facteurs. Si les variables observées sont peu corrélées entre elles, alors il est vain d'utiliser une AFE. La première étape consiste donc à vérifier que les données se prêtent bien à une AFE. Pour vérifier la multi-colinéarité des variables, les indicateurs statistiques suivants peuvent être utilisés (Mulaik, 2009) :

– **Le test de sphéricité de Bartlett**. Il s'agit d'un test statistique permettant de rejeter l'hypothèse que les variables sont toutes indépendantes, c'est-à-dire que la matrice de corrélations est proche de la matrice identité. Le test vérifie alors que la valeur absolue du déterminant de la matrice de corrélations est significativement différente de 1, compte tenu du niveau de test choisi.
– **L'indice de Kaiser-Meyer-Olkin (KMO)**. Pour vérifier qu'une variable est corrélée avec les autres variables, il faut calculer l'indice de KMO partiel, qui est égal au pourcentage de variance de la variable expliquée par les autres variables. L'indice KMO global est alors calculé à partir des indices KMO partiels. Une valeur supérieure à 0,6 est considérée comme acceptable pour rendre l'AFE pertinente (Howard, 2016).

Étape 2 : Sélectionner un modèle factoriel

Pour choisir un modèle factoriel, l'approche exploratoire se décompose en trois étapes :

1) La première étape est l'extraction des facteurs. On estime, à partir des données, les coefficients de saturation pour un nombre prédéfini k de facteurs. Le principe est d'ajuster le modèle factoriel aux données afin qu'il capture la plus grande partie de la variance commune.
2) La deuxième étape concerne l'interprétabilité du modèle factoriel estimé. L'estimation est une procédure purement statistique, mais il faut s'assurer que les facteurs identifiés sont interprétables par le praticien. Autrement dit, il faut que ce dernier puisse les nommer et les relier à des concepts propres au domaine d'étude. Par exemple, en sciences de l'éducation, le praticien cherchera à associer un facteur à une dimension particulière du phénomène étudié ou à un construit du cadre théorique mobilisé.
3) La dernière étape consiste, pour le praticien, à sélectionner parmi les modèles factoriels statistiquement pertinents celui qui est le plus interprétable au regard de sa problématique.

Extraire les facteurs

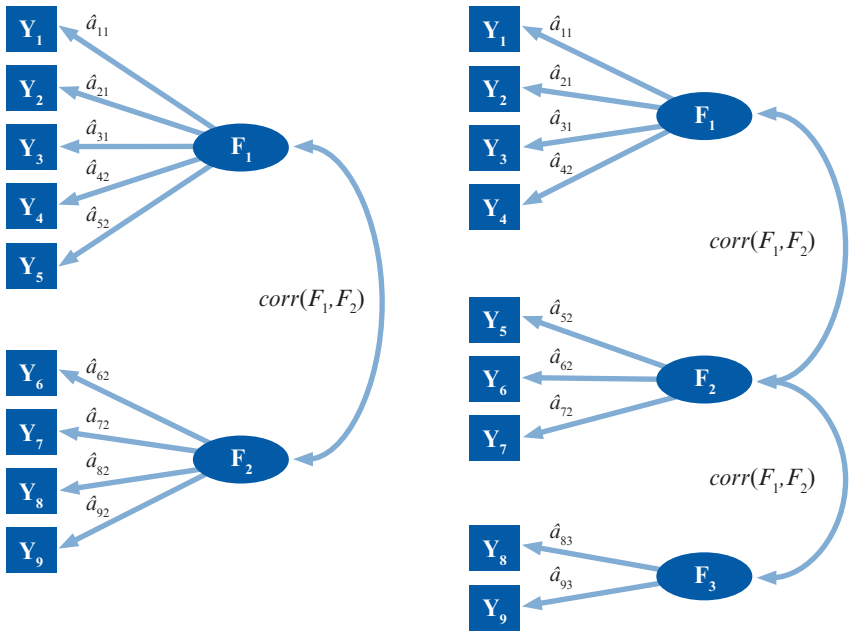
Pour une valeur de k fixée, l'extraction des facteurs consiste à estimer les coefficients de saturation (a_{ij}), à partir des données. Les logiciels statistiques offrent de nombreuses techniques d'estimation. Deux méthodes se distinguent cependant par leur efficacité à estimer les coefficients de saturation, même pour des faibles valeurs de l'indice KMO pour le jeu de

données : la méthode du maximum de vraisemblance (MV), qui peut être utilisée dans le cas où les données sont distribuées selon une loi normale multivariée, et, autrement, la méthode des axes principaux (PA) (Mulaik, 2009). Il est donc essentiel de vérifier la normalité des données avant d'utiliser la méthode du MV, particulièrement avec des données qualitatives ordinales telles que celles issues d'échelles de Likert. Les tests de normalité multivariés les plus populaires sont ceux proposés par Mardia (1974) :

- le test d'asymétrie de Mardia : il s'agit de tester que l'asymétrie multivariée des données est significativement différente de celle d'une distribution normale ;
- le test de kurtosis de Mardia : il s'agit de tester que le coefficient d'aplatissement (*kurtosis*) multivarié des données est significativement différent de celui d'une distribution normale.

Pour chaque valeur du nombre k de facteurs sélectionnés, on obtient un modèle estimé, c'est-à-dire **une structure factorielle latente *a posteriori*** (voir figure 2).

Figure 2
Exemples illustratifs de modèles factoriels pour neuf variables



La figure de gauche présente une structure à deux facteurs, celle de droite, une structure à trois facteurs. Seuls les coefficients élevés en valeur absolue (en pratique $|a_{ij}| > 0{,}3$) sont représentés pour faciliter l'identification de groupes de variables fortement liées aux facteurs.

Interpréter les facteurs

La problématique de l'interprétabilité des facteurs réside dans la capacité du praticien à relier ces facteurs à des concepts compréhensibles et pertinents pour lui. Une méthode pour y parvenir est de rechercher des motifs parmi les coefficients de saturation, afin de regrouper les variables qui présentent des coefficients de saturation élevés avec le même facteur. Une règle généralement acceptée pour sélectionner les variables consiste à considérer comme significatifs les coefficients supérieurs à 0,3 ou 0,4 (Howard, 2016). Les logiciels de statistiques fournissent ainsi des graphiques de variables par facteur (voir figure 2) qui mettent en évidence uniquement les coefficients significatifs, facilitant ainsi l'identification de regroupements pertinents.

Compte tenu du regroupement de variables significativement liées à un facteur, l'objectif est ensuite de nommer ce facteur de manière à refléter un concept, une notion du cadre théorique mobilisé : cette démarche requiert une certaine expertise du domaine étudié. De cette manière, les facteurs identifiés représentent diverses dimensions de la problématique étudiée.

Certains facteurs peuvent ne pas être aisément interprétables, car les variables peuvent charger plusieurs facteurs (c'est-à-dire, avoir des coefficients de saturation élevés avec différents facteurs). Pour y remédier, il est possible d'effectuer une rotation des facteurs (Fabrigar & Wegener, 2011). La décomposition des variables en combinaisons linéaires de facteurs (équation 2) peut être vue comme une projection des variables dans un espace où les facteurs représentent des axes : la position d'une variable par rapport à ces axes est déterminée par ses coefficients de saturation. Par conséquent, une rotation géométrique de ces axes peut augmenter certains coefficients de saturation tout en en diminuant d'autres, permettant ainsi l'émergence de nouveaux facteurs qui sont plus fortement liés aux variables.

Il existe deux types de rotations : orthogonale ou oblique. La rotation orthogonale est utilisée pour obtenir des facteurs indépendants, alors que la rotation oblique permet d'obtenir des facteurs corrélés entre eux. Ce dernier cas est souvent considéré comme plus conforme à la réalité dans le domaine des sciences humaines (Roussel, 2005).

Sélectionner le modèle

Comment choisir le bon modèle, c'est-à-dire la bonne valeur de k ? La citation de George Box « Tous les modèles sont faux, mais certains sont utiles » (1970, p. 202) est un aphorisme courant en statistique. Dans le cadre de l'analyse factorielle exploratoire, le critère pour choisir un modèle repose sur sa capacité à réduire substantiellement la dimension du phénomène étudié tout en expliquant le maximum de variance (utilité computationnelle). Il est également crucial que les facteurs demeurent significatifs et interprétables par le praticien (utilité conceptuelle). Mulaik (2009), parle de « nombre approprié » de facteurs, qui équilibre ces considérations computationnelles et conceptuelles. Fabrigar et Wegener (2011, p. 148) rappellent que cette décision implique aussi un choix conceptuel important.

Les logiciels statistiques mettent à disposition différents critères visant à mesurer l'utilité computationnelle. Les conditions d'utilisation de ces différents critères sont les suivantes :

– Dans le cas de données distribuées selon une loi normale multivariée, le nombre optimal de facteurs latents est celui qui minimise un critère biais-variance. Le nombre optimal de facteurs réalise un compromis entre le biais de modélisation (l'erreur introduite par une modélisation trop simple avec peu de facteurs) et la variance de modélisation (l'erreur introduite par une modélisation trop complexe). Les critères biais-variance les plus fréquemment utilisés sont le critère d'information bayésien (BIC) et le critère d'information d'Akaïke (AIC) (Saporta, 2006). Rappelons que ces critères nécessitent l'hypothèse de normalité des données.
– Sans l'hypothèse de normalité, il est possible d'obtenir divers critères non-paramétriques à partir des valeurs propres de la matrice de corrélations, qui correspondent à la quantité de variance expliquée par chaque facteur extrait :
  - l'*Empirical Kaiser Criterion* (EKC), qui est un critère qui ne retient un facteur que si sa part de variance expliquée est supérieure à celle qui est expliquée par une seule variable. Le critère EKC peut conduire à sélectionner un nombre excessivement élevé de composantes principales (Fabrigar & Wegener, 2011) ;
  - l'analyse parallèle (PA) qui, à l'instar du critère EKC, consiste à ne retenir que les facteurs ayant une part de variance expliquée supérieure à celle d'un jeu de données simulé de manière aléatoire.

Notons que si le critère EKC est pertinent pour choisir le modèle factoriel dans le cadre d'une ACP, puisque les valeurs propres capturent la variance totale des variables, il l'est moins pour l'AFE, où les facteurs capturent la variance commune, et non la variance totale (Fabrigar & Wegener, 2011) : le critère EKC il n'est donc pas recommandé pour l'AFE.

Évaluer l'utilité conceptuelle d'un modèle, c'est déterminer dans quelle mesure les variables ayant des charges factorielles élevées avec un facteur possèdent une certaine cohérence conceptuelle. Lors de la phase d'interprétabilité, l'avis d'un expert sur la pertinence des facteurs identifiés fournit une évaluation qualitative de la cohérence conceptuelle du modèle.

Analyser la structure des individus : la classification non supervisée

Comme nous l'avons vu précédemment, l'AFE permet d'analyser la structure des corrélations et d'identifier des facteurs qu'il est possible de relier à des dimensions spécifiques de la problématique étudiée. Il peut également être utile d'identifier des groupes d'individus significatifs et pertinents par rapport au phénomène étudié. Dans l'étude présentée dans la quatrième partie sur l'adaptation des primo-entrants, il peut être pertinent d'identifier s'il existe des groupes d'étudiants similaires ou différents en matière d'adaptation. Pour cela, on utilise les méthodes de classification non supervisée (ou *clustering*) (Lebart et al., 1995).

Le principe de la classification non supervisée est de construire des classes (ou groupes) d'individus à l'aide d'une mesure de similarité entre individus. Dans le cas de données quantitatives, la similarité est par exemple mesurée par la distance euclidienne : deux individus proches sont considérés comme similaires. Dans le cas de données qualitatives, on peut mesurer la similarité entre deux individus à l'aide des fréquences des réponses : deux individus qui donnent une réponse avec la même fréquence sont considérés comme similaires. Il est important de noter que la classification non supervisée s'appuie sur les coordonnées des individus, obtenues par projection du nuage sur les composantes principales. Lorsque les données sont ordinales, comme celles recueillies à l'aide d'échelles de Likert, les composantes principales sont obtenues par une ACP. En revanche, pour des variables qualitatives nominales, c'est une ACM qui est utilisée (Lebart et al, 1995).

Les méthodes de classification non supervisée les plus couramment utilisées sont l'algorithme des *k*-means et la classification ascendante hiérarchique (CAH), en utilisant la distance euclidienne comme mesure de similarité entre deux individus.

– **L'algorithme des *k*-means** : le nombre de classes k est spécifié à l'avance. Au départ k individus sont choisis et constituent les centres des classes. Puis, les autres individus sont assignés aux classes minimisant la distance avec les centres des classes. Une fois la partition faite, le point moyen des classes est considéré comme un nouveau centre et la procédure est itérée jusqu'à ce que les centres soient stables. L'algorithme des k-means requiert généralement plusieurs itérations pour atteindre une solution optimale. La convergence peut prendre du temps, notamment lorsque le nombre d'individus et de variables ou le nombre de classes est élevé. De plus, l'algorithme est sensible à l'initialisation, ce qui peut être problématique en présence de données aberrantes. Par ailleurs l'algorithme nécessite de fixer *a priori* le nombre k de classes visées.
– **La classification ascendante hiérarchique (CAH)** : au départ, chaque individu est considéré comme une classe distincte. Les classes sont ensuite fusionnées itérativement jusqu'à obtenir une seule classe qui contient tous les individus. Plusieurs principes de fusion existent. Le principe de Ward est le plus populaire : il consiste à fusionner les classes qui minimisent la variance intra-classe après fusion (Saporta, 2006). La CAH produit une structure hiérarchique, ce qui permet une visualisation claire des relations entre les différentes classes à l'aide d'un dendrogramme. Elle ne nécessite pas d'*a priori* sur le nombre de classes. L'avantage de cette méthode est qu'elle permet d'identifier le nombre optimal de classes. En revanche, la CAH est plus exigeante en matière de calcul que l'algorithme des k-means et la structure hiérarchique moins facile à interpréter que les nuages obtenus avec les k-means.

En pratique, on utilise les avantages des deux méthodes en initialisant le processus par une CAH afin de déterminer le nombre approprié de classes, puis en utilisant la configuration obtenue par la CAH pour initialiser l'algorithme des k-means.

Étude de l'adaptation des primo-entrants en IUT : quels sont les constats à la suite des réformes du baccalauréat et du BUT?

En France, les réformes du baccalauréat général en 2019 et du BUT en 2021 ont modifié les profils des lycéens intégrant les IUT, ainsi que les contenus et les méthodes d'évaluation des apprentissages en BUT. Ces changements ont eu des répercussions sur l'adaptation académique des

étudiants. Il est donc est apparu crucial pour les équipes pédagogiques des IUT d'analyser l'adaptation des primo-entrants afin d'apporter des réponses appropriées pour faciliter leur transition du lycée vers l'université. C'est ainsi que l'étude exploratoire présentée ici a été menée pour de formations d'un IUT d'Île de France.

L'enquête et le jeu de données

Pour étudier l'adaptation des primo-entrants, nous avons mené une enquête auprès des étudiants de 1re année de deux formations de BUT dans un IUT d'Île de France. Nous avons utilisé la version courte du *Student Adaptation to College Questionnaire* (SACQ), en français (Carayon & Gilles, 2005). Ce questionnaire est constitué de 41 questions visant à recueillir des informations sur les quatre thèmes suivants :

- l'adaptation académique, pour évaluer l'ajustement académique et la réussite académique des étudiants ;
- l'adaptation sociale, pour évaluer l'ajustement social et les relations avec les pairs et également avec l'équipe pédagogique ;
- l'adaptation émotionnelle, pour évaluer le bien-être émotionnel et psychologique des étudiants ;
- l'adaptation personnelle, pour évaluer l'ajustement personnel de l'étudiant, c'est-à-dire sa capacité à gérer les changements liés à la vie universitaire en dehors du cadre strictement académique.

Le questionnaire a été administré en ligne via *Sphinx*, au printemps 2023. Sur les 628 étudiants sollicités, 362 ont répondu. Pour chaque question, les étudiants ont été invités à se positionner sur une proposition visant à évaluer leur capacité d'adaptation sur les quatre thèmes, en utilisant une échelle de Lickert à 5 points, de -2 (pas du tout d'accord avec la proposition) à +2 (tout à fait d'accord). Pour les propositions formulées de manière négative, les réponses ont été inversées avant les analyses statistiques. Ainsi, une valeur élevée pour une réponse correspond à une adaptation sans difficulté.

La recherche de facteurs expliquant l'adaptation

Nous avons utilisé une AFE pour examiner les relations entre les variables, dans le but de déceler des facteurs susceptibles d'expliquer les difficultés d'adaptation des répondants. Naturellement, nous anticipons de retrouver les dimensions inhérentes au SACQ. Cependant, notre objectif est également de déterminer si des dimensions particulières émergent

spécifiquement pour ces étudiants, dans le cadre des récentes réformes du baccalauréat et du BUT. L'AFE a été mise en œuvre sur ces données à l'aide du *package R Psych* (Revelle, 2021).

```
# Description des données :
# Ce fichier Excel SACQ.xlsx contient les réponses du questionnaire SACQ avec
362 répondants en lignes et 41 questions en colonnes (variables actives)

#Lecture des données
library(readxl)
DATA <- read_excel("SACQ.xlsx",)

#Chargement du package "Psych" pour l'analyse factorielle exploratoire
library(psych)
```

Étape 1 : Vérifier l'adéquation des données

```
# Chargement du package "Hmisc" pour l'analyse de données
library(Hmisc)

# Calcul de la matrice de corrélations type= "spearman" pour des échelles de Likert,
"pearson" sinon)
res_rcorr=rcorr(as.matrix(DATA[,var_actives]), type="spearman")
mat.cor = res_rcorr$r

#test de sphéricité de Bartlett
cortest.bartlett(mat.cor)

#Calcul du score KMO
KMO(mat.cor)
```

Sur ces données, le test de sphéricité de Bartlett valide l'hypothèse que les items ne sont pas tous corrélés, avec un niveau de signification inférieur à 0,001. L'indice KMO global est quant à-lui de 0,888 (supérieur à 0,6), confirmant l'adéquation des données à une AFE. Par ailleurs, on observe qu'aucun item n'est indépendant des autres : les indices KMO partiels varient entre 0,66 et 0,95.

En outre, nous avons effectué des tests de normalité multivariée pour vérifier l'hypothèse de normalité des données (Mardia, 1974).

```
# Chargement du package "MVN" pour les tests de normalité multivariée
library(MVN)

# Tests de Mardia
res.mvn <- mvn(as.matrix(DATA[,var_actives]),mvnTest = "mardia")
print(res.mvn$multivariateNormality)
```

Les résultats de ces deux tests sur les données permettent de conclure au rejet de l'hypothèse de normalité des données étudiées, avec un niveau de signification inférieur à 0,01.

Étape 2 : Sélectionner un modèle factoriel

Les données n'étant pas gaussiennes, les facteurs ont été extraits en utilisant la méthode des axes principaux (PA) avec une rotation oblique, pour obtenir des facteurs potentiellement corrélés entre eux.

```
# Effectuer une analyse parallèle pour déterminer le nombre de facteurs à retenir pour
l'analyse factorielle :
# 'fm = "pa"' spécifie la méthode d'extraction des facteurs principaux,
# 'fa="fa"' indique le type d'analyse factorielle,
# 'n.iter = 100' définit le nombre d'itérations pour l'analyse parallèle.
res.pa = fa.parallel(mat.cor, fm = "pa", fa="fa", n.iter = 100)
res.fa$nfact  #nombre de facteurs préconisés
```

La méthode d'analyse parallèle (PA) préconise une structure à cinq facteurs (tableau 1, ci-après).

```
# Réaliser une analyse factorielle :
# 'nfactors = 5' définit le nombre de facteurs,
# 'fm = "pa"' indique à nouveau la méthode d'extraction des facteurs principaux,
# 'rotate = "oblimin"' spécifie la méthode de rotation oblimin, utile pour les facteurs
corrélés.
res_fa5 = fa(mat.cor, nfactors = 5, fm = "pa", rotate = "oblimin")

# Enregistrement des coefficients de saturation dans un fichier Excel
loadings_df <- as.data.frame.matrix(res_fa5$loadings)
loadings_df$Variables <- rownames(res_fa5$loadings)
library("writexl")
write_xlsx(loadings_df, path = "fa5_loadings.xlsx")
```

Le tableau 1 présente les coefficients de saturation obtenus pour les modèles à cinq facteurs.

Tableau 1
Coefficients de saturation du modèle factoriel à cinq facteurs

Variables	PA1	PA2	PA3	PA4	PA5
AF_Choix	**0,90**	-0,01	-0,04	-0,01	0,02
AF_Choix_inv	**0,78**	0,11	-0,05	0,03	-0,12
AF_Formation	**0,83**	-0,03	0,04	-0,01	0,11
AF_Terminer	**0,60**	-0,03	0,05	0,05	-0,22
AF_Interet	**0,63**	-0,05	-0,04	0,01	0,21
AF_Integration_inv	**0,47**	0,25	**0,31**	-0,02	-0,07
AF_Satisfaction	**0,46**	0,16	0,19	-0,04	0,29
AF_Projet_Pro	**0,38**	0,07	-0,04	0,10	-0,02
AA_Revisions	0,27	0,15	-0,02	0,17	-0,11

Variables	PA1	PA2	PA3	PA4	PA5
AF_Niveau_inv	0,03	**0,80**	0,02	-0,04	-0,10
AF_Rythme_inv	0,01	**0,71**	0,03	0,00	0,13
AF_Comprehension_inv	0,08	**0,61**	0,02	-0,09	0,00
AF_Procrastination_inv	-0,06	**0,50**	0,05	0,07	-0,16
AF_Resultats	0,09	**0,43**	-0,04	0,10	-0,15
AF_Etat_psych_inv	0,02	**0,42**	0,17	0,25	0,01
AF_Competences	0,26	**0,34**	0,02	0,08	-0,13
AF_Gestion_Quotifien_inv	-0,07	**0,42**	0,06	**0,33**	0,18
AF_Gestion_temps	0,13	**0,36**	-0,02	0,26	0,02
AF_Monde_Travail	-0,06	-0,22	0,12	-0,02	-0,08
AA_Participation	0,09	0,21	0,10	0,00	-0,04

Variables	PA1	PA2	PA3	PA4	PA5
AF_Pairs_Amis	-0,06	-0,02	**0,81**	-0,01	-0,07
AF_Pairs_TD	-0,05	0,04	**0,76**	-0,05	0,06
AF_Pairs_Social	0,11	0,05	**0,67**	0,01	0,16
AF_Pairs_Com_Etudiants_inv	0,04	0,05	**0,59**	0,04	-0,09
AF_Pairs_Conflits	-0,04	-0,08	**0,55**	0,10	0,09
AA_Etudier_IUT	0,18	-0,22	0,24	0,10	0,10

Tableau 1
Coefficients de saturation du modèle factoriel à cinq facteurs (suite)

Variables	PA1	PA2	PA3	PA4	PA5
AP_Finances	-0,05	-0,03	-0,12	**0,64**	0,09
AP_Budget	-0,12	0,12	-0,13	**0,58**	0,10
AP_Logements	0,06	-0,06	0,13	**0,57**	-0,15
AP_Sommeil	0,10	0,15	-0,02	**0,43**	0,17
AP_Loisirs	0,09	0,01	0,22	**0,48**	-0,17
AP_EDT	0,03	**0,36**	-0,04	**0,39**	0,23
AP_Nourriture	0,09	-0,19	0,09	**0,52**	-0,08
AP_Soutien_proches	0,16	0,04	0,28	**0,37**	-0,15
AP_Etudier_Lieu	0,28	-0,11	0,18	**0,30**	0,18
AP_Sante_physique	0,01	0,23	0,24	**0,32**	-0,08
AS_EP_Presence	0,02	0,08	0,05	0,26	0,02

Variables	PA1	PA2	PA3	PA4	PA5
AS_EP_Com_Ens_inv	0,05	0,22	0,13	0,03	**0,49**
AS_EP_Accueil	0,23	-0,12	0,11	0,05	**0,48**
AS_EP_Relations_Ens	0,22	0,08	0,15	0,05	**0,41**
AS_EP_Transport_inv	0,15	0,17	-0,13	0,07	0,21

Note. Les coefficients supérieurs à 0,3 en valeur absolue sont surlignés.

Le modèle à cinq facteurs identifiés retrouve globalement les dimensions de l'adaptation en 1re année à l'université proposés dans Carayon et Gilles (2005) et De Clercq et al. (2020) : adaptation académique à travers le facteur PA2, adaptation à la formation à travers le facteur PA1 et adaptation personnelle à travers le facteur PA4. Cependant, la dimension adaptation sociale est décomposée en deux sous dimensions :

– **le facteur PA3,** lié aux questions concernant les interactions avec les pairs, par exemple :
  - AS_Amis : « Je me suis fait des ami·es dans ma formation »
  - AS_Relations_TD : « Je suis globalement satisfait·e de mes relations avec les étudiant·e·s de mon groupe de TD »
  - AS_Social : « Je suis satisfait·e de ma vie sociale à l'IUT ».
  Nous le nommons **Interactions avec les pairs (IP).**

– **le facteur PA5,** lié aux questions concernant les interactions avec les enseignants, par exemple :
  - AS_Com_Enseigants_inv : « Je rencontre des difficultés à communiquer avec les enseignant·es »
  - AS_Accueil : « À la rentrée, je me suis senti·e accueilli·e par l'équipe pédagogique »
  - AS_Relations_Ens : « Je suis globalement satisfait·e de mes relations avec les enseignant·e·s ».

  Nous le nommons **Interactions avec les enseignants (IE)**.

Notons que la question AP_Transport_inv : « Le temps que je passe dans les transports entre chez moi et l'IUT est, pour moi, difficile à supporter » est corrélée à ce facteur et non au facteur PA2 adaptation personnelle. Cela peut s'expliquer par le fait que, dans les formations concernées par cette étude, la tolérance aux retards est à la discrétion des enseignants.

Cette décomposition en cinq facteurs est apparue pertinente pour aborder la problématique de l'accueil et de l'accompagnement des primo-entrants durant leurs premiers jours à l'université, car elle souligne le besoin de concevoir, d'une part, des activités favorisant les interactions entre pairs et, d'autre part, de réfléchir aux interactions avec l'équipe pédagogique. En effet, de nombreux travaux ont montré l'importance des interactions entre pairs, qui favorise l'intégration, renforce le sentiment d'appartenance, fournit un soutien émotionnel entre étudiants, etc., et donc, favorise indirectement la réussite académique (Masson & Ratenet, 2020 ; Tinto, 2012). D'autre part, de nombreux travaux soulignent l'importance de la communication entre les enseignants et les étudiants ainsi que de l'évaluation formative et de la rétroaction pour soutenir l'apprentissage (Raucent et al., 2021 ; Tardif, 1992).

La recherche de profils d'adaptation

Après avoir déterminé la structure factorielle sous-jacente aux données, nous avons mis en œuvre une classification non supervisée afin de distinguer différents profils de primo-entrants en matière d'adaptation à l'IUT. Notons que l'AFE a montré l'existence de cinq groupes de variables liées aux facteurs et que la composition de ces groupes n'est pas uniforme ici : par exemple, le facteur PA2 est lié à 13 variables alors que les facteurs PA3 et PA5 ne sont reliés qu'à quatre variables. L'impact du groupe PA2 est donc artificiellement plus important dans la mesure de la similarité

entre deux étudiants. Afin d'équilibrer l'influence de chaque groupe de variables lors de la classification non supervisée, nous avons tout d'abord mis en œuvre une analyse factorielle multiple (AFM).

Étape 1 : Réduire le bruit d'échantillonnage à l'aide d'une AFM

L'AFM a été mise en œuvre sur ces données à l'aide du *package R FactoMineR* (Husson et al., 2023).

```
# Chargement du package "Factominer"
library(FactoMineR)

# Définition des indices des variables pour chaque groupe conceptuel
var.AA=1:11   # Groupe de variables Adaptation Académique
var.AF=12:19  # Groupe de variables Adaptation à la Formation
var.AP=20:32  # Groupe de variables Adaptation Personnelle
var.IP=33:36  # Groupe de variables Interactions avec les Pairs
var.IE=37:41  # Groupe de variables Interactions avec les Enseignants
var.sup_quanti= 42:46  # Moyennes par groupe : AA, AF, AP, IE, IP
var.sup_quali= 47:49   # Variables qualitatives supplémentaires

# Exécution de l'analyse factorielle multiple
res.mfa <- MFA(DATA[,c(var.AA,var.AF,var.AP,var.IE,var.IP,var_sup_quanti,
var_sup_quali)],
  ncp=33,  # Nombre de composantes à retenir
  group = c(11,8,13,4,5,5,3),  # Taille de chaque groupe de variables
  type = c(rep("s", 5),"s","n"),  # Types des variables dans chaque groupe ("s" pour
symétrique, "n" pour nominal)
  name.group = c("Groupe AA","Groupe AF","Groupe AP","Groupe IP","Groupe
IE","SUP_quali","SUP_quanti"),  # Noms des groupes de    variables
  num.group.sup=c(6,7),  # Groupe de variables supplémentaires
  graph = FALSE  # Ne pas générer de graphiques automatiquement
      )
```

Les 33 premières composantes principales représentent plus de 95% de la variance totale des 41 variables (voir tableau 2).

Il est important de souligner que les dernières composantes principales (de 34 à 41) capturent moins de 5% de la variance totale. Elles peuvent par conséquent être considérées comme du bruit d'observation. Ignorer ces composantes permet de rendre la classification non supervisée plus robuste vis-à-vis de l'échantillon étudié. Cela met en évidence la capacité de l'ACP (et donc de l'AFM qui en découle) à réduire la dimension des données et à éliminer le bruit d'échantillonnage.

Tableau 2
Composantes de l'AFM

Composante	% cumulé de variance expliquée
1	23,1 %
2	30,0 %
3	36,2 %
⋮	⋮
33	95,5 %

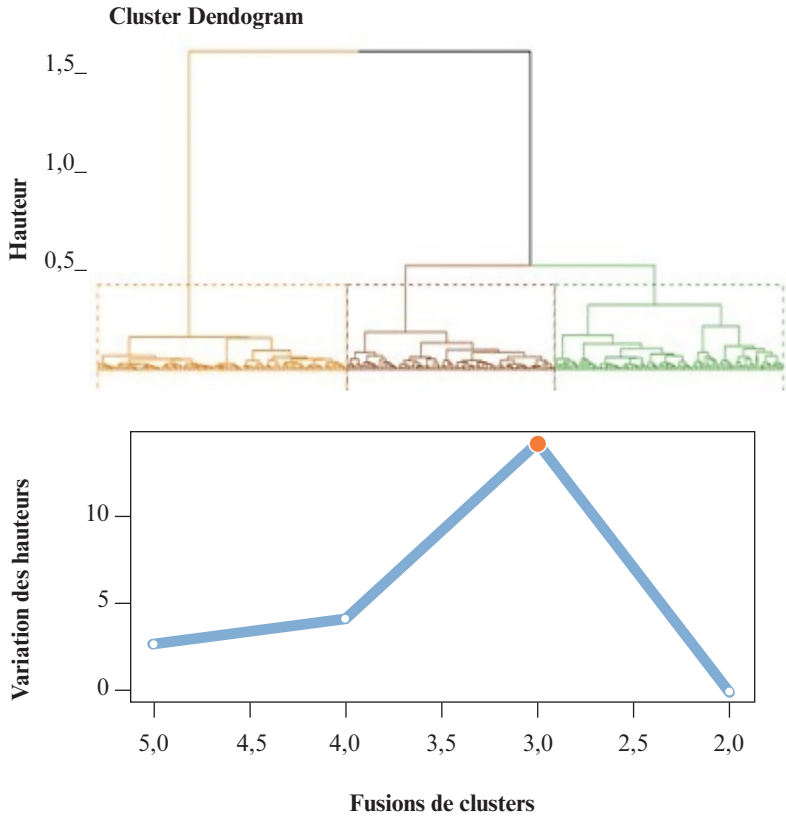
Étape 2 : Construire des profils d'adaptation à l'aide d'une classification non supervisée

La classification non supervisée a été mise en œuvre sur la projection du nuage initial, avec 41 variables, sur les 33 premières composantes principales de l'AFM. Pour cela, nous avons utilisé la fonction HCPC du *package R FactoMineR* (Husson et al., 2023). Comme indiqué dans la troisième partie, les deux méthodes de classification présentent chacune des avantages et des inconvénients. La fonction HCPC les combine afin de tirer profit des avantages de chacune : une CAH est tout d'abord effectuée pour déterminer le nombre optimal de classes.

Le dendrogramme obtenu par la CAH sur les 33 premières composantes de l'AFM est présenté dans la partie supérieure de la figure 3. L'analyse du gain de variance-intra après fusion (voir figure 3, partie inférieure) montre que celui-ci est optimal lorsque l'on passe de quatre à trois classes. Nous avons donc retenu une structure latente d'individus à trois classes.

```
# Classification non supervisée : une CAH est effectuée sur les 33 première
composantes principales de l'AFM afin d'obtenir un nombre optimal de classes, puis
le résultat est affiné à l'aide de la méthode des k-means
res.hcpc.mfa <- HCPC(res.mfa, graph = FALSE)
fviz_dend(res.hcpc.mfa,
        cex = 0.7,    # Taille du texte
        palette = c("#ff7f00","#a65628","#4daf4a"), # couleurs des classes
        rect = TRUE, rect_fill = TRUE, # Rectangle autour des classes
        rect_border = c("#ff7f00","#a65628","#4daf4a"),# Couleur du rectangle
        ylim = c(-0.1, 1.0),
        horiz = F,
        labels_track_height = 0.8      # Augmente l'espace pour le texte
        )
```

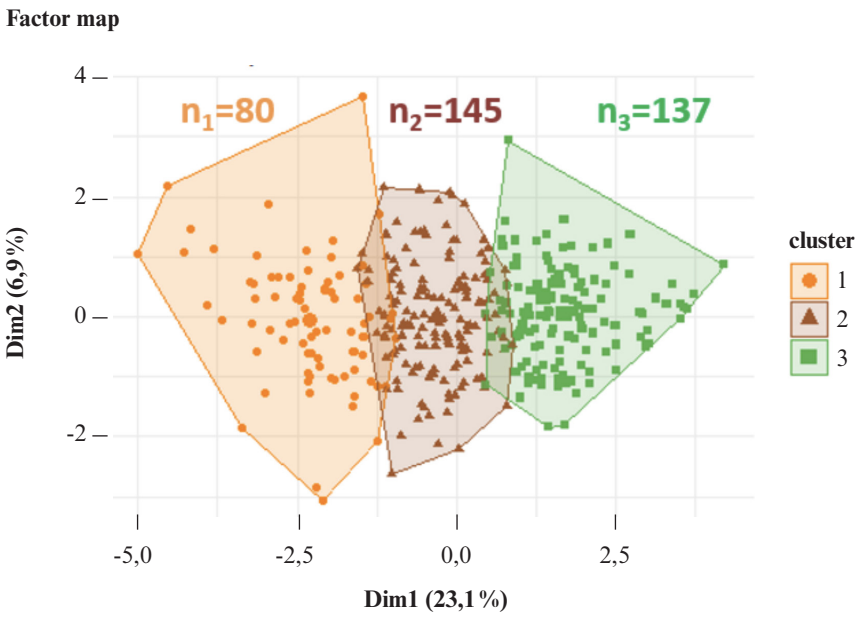
Figure 3
Classification hiérarchique et sélection du nombre de clusters



La partie supérieure présente le dendrogramme de la CAH sur les 33 premières composantes principales. L'ordonnée représente la variance-intra après fusion. Le graphique présenté dans la partie inférieure montre le gain de variance-intra en fonction du nombre de classes. Le gain est optimal en passant de quatre à trois classes.

Puis, les trois classes obtenues par la CAH servent à initialiser l'algorithme des *k*-means qui permet d'obtenir des classes plus homogènes et mieux séparées. La figure 4 présente les trois classes ainsi obtenues, projetées sur le plan principal de l'AFM.

Figure 4
Projection des trois classes sur le premier plan de l'AFM



Étape 3 : Identifier les profils d'adaptation à l'aide d'une caractérisation des classes

Les trois classes regroupent respectivement 80, 145 et 137 étudiants : un étudiant ayant une coordonnée élevée sur la composante 1 a des valeurs élevées pour l'ensemble des réponses aux questions. La projection des classes sur le plan principal de l'AFM (voir figure 4 ci-dessus) laisse à penser que les profils obtenus correspondent à des groupes d'individus ayant des scores d'adaptation plus ou moins élevés.

Pour décrire les différentes classes, il est possible d'utiliser des indicateurs synthétiques ou des variables supplémentaires (c'est-à-dire non utilisées pour faire l'AFM et la classification). Dans ce contexte, pour chaque individu, nous avons calculé la moyenne des variables des cinq groupes de l'AFE. Pour évaluer les différences entre les trois classes, nous avons utilisé le test de Kruskal-Wallis, choisi pour sa capacité à comparer des classes sans hypothèse sur la distribution des données (Lebart et al., 1995). Le test est appliqué sur chacune des moyennes des cinq groupes de l'AFE.

Il s'agit décider entre les hypothèses :

1) hypothèse nulle (H0) : les médianes des classes sont égales ;
2) hypothèse alternative (H1) : au moins une des médianes diffère significativement des autres.

Nous avons utilisé la fonction *kruskal.test* du *package stats* de R.

```
# Chargement du package "stats"
library(stats)
#test de Kruskall-Wallis pour sur les moyennes par groupe de variables
kruskal.test(res.hcpc.mfa$data.clust$Moy_AA~res.hcpc.mfa$data.clust$clust)
kruskal.test(res.hcpc.mfa$data.clust$Moy_AF~res.hcpc.mfa$data.clust$clust)
kruskal.test(res.hcpc.mfa$data.clust$Moy_AP~res.hcpc.mfa$data.clust$clust)
kruskal.test(res.hcpc.mfa$data.clust$Moy_AS_Pairs~res.hcpc.mfa$data.clust$clust)
kruskal.test(res.hcpc.mfa$data.clust$Moy_AS_EP~res.hcpc.mfa$data.clust$clust)
```

Les cinq mises en œuvre du test de Kruskal-Wallis ont permis de conclure, avec à chaque fois un niveau de signification inférieure à 0,01, à l'existence d'une différence entre les trois classes. Pour préciser la nature de cette différence, après chaque test de Kruskal-Wallis, nous avons procédé à un test *post-hoc*. Ce dernier compare la moyenne de chaque classe à la moyenne générale, en se basant sur des valeurs-tests calculées pour chaque classe (Lebart et al., 1995) :

$$Z_{\text{classe}} = \frac{\overline{X}_{\text{classe}} - \overline{X}}{\sqrt{s^2_{\text{classe}}}} \quad \text{où} \quad s^2_{\text{classe}} = \frac{n - n_{\text{classe}}}{n - 1} \frac{s^2(X)}{n_{\text{classe}}}$$

Ces valeurs-tests peuvent être interprétées comme la statistique d'un test de comparaison de moyennes, où l'hypothèse nulle stipule que les observations dans cette classe sont tirées au hasard dans l'ensemble de la population. Dans l'hypothèse d'un tirage au hasard, la valeur-test, Z_{classe}, a 95 chances sur 100 d'être comprise dans l'intervalle [-1.96 et +1.96]. Si la valeur de Z_{classe} est inférieure à -1,96 (resp. supérieure à 1,96), alors nous pouvons conclure que la moyenne de la classe est significativement inférieure (resp. supérieure) à celle de l'ensemble de la population ; sinon (c'est-à-dire valeur entre -1,96 et 1,96), l'écart entre les moyennes n'est pas significatif.

Les calculs des valeurs-tests sont effectués à l'aide de la fonction catdes du package FactoMineR.

```
#Tests post hoc : calcul des valeurs-tests sur les variables supplémentaires
res.desc = catdes(res.hcpc.mfa$data.clust[c(var_sup_quanti, var_sup_quali),], num.
var=var.clust, proba=0.05)
```

Les résultats sont présentés dans le tableau 3. Chaque sous tableau indique la moyenne globale, la moyenne par classe et la valeur test. Lorsque la valeur test est inférieure à -1,96 (resp. supérieure à 1,96), on peut conclure que la moyenne du groupe est significativement inférieure (resp. supérieure) à celle de l'ensemble de la population interrogée ; sinon (valeur entre -1,96 et 1,96) l'écart entre les moyennes n'est pas significatif.

L'analyse des valeurs-tests du tableau 3 montre que :

– la classe 1 regroupe des individus ayant, en moyenne, des valeurs inférieures à la moyenne de la population interrogée pour tous les groupes de variables résultants de l'AFE. Ce profil correspond à celui de individus qui rencontrent potentiellement des difficultés d'adaptation dans toutes les dimensions de l'AFE ;

– la classe 2 regroupe des personnes ayant, en moyenne, des valeurs inférieures à la moyenne de la population interrogée pour les trois groupes de variables Adaptation personnelle, Interactions avec les pairs et Interactions avec les enseignants. Les individus de ce profil déclarent des difficultés pour gérer le quotidien de la vie étudiante en dehors du cadre universitaire et dans leur intégration sociale. Par contre, il n'y a pas de différence significative pour les groupes de variables Performance académique et Adaptation à la formation. Donc, les individus de ce groupe ne rencontrent potentiellement pas de difficultés dans ces deux dimensions ;

– la classe 3 regroupe des individus ayant, en moyenne, des valeurs supérieures à la moyenne de la population interrogée pour tous les groupes de questions de l'AFE. Les individus ayant ce profil déclarent significativement plus que les autres ne pas rencontrer de difficultés d'adaptation à l'IUT.

Tableau 3
Caractérisation des trois classes pour chaque groupe de variables de l'AFE

Moyenne des variables du groupe Adaptation académique (AA)		
Groupe	**Moyenne**	Z_{classe}
Population	0,42	
Classe 1	-0,24	-11,86
Classe 2	0,38	-0,97
Classe 3	0,83	11,12

Moyenne des variables du groupe Interactions avec les pairs (IP)		
Groupe	**Moyenne**	Z_{classe}
Population	0,96	
Classe 1	-0,235	-11,86
Classe 2	0,84	-2,52
Classe 3	1,48	9,47

Moyenne des variables du groupe Adaptation à la formation (AF)		
Groupe	**Moyenne**	Z_{classe}
Population	0,43	
Classe 1	-0,295	-12,51
Classe 2	0,71	0,26
Classe 3	1,27	10,41

Moyenne des variables du groupe Interactions avec les enseignants (IE)		
Groupe	**Moyenne**	Z_{classe}
Population	0,51	
Classe 1	-0,115	-8,79
Classe 2	0,36	-3,45
Classe 3	1,04	11,02

Moyenne des variables du groupe Adaptation personnelle (AP)		
Groupe	**Moyenne**	Z_{classe}
Population	0,43	
Classe 1	-0,18	-9,20
Classe 2	0,22	-4,95
Classe 3	1,01	12,89

Pour enrichir la description des profils identifiés, il est possible d'intégrer des variables supplémentaires telles que des détails sur le parcours académique de l'étudiant (par exemple, la série du baccalauréat) ou des données issues d'études qualitatives comme des rapports d'étonnement des primo-entrants (Pamphile et al., 2024). Cependant, l'intention première de cette étude est un objectif didactique lié à la mise en œuvre d'analyses statistiques exploratoires sur des données en éducation.

Conclusion et discussion

L'analyse de données multivariées est courante dans de nombreuses disciplines, y compris en sciences de l'éducation. Les défis principaux lors de l'exploration de données multivariées résident dans l'identification de la diversité des mécanismes sous-jacents, et surtout dans l'analyse de la complexité de leurs interactions. Les analyses statistiques exploratoires sont des outils statistiques particulièrement performants pour identifier et analyser la structure cachée sous-jacente de données multivariées : d'une part, les analyses factorielles permettent d'explorer les liaisons entre les variables, et, d'autre part, les méthodes de classification non supervisée permettent d'explorer les similitudes entre individus. La mise en œuvre de ces méthodes nécessite toutefois une expertise afin d'en éviter les écueils et d'obtenir des analyses pertinentes et fiables. L'objectif de cet article est de guider les chercheurs en sciences de l'éducation qui souhaitent utiliser ces méthodes pour explorer leurs données.

Nous avons dans un premier temps présenté les diverses analyses statistiques exploratoires en précisant leurs objectifs et leurs conditions d'utilisation. Nous avons illustré leur mise en œuvre sur des données recueillies à l'aide d'un questionnaire sur l'adaptation académique, administré auprès des primo-entrants dans un IUT, à la suite des réformes du bac général en 2019 et du BUT en 2021. Une AFE a permis d'obtenir une représentation synthétique des relations entre les 41 variables du questionnaire. Nous avons retenu un modèle à cinq facteurs, que nous avons identifiés comme étant des dimensions expliquant les difficultés d'adaptation des primo-entrants. Après une AFM préalable sur les groupes de variables liées aux facteurs identifiés lors de l'AFE, une classification non supervisée des individus, combinant une CAH et un *k*-means, nous a permis de sélectionner trois classes. Ces classes ont été caractérisées en comparant les moyennes des groupes de variables de l'AFE. Les trois classes obtenues correspondent à des profils différents d'adaptation de primo-entrants : 1) un profil d'individus déclarant avoir eu des difficultés d'adaptation dans les cinq dimensions identifiées, 2) un profil d'individus déclarant avoir eu des difficultés d'adaptation dans deux dimensions et enfin 3) un profil d'individus déclarant n'avoir pas éprouvé de difficultés d'adaptation. Si le principal objectif visé par les activités pédagogiques proposées dans les enseignements est la performance académique des étudiants, notre étude

permet aux équipes pédagogiques de prendre conscience du fait que la mise en place d'activités favorisant l'intégration sociale des primo-entrants en IUT pourraient également leur être bénéfique.

En conclusion, les méthodes d'analyses statistiques exploratoires se sont montrées efficaces pour identifier des facteurs latents expliquant l'adaptation des primo-entrants à l'IUT et mettre à jour divers profils de primo-entrants en ce qui a trait aux difficultés d'adaptation. Les résultats contribuent ainsi à une compréhension holistique de l'adaptation de ces primo-entrants, ce qui permet aux équipes pédagogiques de prendre des décisions éclairées et de concevoir des activités adaptées.

Les apports des analyses statistiques exploratoires pour les praticiens en sciences de l'éducation

Les analyses statistiques exploratoires constituent un ensemble d'instruments efficaces pour identifier et pour analyser la structure sous-jacente de données multivariées. Elles permettent d'aborder les problématiques suivantes :

- identifier la structure latente des variables mesurées : les analyses factorielles permettent d'analyser la structure des corrélations entre les variables et ainsi, d'identifier des facteurs clés expliquant le phénomène étudié quant au cadre théorique mobilisé ;
- mettre à jour des profils caractéristiques des individus étudiés : les méthodes de classification non supervisée permettent de segmenter la population étudiée en groupes distincts d'individus similaires. Il est alors possible de construire des indicateurs synthétiques et d'utiliser des méthodes d'analyses supervisées sur ces groupes homogènes d'individus ;
- guider les chercheurs vers de nouvelles questions de recherche : en identifiant des facteurs ou des groupes d'individus inattendus ou atypiques, les analyses non supervisées peuvent faire émerger de nouveaux questionnements.

Les limites de cet article

Cet article a été rédigé avec un objectif didactique : guider le praticien dans l'utilisation de l'AFE et la classification non supervisée afin d'explorer ses données. Nous avons délibérément omis de discuter des analyses factorielles pour les données qualitatives non ordinales ou les données textuelles. Pour ce qui est des données qualitatives, les lecteurs

intéressés peuvent se tourner vers l'analyse des correspondances multiples (ACM). Concernant les données textuelles, obtenues par exemple à partir de commentaires lors d'évaluations de formations, une approche consiste à convertir les mots en vecteurs numériques en utilisant des réseaux de neurones, puis à effectuer une classification afin d'identifier les thèmes abordés (Hardeniya et al, 2016 ; Pamphile et al., 2024). Du point de vue pratique, afin d'analyser la fidélité des facteurs identifiés, nous n'avons pas abordé la phase d'analyse factorielle confirmatoire (AFC), ni les analyses prédictives (Mulaik, 2009).

Révision linguistique : Marie-Claire Legaré

Mise en page : Emmanuel Gagnon

Résumé en portugais : Eusébio André Machado

Réception : 05 septembre 2023

Version finale : 18 novembre 2024

Acceptation : 15 avril 2025

LISTE DES RÉFÉRENCES

Box, G. E. (1979). Robustness in the strategy of scientific model building. Dans *Robustness in statistics* (pp. 201-236). Academic Press.

Carayon, S. & Gilles, P. Y. (2005). Développement du questionnaire d'adaptation des étudiants à l'université (QAEU). *L'orientation scolaire et professionnelle*, (34/2), 165-189. https://doi.org/10.4000/osp.463

Carifio, J. & Perla, R. J. (2007). Ten common misunderstandings, misconceptions, persistent myths and urban legends about Likert scales and Likert response formats and their antidotes. *Journal of social sciences*, *3*(3), 106-116. https://doi.org/10.3844/jssp.2007.106.116

De Clercq, M., Roland, N., Dangoisse, F. & Frenay, M. (2023). La transition vers l'enseignement supérieur : comprendre pour mieux agir sur l'adaptation des étudiants en première année. Peter Lang

Escoffier, B. & Pagès, J. (1998). *Analyses factorielles simples et multiples*. Dunod.

Fabrigar, L. R. & Wegener, D. T. (2011). *Exploratory factor analysis*. Oxford University Press.

Hardeniya, N., Perkins, J., Chopra, D., Joshi, N. & Mathur, I. (2016). *Natural language processing: python and NLTK*. Packt Publishing Ltd.

Howard, M. C. (2016). A review of exploratory factor analysis decisions and overview of current practices: What we are doing and how can we improve? *International Journal of Human-Computer Interaction*, *32*(1), 51-62. https://doi.org/10.1080/10447318.2015.1087664

Husson, F., Josse, J., Lê, S. & Mazet, J. (2023). *Package 'FactoMineR'. Multivariate Exploratory Data Analysis and Data Mining*. Available on CRAN. https://cran.r-project.org/web/packages/FactoMineR/FactoMineR.pdf

Husson, F., Lê, S. & Pagès, J. (2011). Exploratory multivariate analysis by example using R. CRC press.

Lebart, L., Morineau, A. & Piron, M. (1995). *Statistique exploratoire multidimensionnelle* (Vol. 3). Dunod.

Masson, J. & Ratenet, L. (2020). Relation entre sentiment d'efficacité personnelle à entrer à l'université chez les étudiants de 1er cycle et stratégies de coping : construction et validation d'une échelle. *Revue internationale de pédagogie de l'enseignement supérieur*, *36*(1). https://doi.org/10.4000/ripes.2319

Mardia, K. V. (1974). Applications of some measures of multivariate skewness and kurtosis in testing normality and robustness studies. *Sankhyā: The Indian Journal of Statistics*, Series B, 115-128. https://www.jstor.org/stable/25051892

Mulaik, S. A. (2009). *Foundations of factor analysis*. CRC press.

Pamphile, P., Bournaud, I. & Clavel, C. (2024). Identifier et comprendre les difficultés d'adaptation des primo entrantes à l'université : utilisation d'une méthode mixte quantitative-qualitative avec des méthodes statistiques d'apprentissage automatique. Communication au *Colloque International Diversité et Réussite dans l'Enseignement Supérieur DIRES*. 3-5 Avril, Nantes. https://hal.science/hal-04489836

Raucent, B., Verzat, C., Van Nieuwenhoven, C. & Jacqmot, C. (2021). Accompagner des étudiants-Quels rôles pour l'enseignant ? Quels dispositifs ? Quelles mises en œuvre ? De Boeck Supérieur.

Revelle, W. (2021). *Psych: Procedures for Psychological, Psychometric, and Personality Research* (Version 2.2.6) [Logiciel]. R package. https://cran.r-project.org/web/packages/psych/index.html

Roussel, P. (2005). Chapitre 9. Méthodes de développement d'échelles pour questionnaires d'enquête. Dans P. Roussel et F. Wacheux (dir), *Management des ressources humaines : Méthodes de recherche en sciences humaines et sociales* (p. 245-276). De Boeck Supérieur.

Saporta, G. (2006). Probabilités, analyse des données et statistique. Éditions Technip.

Tardif, J. (1992). *L'enseignement stratégique*. Éditions Logiques.

Tinto, V. (2012). Leaving college: Rethinking the causes and cures of student attrition. University of Chicago Press.