# Etudier la structure latente de données multivariées à l'aide d'analyses non supervisées, en science de l'éducation.

Patrick Pamphile, Isabelle Bournaud

# Etudier la structure latente de données multivariées à l'aide d'analyses non supervisées, en science de l'éducation.

**Patrick PAMPHILE**, enseignant-chercheur, Laboratoire de Mathématiques d'Orsay - CNRS UMR 8628, Inria - Team CELESTE, Université Paris-Saclay, France

ORCID : https://orcid.org/ 0000-0002-4560-9069

patrick.pamphile@universite-paris-saclay.fr

**Isabelle BOURNAUD**, enseignante-chercheuse, UR Etudes sur les Sciences et les Techniques, Université Paris-Saclay, France

ORCID : https://orcid.org/0000-0002-9819-2789

isabelle.bournaud@universite-paris-saclay.fr

**Mots clés**

Analyses non supervisées, Analyse Factorielle Exploratoire, Clustering, Structures cachées, Adaptation académique, Primo-entrant·es, Transition Lycée-Université

**Résumé en français**

La présence de structures latentes au sein de données multivariées est très fréquente en science de l'éducation : profils inconnus d'étudiants ou facteurs cachés expliquant la réussite académique, l'adaptation à l'université, etc. Ces structures ne sont pas directement observables, mais peuvent jouer un rôle significatif sur le phénomène étudié. Les analyses statistiques non supervisées sont des outils adaptés à l'identification des structures latentes, mais leur utilisation requiert une expertise pour en éviter les écueils et garantir des analyses fiables. L'objectif de ce travail est de guider les chercheurs en science de l'éducation qui souhaitent les utiliser. Nous présentons en détail leur mise en œuvre sur

la problématique de l'adaptation des primo-entrant·es en IUT, suite aux réformes du bac et du BUT. Une analyse factorielle exploratoire permet d'identifier cinq facteurs expliquant les difficultés d'adaptation et un clustering permet de distinguer trois profils de primo-entrant·es en termes de difficultés d'adaptation à l'université.

**Keywords**

Unsupervised analysis, exploratory factorial analysis, clustering, hidden structures, Academic adjustment, first-year students, high school-university transition

**Abstract**

The presence of latent structures in multivariate data is very common in education: unknown student profiles or hidden factors that explain academic success, adaptation to university, etc. These structures are not directly observable but may play a significant role in the phenomenon under study. Unsupervised statistical analysis is a suitable tool for identifying latent structures in data, but its use requires expertise to avoid pitfalls and ensure reliable analysis. The aim of this paper is to provide guidance to educational researchers who wish to use them. We present in detail their application to the problem of the adaption of first-year students in IUT, following the secondary school diploma and BUT reforms. An exploratory factorial analysis identifies five factors that explain adaptation difficulties, and a clustering analysis distinguishes three profiles of first-year students in terms of adaptation difficulties.

# 1. Introduction

## 1.1. Structures cachées des données multivariées et analyses statistiques non supervisées

Le recueil de données multivariées est courant dans de nombreuses disciplines, y compris en sciences de l'éducation. Lors d'un recueil de données auprès d'une promotion d'étudiant·es, il est fréquent de mesurer plusieurs variables simultanément afin d'examiner les relations potentielles entre elles et comprendre les facteurs qui influencent les résultats académiques, les apprentissage, l'efficacité d'un programme de formation, etc.

La présence de structures latentes au sein de données multivariées est très fréquente dans les sciences humaines et en science de l'éducation en particulier. Les étudiant·es inscrit·es dans la même formation sont susceptibles d'avoir des parcours antérieurs différents, des pratiques d'étude différentes, des enseignant·es différent·es au sein de la même formation avec des activités pédagogiques différentes, des évaluations différentes, etc. ; sans parler de leur contexte culturel, du statut socio-économique familial, de leurs conditions de vie étudiante (budget, logement, trajet pour l'université, job étudiant, …), de leurs motivations et aspirations vis à vis de la formation, ... . Tous ces facteurs contribuent à l'existence de profils différents d'étudiant·es. Cette hétérogénéité est cachée, mais peut avoir une influence sur les apprentissages des étudiant·es, leur expérience d'étudiant et leur réussite académique, en particulier chez les primo-entrant·es à l'université.

Par ailleurs, à l'ère des *big data*, la disponibilité croissante et massive de données numériques est une tendance importante y compris en sciences de l'éducation, avec les questionnaires en ligne, les enregistrements de l'activité en ligne des étudiant·es, la numérisation des données administratives, des données d'évaluation, etc. Les données massives sont elles aussi susceptibles d'hétérogénéité car elles sont recueillies auprès d'un large éventail d'étudiant·es ou collectées sur de longues périodes. Le volume de données est alors tel que cette hétérogénéité ne peut être appréhender *a priori*.

L'hétérogénéité latente est donc une notion importante en sciences humaines car elle conceptualise le fait que les données collectées peuvent être complexes avec de multiples variables interdépendantes et

la présence de facteurs sous-jacents qui ne sont pas directement observables et qui peuvent cependant jouer un rôle significatif dans les phénomènes étudiés.

Lors de l'analyse statistique des données recueillies, l'hétérogénéité latente peut entraîner des violations des conditions d'utilisation des méthodes classiques d'analyses supervisées, telles que la régression linéaire multiple ou la régression logistique, très populaires en science de l'éducation. En effet les résidus ne sont plus forcément gaussiens et les variances peuvent ne plus être constantes. Ne pas tenir compte de l'hétérogénéité latente des données rend ainsi peu fiable les résultats de l'étude et peut conduire à des conclusions erronées.

Les méthodes statistiques non supervisées visent quant à elles, à analyser la structure latente des données, soit en termes de liaisons entre les variables, soit en termes de regroupements pertinents des individus. Ces méthodes permettent ainsi de quantifier les dimensions latentes de la structure des variables mesurées, d'identifier les facteurs latents expliquant le phénomène étudié et enfin de mettre à jour des profils caractéristiques des d'individus de l'étude. En identifiant des facteurs ou des groupes d'individus inattendus ou atypiques, les analyses non supervisées peuvent alors guider les chercheurs vers de nouvelles questions de recherche. D'autre part, en intégrant des profils latents d'individus, il devient moins probable de transgresser les présupposés nécessaires à l'utilisation des modèles de régression ou de classification. L'intégration de dimensions latentes pertinentes permet également à ces modèles de mieux saisir la diversité des données.

Utilisées en science de l'éducation, les méthodes d'analyses non supervisée permettent d'explorer de manière holistique la diversité des profils d'étudiant·es et des contextes pédagogiques, en prenant en compte simultanément les facteurs explicatifs. Les comportements des apprenant·es sont souvent multifactoriels et interdépendants : les analyses non supervisées peuvent aider à décomposer ces comportements complexes en sous-composantes plus simples à appréhender, facilitant ainsi la compréhension globale du phénomène. Cette approche permet ainsi aux équipes pédagogiques d'avoir une compréhension plus approfondie des facteurs influents du phénomène étudié et ainsi mieux adapter

la fréquence de leurs interventions et leurs activités pédagogiques pour répondre aux besoins spécifiques des différents profils d'étudiant·es.

### 1.2. Le cas de l'adaptation des primo-entrant·es en IUT[1]

En France, la réforme du Baccalauréat général en 2019 et celle du BUT en 2021 ont engendré des changements dans les profils des lycéen·nes intégrant les IUT, et 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-entrant·es. Dans cette situation, il est apparu crucial d'appréhender les interactions complexes entre les différentes variables jouant un rôle dans l'adaptation en IUT et d'identifier différents profils d'adaptation parmi les étudiant·es interrogé·es. L'objectif pour les praticiens est de mieux appréhender les facteurs clés qui facilitent l'adaptation des primo-entrant·es et permettre ainsi d'améliorer leur expérience étudiante.

Pour étudier ce phénomène, nous avons utilisé la version courte du questionnaire *Student Adaptation to College Questionnaire* (SACQ), en français (Carayon et Gilles, 2005). Ce questionnaire, constitué de 41 questions, vise à recueillir des informations sur quatre thèmes : Adaptation académique, Adaptation sociale, Adaptation émotionnelle, Adaptation personnelle. Cette enquête et les résultats obtenus sont présentés en détail dans la partie 4, seuls sont fournis ici les éléments permettant d'illustrer les propos sur les méthodes statistiques.

**En quoi les analyses non supervisées peuvent être utiles pour comprendre le phénomène d'adaptation des primo-entrant·es en IUT ?**

Tout d'abord, une analyse factorielle permet de faire émerger des facteurs cachés dits « latents», corrélés à des groupes de variables du questionnaire. La structure factorielle obtenue fournit une représentation simplifiée et structurée des données, facilitant ainsi l'interprétation des résultats.

---

[1] Les IUT (Instituts Universitaires de Technologie) sont des composantes des universités françaises. Ils proposent principalement des formations professionnalisantes de niveau Bac+3, le Bachelor Universitaire de Technologie (BUT).

En outre, en comparant la structure *a priori* du questionnaire à la structure factorielle obtenue *a posteriori*, l'analyse factorielle permet de valider le questionnaire, c-à-d de vérifier que les questions posées permettent bien de capter les thèmes visés.

Enfin, les méthodes de clustering permettent de construire des profils de primo-entrant·es en termes d'adaptation.

### 1.3. Objectif de l'article

Les analyses statistiques non supervisées sont ainsi adaptées pour étudier la structure latente de données multivariées. Si l'avènement des Big Data, du Deep Learning et du Machine Learning est une avancée pour l'analyse de données, toutefois, l'utilisation de ces outils nécessite une expertise afin d'en éviter les écueils et d'obtenir des analyses fiables et pertinentes. L'objectif de cet article est de présenter la pertinence et la mise en œuvre des analyses statistiques non supervisées pour analyser des données multivariées en science de l'éducation. La partie 2 donne une vue d'ensemble des méthodes d'analyses factorielles en présentant leurs objectifs, les concepts associés et les différentes étapes de leur mise en œuvre. La partie 3 aborde les différentes méthodes de clustering avec leurs avantages et inconvénients respectifs. Dans la partie 4, nous illustrons la mise en œuvre de ces méthodes sur des données issues d'une enquête menée sur l'adaptation de primo-entrant·es en IUT. Nous concluons par une réflexion sur l'apport potentiel des analyses non supervisées pour l'étude de données multivariées en science de l'éducation.

## 2. Les analyses factorielles pour l'analyse non-supervisée de variables

Dans cette partie, nous présentons les fondements statistiques de l'analyse factorielle. Les notions présentées sont illustrées sur les données de l'enquête menée sur l'adaptation des primo-entrant·es en IUT.

### 2.1. Définition et concepts clés

Les analyses factorielles sont des techniques statistiques utilisées pour explorer des liaisons entre des variables à l'aide d'un ensemble de données. Elles visent à résumer la structure sous-jacente des

liaisons entre variables à l'aide d'un nombre plus restreint de variables non observées, appelés « facteurs ». Ces facteurs représentent des dimensions intrinsèques qui expliquent la covariance entre les variables observées. L'objectif principal des analyses factorielles est ainsi de faciliter l'interprétation des résultats en fournissant une représentation simplifiée et structurée des données (Fabrigar et Wegener, 2011 ; Mulaik, 2009 ; Escofier et Pagès, 1998 ; Lebart et al., 1995).

Avant d'expliquer le principe des différentes analyses factorielles, il convient de rappeler ses principaux concepts.

**Variations :** lorsque l'on observe $p$ variables quantitatives sur $n$ individus, les variations du nuage de $n$ points dans $\mathbb{R}^p$ peuvent s'expliquer par des différences entre les individus, ou par des liaisons entre les variables. Si, par exemple, on observe $p$ questions d'un questionnaire administré auprès de $n$ étudiant·es, les variations du nuage peuvent être dues soit aux différences entre individus –deux étudiant·es peuvent, selon leur parcours et leurs pratiques d'étude, répondre différemment à une question sur l'adaptation académique ; soit aux liaisons entre les réponses aux questions –une question sur l'adaptation à l'allongement du temps de transport pour se rendre sur le lieu d'étude peut être fortement corrélée à une question sur l'adaptation académique. Les variations sont donc sources d'informations : elles mettent en évidence des différences ou similarités entre les individus, ou des liens entre les variables observées sur les individus. Les variations sont mesurées par les variances des variables (une forte variance pour une variable indique que les réponses des individus à cette question sont très hétérogènes) et par les corrélations entre les variables (une forte corrélation positive entre deux variables indique que ces variables ont tendance à avoir des valeurs similaires, une forte corrélation négative qu'elles ont tendance à avoir des valeurs opposées et une faible corrélation qu'il n'y a pas de lien de linéarité entre les variables). L'étude des variations est ainsi un point central dans l'analyse de données multivariées.

**Erreur de mesure :** tout instrument de mesure est susceptible d'entraîner des erreurs de mesure, c-à-d un écart entre ce qui est mesuré $Y$ (les réponses aux questions sur l'adaptation académique) et ce

que l'on cherche $X$ (le niveau d'adaptation académique que l'on ne peut pas obtenir avec une simple question). On a alors :

$$Y = X + (Y - X) = X + E,$$

où $E$ est l'erreur de mesure, appelée également bruit résiduel.

On distingue deux types d'erreur : le *biais* qui est l'erreur systématique due à l'outil de mesure et *l'erreur aléatoire* liée à utilisation de l'outil (Roussel, 2005). Mesurer la qualité d'un instrument de mesure (un questionnaire par exemple), consiste donc à :

- **valider sa structure** : l'instrument permet-il de mesurer sans biais ce qu'il est censé mesurer ? D'un point de vue statistique, il s'agit de vérifier que la moyenne du bruit $E$ est nulle.
- **évaluer sa fiabilité** : l'erreur de mesure aléatoire lors des diverses utilisations de l'instrument de mesure est-elle négligeable par rapport à ce que l'on cherche à mesurer ? D'un point de vue statistique, il s'agit de vérifier que la variance du bruit $E$ est faible devant celle de $X$.

**Facteurs :** Le phénomène étudié peut être mesuré par la variable $X$ qui peut s'écrire comme une fonction $\Psi$ de variables $F_1, \cdots, F_k$, que l'on nomme « facteurs » :

$$\mathrm{X} = \Psi(F_1, \cdots, F_k).$$

La fonction $\Psi$ est un « modèle factoriel » des données et les facteurs $F_1, \cdots, F_k$ représentent les dimensions sous-jacentes du phénomène étudié.

**Linéarité :** On parle de modèle linéaire quand la fonction $\Psi$ est linéaire, on a donc : $X = a_1 F_1 + \cdots + a_k F_k$. Le poids $a_{\mathrm{i}}$ est appelé « poids factoriel » et mesure la contribution du facteur $F_{\mathrm{i}}$ à expliquer les variations de $X$.

**Variance expliquée :** La variance des observations $Y$ appelée « variance totale » se décompose en la somme d'une part de la variance du modèle $\Psi(F_1, \cdots, F_k)$ appelée « variance expliquée par le modèle » et d'autre part la variance des erreurs $E$ appelée « variance non-expliquée » ou variance du bruit résiduel. Le rapport entre la « variance expliquée » et la « variance totale » permet d'évaluer la

proportion d'information capturée par les facteurs et ainsi, la propension du modèle à pouvoir expliquer l'information intrinsèque des données.

### 2.2. Les différentes analyses factorielles

On distingue quatre types d'analyse factorielle, selon l'objectif poursuivi :

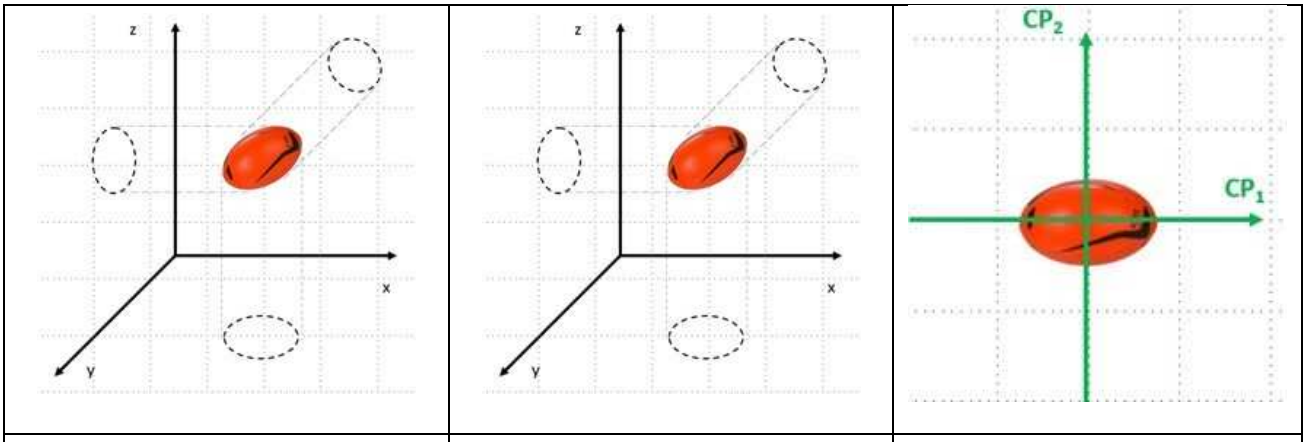
**L'analyse en composantes principales (ACP)** : elle vise à réduire la dimension des données en conservant un maximum d'information. Pour cela, l'ACP construit $p$ facteurs ($CP_i$) non corrélés, chacun étant une combinaison linéaire des $p$ variables observées. Les facteurs sont appelés « composantes principales ». Le modèle factoriel est le suivant :

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

**Equation 1 :** Modèle factoriel d'une ACP

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). En pratique, les composantes principales sont obtenues en diagonalisant la matrice des 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. On obtient ainsi $p$ composantes principales expliquant les $p$ variables initiales. L'intérêt principal de l'ACP est de permettre de réduire la dimension du phénomène étudié tout en limitant la perte d'information : il suffit pour cela de sélectionner les $k$ premières composantes expliquant plus de 95% de la variance totale des $p$ variables initiales (avec $k < p$ ). L'information des données obtenues à partir des $p$ variables initiales est ainsi expliquée à l'aide d'un nombre réduit $k$ de « composantes principales ». Les $p - k + 1$ composantes non sélectionnées expliquant moins de 5% de la variance totale, peuvent être associées au bruit lié à l'échantillon. Se limiter aux $k$ premières composantes expliquant 95% de la variance totale rend ainsi les analyses plus robustes au bruit d'échantillonnage.

Par ailleurs le nuage des $n$ individus est dans $\mathbb{R}^p$. L'ACP permet alors de visualiser ce nuage dans un plan défini par deux composantes, tout en minimisant les distorsions (voir Figure 1 ci-dessous) :




| a) | b) | c) |
|---|---|---|

**Figure 1 :** 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.

Cette visualisation est très utile pour repérer d'éventuelles données aberrantes ou des groupes d'individus qui présentent des similarités latentes.

**L'analyse factorielle multiple (AFM) :** Si l'ACP permet d'analyser les corrélations entre les variables, de réduire la dimension et de visualiser le nuage des individus, 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 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. 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. En effet, si un groupe est composé de dix questions et un autre de deux questions, alors le groupe de dix questions va avoir, 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 inter-groupes. S'agissant d'une ACP, elle permet en outre de réduire la dimension du phénomène étudié et de réduire le bruit d'échantillonnage en se limitant à une partie des composantes principales. Enfin, elle permet de visualiser le nuage des individus tout en étudiant leurs similarités/différences à travers les groupes de variables.

**L'analyse factorielle exploratoire (AFE) :** Si l'ACP cherche à expliquer les variables observées $(Y_i)$ à l'aide de $k$ composantes principales maximisant la variance totale, l'AFE cherche à expliquer la structure des corrélations entres les $(Y_i)$ à l'aide de $k$ facteurs $(F_{\mathrm{i}}, 1 \leq i \leq k)$. Le modèle factoriel est alors plus complexe que celui de l'ACP car il introduit du bruit (Equation 2) :

$$\begin{array}{lll} Y_1 & = a_{11}F_1 + \cdots + a_{1k}F_k & +E_1 \\ \vdots & \vdots & \vdots \\ Y_i & = a_{i1}F_1 + \cdots + a_{ik}F_k & +E_i \\ \vdots & \vdots & \vdots \\ Y_p & = a_{p1}F_1 + \cdots + a_{pk}F_k & +E_p \end{array}$$

**Equation 2 :** Modèle factoriel d'une AFE

Dans l'AFE, le poids factoriel $a_{ij}$ explique la corrélation entre $Y_{\mathrm{i}}$ et le facteur $F_j$, indépendamment du bruit résiduel. Les facteurs $(F_{\mathrm{i}})$ sont des variables « latentes » correspondant à différentes dimensions du phénomène étudié. Par ailleurs, dans l'ACP les composantes principales sont construites afin d'être indépendantes, alors que dans l'AFE les facteurs sont susceptibles d'être corrélés entre eux. Notons que contrairement aux composantes principales de l'ACP, les facteurs de l'AFE ne permettent pas de reconstruire les variables mais uniquement leurs corrélations. Dans le cas d'un questionnaire pour lequel on dispose d'une structure *a priori,* l'AFE permet de valider le questionnaire en vérifiant que la structure *a priori* correspond à la structure obtenue *a posteriori* par l'AFE (voir partie 4).

**L'analyse factorielle confirmatoire (AFC) :** Si l'objectif de l'AFE est d'explorer les données afin d'identifier une structure factorielle *a posteriori*, l'AFC vise à valider une structure factorielle *a priori* à partir des données. Elle implique de spécifier un modèle factoriel avec un nombre choisi de facteurs, et de comparer les données observées avec le modèle spécifié à l'aide de tests statistiques. L'AFC est

souvent utilisée pour valider des modèles obtenus au préalable avec une AFE sur un autre jeu de données.

### 2.3. Hypothèses et conditions d'application de l'analyse factorielle

Différentes hypothèses et conditions sont à prendre en compte lors de la mise en œuvre de l'analyse factorielle afin que l'interprétation des résultats obtenus soit fiable :

**Linéarité :** on suppose que les relations entre les variables observées et les facteurs sont linéaires.

**Corrélation :** on suppose que les variables observées ne sont pas toutes indépendantes ou tout au moins qu'il existe un nombre non négligeable de variables corrélées.

**Normalité :** l'analyse factorielle ne requiert pas nécessairement l'hypothèse que les variables soient distribuées selon une loi normale. Toutefois, la normalité des données peut faciliter la mise en œuvre.

### 2.4. L'Analyse Factorielle Exploratoire pour identifier la structure latente des variables

L'analyse factorielle exploratoire vise à identifier la structure latente des données. Sa mise en œuvre procède en plusieurs étapes qui requièrent chacune différentes modélisations et méthodologiques statistiques (Howard, 2016 ; Mulaik, 2009). Les principales étapes sont les suivantes :

#### Etape 1 : Vérifier l'adéquation des données

Le modèle factoriel de l'AFE suppose que les corrélations des 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 analyse factorielle. La première étape consiste donc à vérifier que les données se prêtent bien à une analyse factorielle. 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-à-d que la matrice de corrélation est proche de la matrice identité. Le test vérifie alors que la valeur absolue du déterminant de la matrice de corrélation est significativement différente de 1.

- **l'indice de Kaiser-Meyer-Olkin (KMO) :** pour vérifier qu'une variable est corrélée avec les autres variables on calcule 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 (Fabrigar et Wegener, 2011).

### Etape 2 : Extraire les facteurs

Une structure factorielle est dimensionnée par $k$, le nombre de facteurs latents et est définie par les poids factoriels $(a_{ij})$. L'extraction des facteurs consiste alors à estimer les poids factoriels $(a_{ij})$ à partir des données, pour une valeur de $k$ fixé. On verra, dans l'étape qui suit, comment sélectionner la valeur « optimale » de $k$. De nombreuses techniques d'estimation des poids factoriels sont disponibles dans les logiciels statistiques. Deux méthodes se distinguent cependant par leur efficacité à estimer les poids factoriels même pour des faibles valeurs de l'indice KMO pour le jeu de données (Mulaik, 2009 p. 52) : la méthode du maximum de vraisemblance (MV) qui peut être utilisée dans le cas où les données sont distribuées selon une loi normale et la méthode des axes principaux (PA) dans le cas contraire.

### Etape 3 : Sélectionner une structure factorielle

Pour chaque valeur du nombre $k$ de facteurs sélectionnés, on obtient un modèle estimé, c-à-d **une structure factorielle latente *a posteriori*** (voir exemple Figure 2).

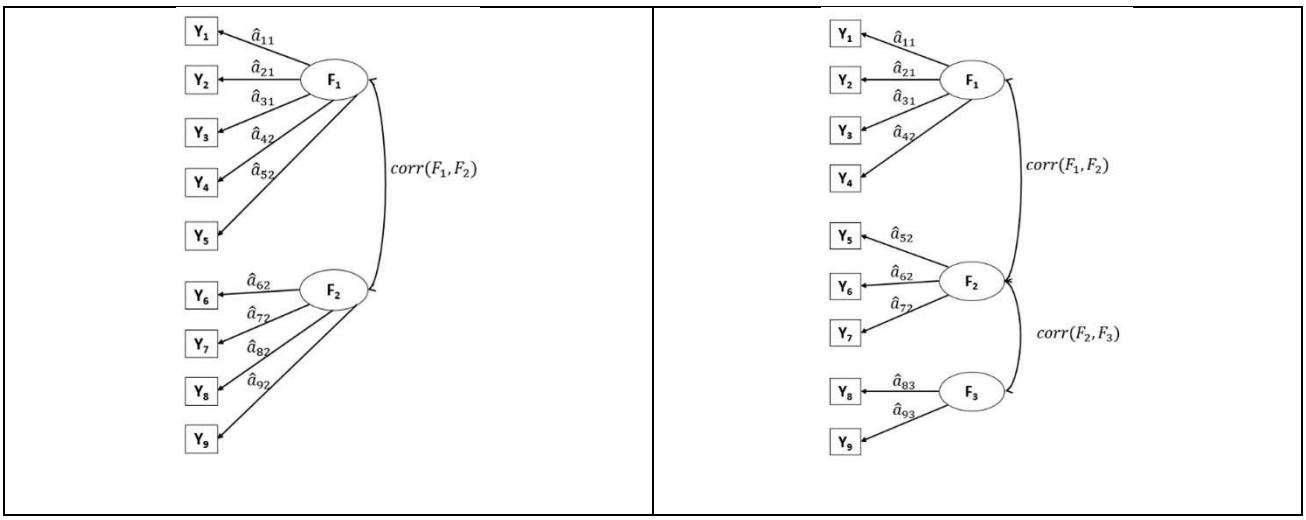


**Figure 2 :** Exemples de structures factorielles à neuf variables. Les poids factoriels ont été estimés à partir du même jeu de données. A gauche une structure à deux facteurs, à droite une structure à trois facteurs.

Comment choisir le « bon » modèle, c-à-d la bonne valeur de $k$ ? Différents critères de sélection sont disponibles dans les logiciels statistiques. Cependant, ces critères visant des objectifs différents, les logiciels peuvent recommander différentes valeurs de $k$. Les conditions d'utilisation de ces différents critères sont présentées ci-après.

- Dans le cas de données « gaussiennes » (les variables sont distribuées selon une loi normale), le nombre optimal de facteurs latents est obtenu à l'aide de la minimisation d'un critère biais-variance : le nombre optimal de facteurs réalise un compromis entre le biais de modélisation (c-à-d « l'erreur » introduite par une modélisation trop simple) et la variance de modélisation (c-à-d « 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'Akaike (AIC) (Saporta, 2006). Rappelons que ces critères nécessitent l'hypothèse de normalité des données.
- Sans l'hypothèse de normalité, deux critères non-paramétriques peuvent être obtenus à partir des valeurs propres de la matrice de corrélation, qui correspondent à la quantité de variance expliquée par chaque facteur extrait :
  - l'Empirical Kaiser Criterion (EKC) : ce critère ne retient un facteur que si sa part de variance expliquée est supérieure à celle expliquée par une seule variable ;
  - l'analyse parallèle (PA) : à l'instar du critère EKC, l'analyse parallèle 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.

Que faire quand ces différents critères ne donnent pas la même structure optimale ? « Tous les modèles sont faux, mais certains sont utiles » est un aphorisme courant en statistique. L'objectif de l'analyse factorielle exploratoire est d'obtenir des facteurs expliquant les corrélations entre les variables mesurées : la structure à retenir est alors celle dont le nombre de facteurs réduit substantiellement la

dimension du phénomène étudié (utilité computationnelle) et dont les facteurs sont les plus interprétables au regard des variables considérées (utilité conceptuelle). Mulaik (2009) parle ainsi du « nombre approprié » de facteurs.

### Etape 4 : Effectuer une rotation des facteurs

Lorsque les facteurs extraits ne sont pas clairement interprétables au regard des variables considérées, on effectue une rotation des facteurs. Cela permet de minimiser d'une part le nombre de variables liées à un facteur et d'autre part le nombre de facteurs liés à une même variable. Il existe deux types de rotations : orthogonale ou oblique. La rotation orthogonale est utilisée lorsque l'on cherche à obtenir des facteurs latents indépendants, la rotation oblique lorsque l'on cherche à obtenir des facteurs latents corrélés.

### Etape 5 : Nommer les facteurs

Cette étape consiste à identifier le « thème » lié à un facteur à partir des variables qui lui sont liées. Dans le cas où les données sont des réponses à un questionnaire (comme présenté en partie 4), cette étape correspond à valider le questionnaire. Il s'agit de vérifier que les questions du questionnaire mesurent sans biais ce qu'elles sont censées mesurer ; cela se traduit par le fait que les variables corrélées à un facteur mesurent le même thème. Les variables qui ne sont liées à aucun facteur peuvent être éventuellement éliminées du questionnaire. Une fois ce processus effectué, le questionnaire est considéré comme « validé »

### Etape 6 : Vérifier la fiabilité du modèle factoriel

Une fois le modèle sélectionné, il reste à vérifier qu'il est fiable, c-à-d que les résultats obtenus sont robustes vis-à-vis de l'échantillon. Dans le cas d'un questionnaire, cela signifie que les résultats obtenus lors de différentes passations du questionnaire seront consistants.

Reprenons le modèle de l'erreur de mesure :

$$Y = X + (Y - X) = X + E,$$

dans lequel $X$ est la valeur que l'on cherche, $Y$ la mesure et $E$ l'erreur de mesure. Le questionnaire est consistant si la variance de $Y$ et de $X$ sont proches, ce que l'on peut mesurer à l'aide du rapport :

$$\rho^2 = \frac{Var(X)}{Var(Y)} \simeq 1.$$

Il est cependant impossible que connaître la variance de $X$. Les divers indices de consistance visent ainsi à estimer le coefficient $\rho^2$. Le plus connu de ces indices est l'alpha de Cronbach. Malheureusement, ce coefficient présente de nombreux inconvénients : il faut supposer que les erreurs de mesure sont homogènes entre les questions, ce qui n'est pas réaliste. Par ailleurs il peut être influencé par le nombre d'items dans l'échelle : des questionnaires avec un grand nombre d'items tendent à avoir des valeurs d'alpha plus élevées. Il est alors préférable d'utiliser l'indice Omega de Mc Donald ou l'indice Lambda-6 de Guttman (voir Bourque et al., 2019 pour une étude détaillée).

## 3. Le clustering pour l'analyse non-supervisé des individus

Comme nous l'avons vu précédemment, l'analyse factorielle exploratoire permet de décrire la structure cachée des variables à l'aide de facteurs expliquant chacun un groupe de variables. Dans l'analyse de tendances émergentes, il peut également être utile d'identifier des groupes d'individus significatifs et pertinents au regard du phénomène étudié. Dans notre cas, on peut s'interroger sur une éventuelle hétérogénéité des primo-entrant·es en termes de capacité à s'adapter à l'IUT : existe-t-il des groupes d'individus similaires ou différents en termes d'adaptation ? Pour cela, on utilise les méthodes de clustering (Lebart et al., 1995).

Le principe du clustering et de construire des clusters –ou groupe 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.

Les méthodes de clustering les plus couramment utilisées sont l'algorithme des $k$-means et la Classification Ascendante Hiérarchique : ces méthodes utilisent une distance entre individus et nécessite donc des variables quantitatives ou tout au moins une conversion des variables qualitatives en variables numériques en effectuant, par exemple, une Analyse des Correspondances Multiples (Saporta 2006 ; Lebart et al., 1995).

- **L'algorithme des $k$-means :** le nombre de clusters $k$ est spécifié à l'avance. Au départ $k$ indvidus sont choisis et constituent les centres des clusters. Puis les autres individus sont assignés au cluster minimisant la distance avec les centres des clusters. Une fois la partition faite, le point moyen des clusters 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 clusters, 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 prio*ri le nombre $k$ de clusters visés.
- **La Classification Ascendante Hiérarchique (CAH) :** au départ, chaque individu est considéré comme un cluster distinct. Les clusters sont ensuite fusionnés itérativement jusqu'à obtenir un seul cluster qui contient tous les individus. Plusieurs principes de fusion existent. Le principe de Ward est le plus populaire : il consiste à fusionner les clusters qui minimisent la variance intra-cluster après fusion. La CAH produit une structure hiérarchique, ce qui permet une visualisation claire des relations entre les différents clusters à l'aide d'un dendrogramme. Elle ne nécessite pas d'*a priori* sur le nombre de clusters. L'avantage de cette méthode est qu'elle permet d'identifier le nombre optimal de clusters. En revanche, la CAH est plus exigeante en termes 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 de clusters approprié, puis en utilisant la configuration obtenue par la CAH pour initialiser l'algorithme des $k$-means.

## 4. Etude de l'adaptation des primo-entrant·es en IUT : quels constats suite aux 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 lycéen·nes intégrant les IUT, ainsi que les contenus et les méthodes d'évaluation des apprentissages Ces changements ont des répercussions sur l'adaptation académique des étudiant·es. Il est donc est apparu crucial pour les équipes pédagogiques d'analyser l'adaptation des primo-entrant·es 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 deux formations dans un IUT d'Ile de France.

### 4.1. L'enquête et le jeu de données

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

- Adaptation académique : évalue l'ajustement académique et la réussite académique des étudiant·es ;
- Adaptation sociale : évalue l'ajustement social et les relations avec les pairs et également avec l'équipe pédagogique ;
- Adaptation émotionnelle : évalue le bien-être émotionnel et psychologique des étudiant·es ;
- Adaptation personnelle : évalue l'ajustement personnel de l'étudiant·e, c-à-d 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 ; 343 étudiant·es ont répondu. Pour chaque question, les étudiant·es étaient invité·es à se positionner sur une proposition visant à évaluer leur capacité d'adaptation sur les différents thèmes, en utilisant une échelle de Lickert à 5 valeurs, 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 ».

### 4.2. Mise en œuvre de l'analyse factorielle

L'AFE a été mise en œuvre sur ces données à l'aide du package R Psych (Revelle, 2021).

#### Etape 1 : Vérifier l'adéquation des données

Sur ces données, le test de sphéricité de Bartlett valide l'hypothèse que les items ne sont pas tous corrélés, avec une p-value <0,001. L'indice KMO global est quand-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) :

- 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 kurtosis multivarié (coefficient d'aplatissement) des données est significativement différent de celui d'une distribution normale.

Les résultats de ces deux tests sur les données permettent de conclure au rejet de l'hypothèse de normalité des données étudiées, avec une p-value inférieure à 0,01.

### Etape 2 : Extraire les facteurs

Les données n'étant pas gaussiennes, les facteurs ont été extraits en utilisant la méthode des axes principaux (AP). Selon le critère utilisé, on obtient une structure à cinq, six ou sept facteurs(voir Tableau 1).

**Tableau 1 :** Nombre de facteurs retenus selon le critère de sélection utilisé

| **Critère d'optimalité** | **Nombre de facteurs retenus** |
|---|---|
| Akaike Information Criterion | 6 |
| Empirical Kaiser criterion | 7 |
| Parallel Analysis | 5 |

Le graphe des structures factorielles latentes des modèles à cinq et six facteurs sont présentés dans la figure 3 ci-après. Sur les graphes, les poids factoriels (part de variance de la variable expliquée par le facteur) sont indiqués sur les flèches des facteurs vers les variables : par exemple la part de variance de la variable AA_Choix, expliquée par le facteur PA1 dans le modèle à cinq facteurs (Figure 3 à gauche), est 0,9. Seules les valeurs supérieures à 0,3 sont indiquées. Les valeurs sur les flèches entre deux facteurs correspondent à la corrélation entre les deux facteurs : par exemple la corrélation entre les facteurs PA1 et PA3 sur le modèle à cinq facteurs (Figure 3 à gauche) est 0,4. Le modèle à sept facteurs n'est pas présenté car il est moins parcimonieux que les deux autres, ce qui va à l'encontre de l'objectif recherché de simplification.

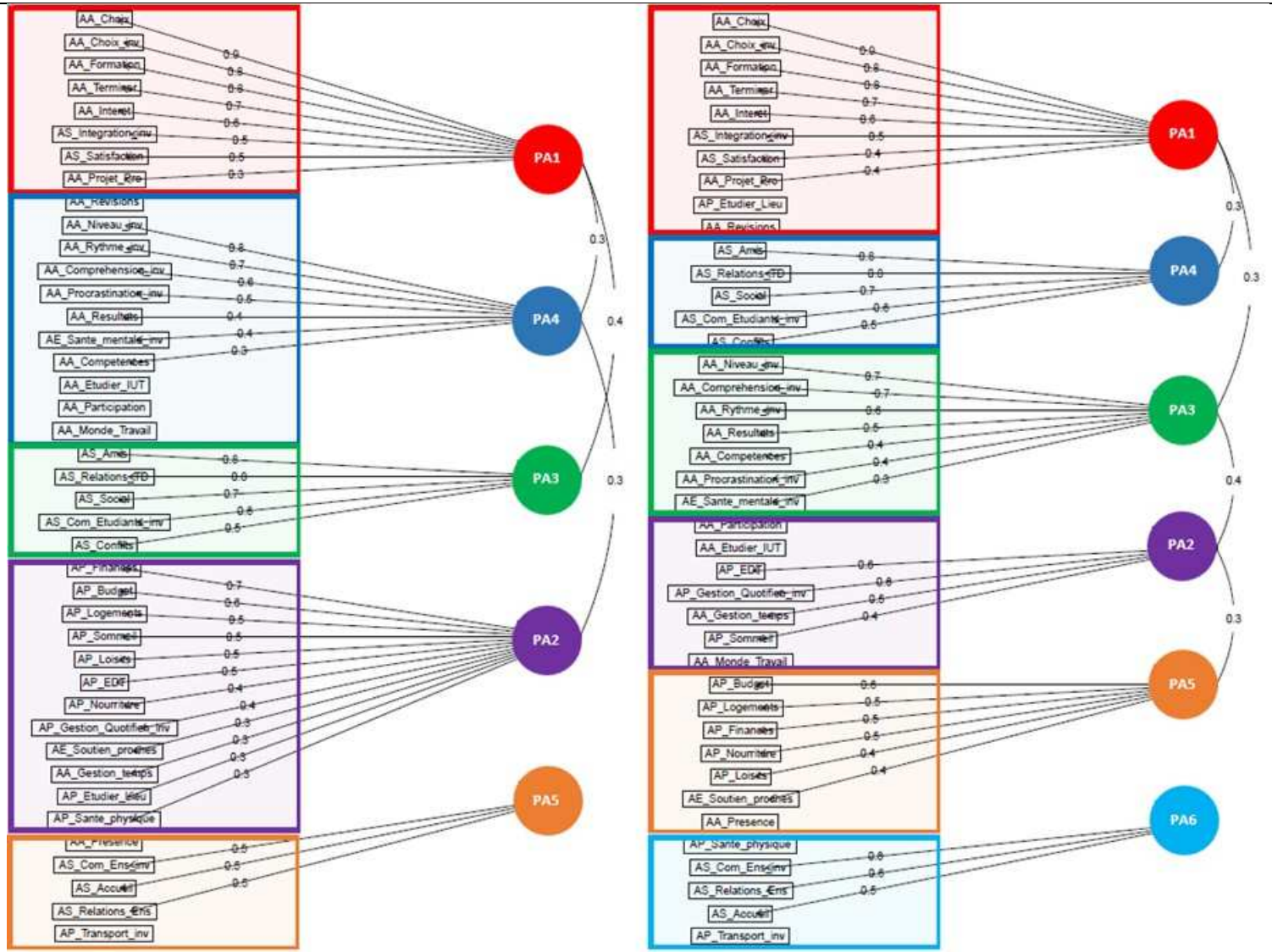



**Figure 3** : structures factorielles des modèles à 5 et 6 facteurs. Les poids sur les flèches des facteurs vers les variables correspondent à la part de variance de la variable expliquée par le facteur. Les poids sur les flèches entre deux facteurs correspondent à la corrélation entre les deux facteurs.

### Etape 3 : Sélectionner la structure factorielle

Les modèles à cinq ou six facteurs ont des consistances internes similaires, comme cela est présenté dans le tableau 2 ci-dessous :

**Tableau 2 : Indices de fiabilité par modèle**

| **Modèle à 5 facteurs** | | **Modèles à 6 facteurs** | |
|---|---|---|---|
| Alpha de Cronbach | 0,91 | Alpha de Cronbach | 0,91 |
| Lambda G.6 | 0,94 | Lambda G.6 | 0,94 |
| Omega Total | 0,94 | Omega Total | 0,93 |
| Omega partiels | | Omega partiels | |
| max | 0,85 | max | 0,88 |
| min | 0,8 | min | 0,77 |

Nous avons sélectionné le modèle à cinq facteurs car d'une part il est plus parcimonieux, et d'autre part, les facteurs sont facilement interprétables au regard du phénomène étudié.

### Etape 4 : Effectuer une rotation

Cette étape n'est pas nécessaire car les facteurs extraits sont interprétables facilement.

### Etape 5 : Nommer les facteurs et valider le questionnaire

L'étude des variables liées aux facteurs permet d'identifier et de nommer les facteurs latents :

- **Le facteur PA1** est lié aux questions concernant la formation, comme par exemple :
  AA_Choix : « Je suis satisfait·e de mon choix de formation » ;
  AA_Formation : « La formation que je suis correspond à mes attentes » ;
  AS_Integration_inv : « J'ai l'impression de ne pas avoir ma place dans ma formation » ;
  AS_Satisfaction : « Je suis globalement satisfait·e de mon expérience à l'IUT » ;
  Nous le nommons **Adaptation à la Formation (AF).**

- **Le facteur PA4** est lié aux questions concernant la performance académique, comme par exemple :
  AA_Revisions : « Je me suis bien préparé·e pour les examens » ;
  AA_Niveau_inv : « Je me sens dépassé·e par le niveau de difficulté des enseignements » ;
  AA_Resultats : « Je suis satisfait·e de mes résultats au premier semestre » ;
  Nous le nommons **Ajustement Académique (AA).**

- **Le facteur PA3** est lié aux questions concernant les interactions avec les pairs, comme 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 PA2** est lié aux questions concernant l'adaptation personnelle à la gestion du quotidien :
  AP_Finances : « Je suis satisfait·e de ma situation financière » ;
  AP_Logement : « Je suis satisfait·e de la qualité de mon logement » ;
  AP_Gestion_Quotidien_inv: « Je rencontre des difficultés à gérer mon quotidien (repas, RDV médicaux, démarches administratives…) » ;

AE_Soutien_proches : « Je me sens soutenu·e par mes proches, famille et amis » ;
AA_Gestion_Temps : « Je suis satisfait·e de ma capacité à gérer mon temps » ;
Nous le nommons **Adaptation Personnelle (AP).**

- **Le facteur PA5** est lié aux questions concernant les interactions avec les enseignant·es, comme 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 » ;
  Nous le nommons **Interactions avec les Enseignants (IE)**.

Notons que la question : AP_Transport_inv : « Le temps que je passe dans les transport 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, la tolérance aux retards est à la discrétion des enseignant·es.

A partir des données, la décomposition *a priori* en quatre thèmes du questionnaire SACQ a ainsi évoluée en un modèle à cinq facteurs.

### 4.3. Mise en œuvre du clustering

Une fois la structure factorielle latente identifiée, nous avons mis en œuvre un clustering pour identifier des profils de primo-entrant·es en termes d'adaptation à l'IUT. Toutefois, l'existence de groupes de variables liées aux facteurs conduit à vouloir chercher des similarités entre étudiant·es :

- au travers de l'ensemble des variables ;
- mais également au travers de chacun des groupes de variables liées : deux étudiant·es ayant des réponses similaires en termes d'adaptation personnelle (facteur PA2), ont-ils des réponses similaires ou différentes pour un autre facteur ?

La composition des groupes de variables liées n'est pas uniforme ici : par exemple le facteur PA2 est lié à 12 variables alors que le facteur PA5 n'est lui lié qu'à 5 variables. L'impact du groupe PA2 est donc artificiellement plus important que celui du groupe PA2 dans la mesure de la similarité entre

deux étudiant·es. Afin d'équilibrer l'influence de chaque groupe de variables lors du clustering, nous avons tout d'abord mis en œuvre une Analyse Factorielle Multiple (AFM).

### Mise en œuvre de l'AFM

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

Les 33 premières composantes principales représentent plus de 95% de la variance totale des 41 variables (voir Tableau 3). Cela met en évidence la capacité des analyses factorielles à réduire la dimension des données et à éliminer le bruit d'échantillonnage.

**Tableau 3 : Composantes de l'AFM**

| Composante | % cumulé de variance expliquée |
|---|---|
| 1 | 23,1% |
| 2 | 30,0% |
| 3 | 36,2% |
| ⋮ | ⋮ |
| 16 | 75,4 % |
| ⋮ | ⋮ |
| 33 | 95,5% |

Il est important de souligner que les dernières composantes principales contiennent peu d'informations significatives ; elles peuvent par conséquent être considérées comme du bruit d'observation. Ignorer ces composantes permet de rendre le clustering plus robuste vis-à-vis de l'échantillon étudié.

L'AFM a ainsi permis de réduire la dimension du modèle à analyser (de 41 à 33 variables), et de le rendre plus robuste au bruit d'échantillonnage.

### Résultats de l'AFM

Les corrélations entre les variables et les composantes principales de l'AFM sont données dans le Tableau 4 : on observe tout d'abord que toutes les variables sont corrélées avec la composante 1 de

l'AFM, on parle d'« effet taille » : une valeur élevée de la composante 1 implique une valeur élevée pour toutes les réponses aux questions. La composante 1 (combinaison linéaire des variables) correspond donc à un score global d'adaptation à l'IUT. Quand on projette le nuage des individus sur le « plan principal de l'AFM », c-à-d le plan constitué de la composante 1 et de la composante 2 (voir Figure 5), un individu ayant une cordonnée élevée sur l'axe de la composante 1 a des valeurs élevées pour toutes les réponses aux questions.

Les variables Performance Académique (PA) et Adaptation Personnelle (AP) sont corrélées positivement avec la composante 2 (Tableau 4) ; les variables Interactions avec les Pairs (IP), Interactions avec les Enseignants (IE) et Adaptation à la Formation (AF) sont, elles, corrélées négativement avec la composante 2. Quand on projette le nuage des individus sur le plan principal de l'AFM (voir Figure 5), un individu ayant une coordonnée élevée sur l'axe de la composante 2, a des valeurs élevées pour les variables des groupes PA et AP ; un individu ayant une coordonnée faible sur l'axe de la composante 2 aura en revanche des valeurs élevées pour les variables des groupes IP, IE et AF.

**Tableau 4 :** Corrélations des variables avec les composantes CP1 et CP2. (pour chaque groupes seules les trois plus corrélées avec CP1 sont indiquées)

| Variables | Corrélation avec CP1 | Corrélation avec CP2 |
|---|---|---|
| IP Social | 6,5 | -4,4 |
| IP_Relations_TD | 5,1 | -4,8 |
| IP_Amis | 4,4 | -5,0 |
| ⋮ | ⋮ | ⋮ |
| **PA_Rythme_inv** | **6,2** | **3,5** |
| **PA_Niveau_inv** | **5,3** | **3,9** |
| **PA_Procrastination_inv** | **4,9** | **3,0** |
| ⋮ | ⋮ | ⋮ |
| AF_Integration_inv | 6,9 | -1,7 |
| AF_Satiffaction | 6,9 | -1,7 |
| AF_Formation | 6,5 | -1,9 |
| ⋮ | ⋮ | ⋮ |
| **AP_EDT** | **5,7** | **3,2** |
| **AP_Gestion_Quotifien_inv** | **5,5** | **3,1** |
| **AP_Sommeil** | **5,1** | **2,2** |
| ⋮ | ⋮ | ⋮ |
| IE_Relations_Ens | 6 | -1, |
| IE_Com_Ens_inv | 5,9 | -0,2 |
| IE_Accueil | 4,7 | -2,4 |
| ⋮ | ⋮ | ⋮ |

## Clustering des individus

Le clustering a été mis en œuvre sur ces données à l'aide du package R Factominer (Husson et al., 2023). Les deux méthodes de clustering présentées dans la partie 3, présentent chacune des avantages et des inconvénients. Nous les avons combinées afin de tirer profit des avantages de chacune : on commence par une CAH pour déterminer le nombre optimal de clusters ; les clusters obtenus par la CAH servent alors à initialiser l'algorithme des $k$-means qui permet d'obtenir des clusters plus homogènes que la CAH et bien séparés.

Le dendrogramme obtenu par la CAH sur les 33 premières composantes de l'AFM est présenté sur la Figure 4 à gauche. L'analyse du gain de variance-intra après fusion (Figure 4 à droite) montre que celui-ci est optimal lorsque l'on passe de quatre à trois clusters. Nous avons donc retenu une structure latente d'individus à trois clusters.

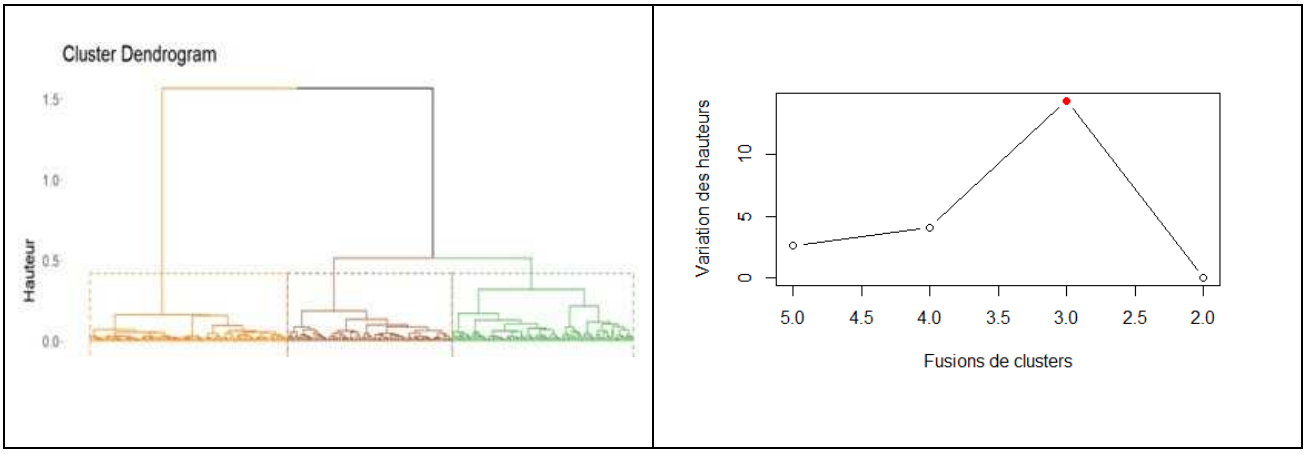



**Figure 4 :** A gauche, dendrogramme de la CAH sur les 33 premières composantes principales. En ordonnée est représentée la variance-intra après fusion. A droite, le gain de variance-intra en fonction du nombre de clusters : le gain est optimal en passant de quatre à trois clusters.

Nous avons ensuite utilisé l'algorithme des $k$-means avec $k = 3$. La figure 5 présente les trois clusters obtenus projetés sur le plan principal de l'AFM.

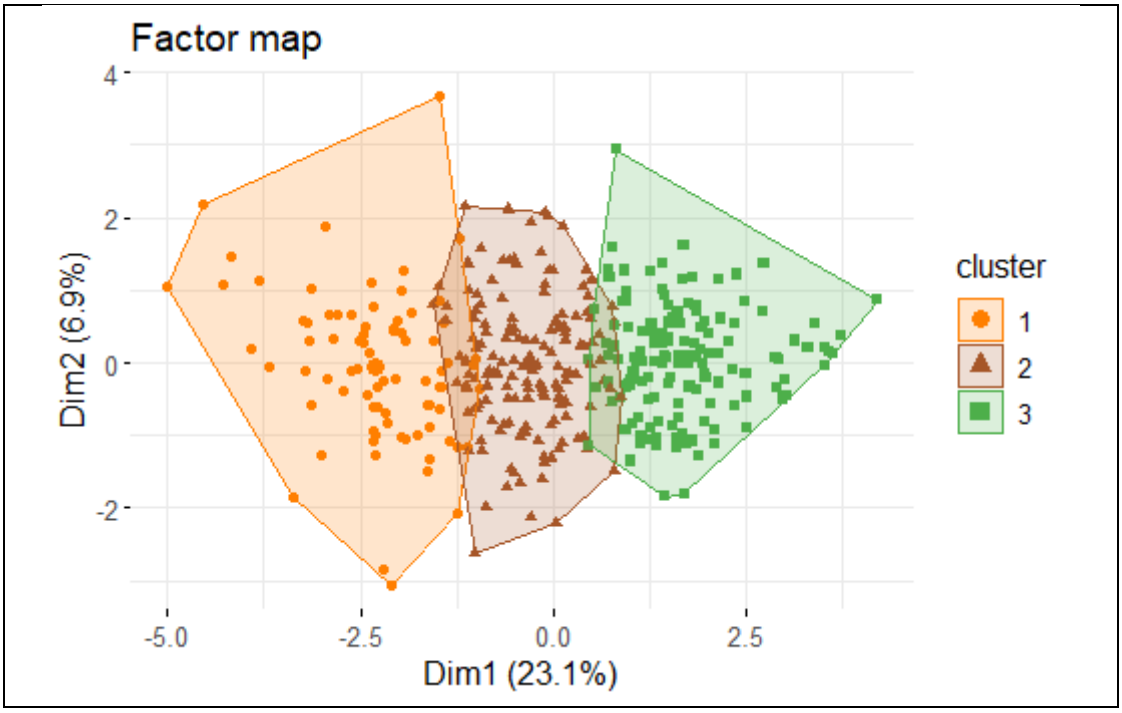



**Figure 5 :** projection des trois clusters sur le premier plan de l'AFM

### 4.4. Analyse des clusters obtenus

Rappelons que la composante 1 de l'AFM reflète un score global d'adaptation à l'IUT : un·e étudiant·e ayant une coordonnée élevée sur la composante 1 a des valeurs élevées pour l'ensemble des réponses aux questions. La projection des clusters sur le plan principal de l'AFM (Figure 5 ci-dessus), laisse à penser que les clusters obtenus correspondent à des groupes d'individus ayant des scores d'adaptation plus ou moins élevés.

Pour caractériser les individus de ces clusters, une comparaison est réalisée en évaluant les moyennes des variables pour chacun des cinq groupes de variables issus de l'AFE. Le test de comparaison de Kruskall-Wallis (Lebart et al., 1995) a été utilisé pour identifier les écarts significatifs entre les groupes. Les résultats obtenus sont présentés dans le tableau 5 ci-après. Chaque ligne du tableau 5 est consacrée à un groupe de variables (Adaptation Personnelle, Interaction avec les Pairs, Interaction avec les Enseignants, Performance Académique et Adaptation à la Formation) :

- le graphique dans la colonne de gauche représente les boîtes-à-moustaches des moyennes des variables du groupe de variables en question, et ce pour chacun des trois clusters (cluster1 à gauche, cluster2 au milieu et cluster3 à droite). Le trait pointillé rouge représente

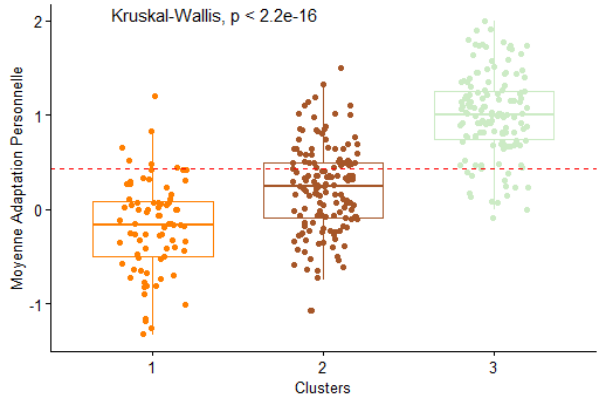
la valeur de la moyenne de toutes les variables du groupe en question pour l'ensemble des personnes interrogées. Elle sert de référence à l'hypothèse « pas de différence en moyenne entre les clusters ». Une *p-value* du test de comparaison de Kruskall-Wallis inférieure à 0,05, confirme que la différence entre les moyennes des clusters est significative.

- le tableau de colonne de droite indique la moyenne pour l'ensemble des étudiants interrogées (symbolisée par les traits pointillés rouge du graphe), la moyenne par cluster et la valeur *z-test* : il s'agit de l'écart normalisé entre la moyenne sur le cluster et la moyenne sur l'ensemble. Lorsque la valeur 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 (c-à-d valeur entre -1,96 et 1,96) l'écart entre les moyennes n'est pas significatif.

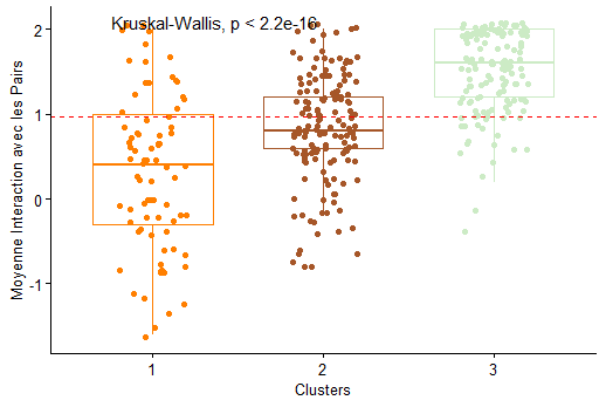
**Tableau 5 :** Caractérisation des trois clusters pour chaque groupe de variables résultant de l'AFE.

Sur les graphiques à gauche figurent pour chaque cluster les boîtes-à-moustaches des moyennes de l'ensemble des variables du groupe de variable. Le trait pointillé rouge représente la valeur de la moyenne sur l'ensemble de la population interrogée. Une p-value du test de comparaison de Kruskall-Wallis inférieure à 0,05, confirme que la différence entre les moyennes des clusters est significative.

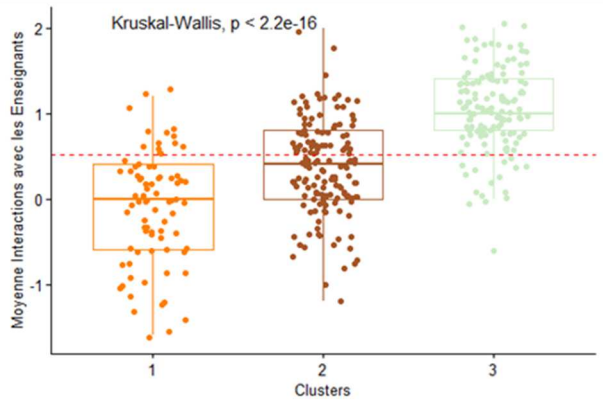
Les tableaux de droite indiquent la moyenne globale, la moyenne par cluster et la valeur z-test. Lorsque la valeur est inférieure à -1,96 (resp. supérieure à 1,96), on peut conclure que la moyenne du groupe est significativement inférieure (resp. supérieure) à celle de l'ensemble de la population interrogée ; sinon (c-à-d valeur entre -1,96 et 1,96) l'écart entre les moyennes n'est pas significatif.



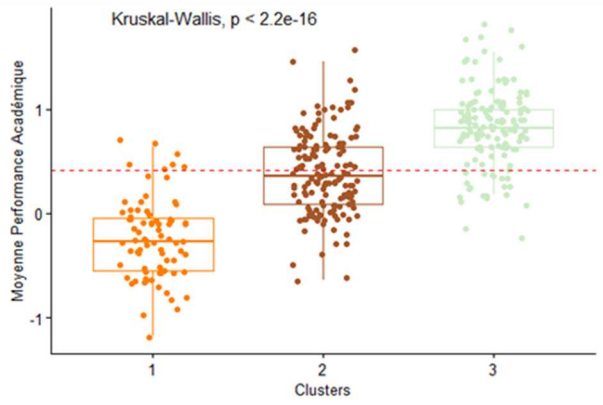

| Adaptation Personnelle (AP) | | |
|---|---|---|
| **Groupe** | **Moyenne** | **z-test** |
| Population | 0,43 | |
| Cluster 1 | -0,18 | -9,20 |
| Cluster 2 | 0,22 | -4,95 |
| Cluster 3 | 1,01 | 12,89 |



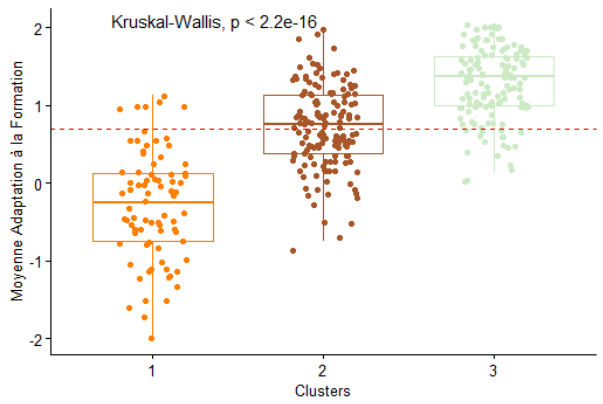

| Interaction avec les Pairs (IP) | | |
|---|---|---|
| **Groupe** | **Moyenne** | **z-test** |
| Population | 0,96 | |
| Cluster 1 | -0.235 | -11,86 |
| Cluster 2 | 0.84 | -2,52 |
| Cluster 3 | 1,48 | 9,47 |




| Interaction avec les Enseignants (IE) | | |
|---|---|---|
| **Groupe** | **Moyenne** | **z-test** |
| Population | 0,51 | |
| Cluster 1 | -0,115 | -8,79 |
| Cluster 2 | 0,36 | -3,45 |
| Cluster 3 | 1,04 | 11,02 |




| Ajustement Académique (AA) | | |
|---|---|---|
| **Groupe** | **Moyenne** | **z-test** |
| Population | 0,42 | |
| Cluster 1 | -0,24 | -11,86 |
| Cluster 2 | 0,38 | -0,97 |
| Cluster 3 | 0,83 | 11,12 |




| Adaptation à la Formation (AF) | | |
|---|---|---|
| **Groupe** | **Moyenne** | **z-test** |
| Population | 0,43 | |
| Cluster 1 | -0,295 | -12,51 |
| Cluster 2 | 0,71 | 0,26 |
| Cluster 3 | 1,27 | 10,41 |

Avec une p-value <0,01, on peut conclure que :

- le Cluster1 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 des étudiant·es qui rencontrent potentiellement des difficultés d'adaptation dans toutes les dimensions de l'AFE ;
- le Cluster2 regroupe des individus 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 étudiant·es 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, on constate qu'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 ;
- le cluster3 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 de ce profil déclarent ne pas rencontrer de difficultés d'adaptation à l'IUT.

## 5. Conclusion et discussion

L'analyse de données multivariées est courante dans de nombreuses disciplines, y compris en science 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 non supervisées 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 données afin de révéler des motifs cachés de variables, d'autre part les méthodes de clustering permettent d'explorer les données afin d'extraire des regroupements d'individus significatifs et pertinents. 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 ce travail est alors de guider les chercheurs en science de l'éducation qui souhaitent les utiliser. Nous avons dans un premier temps présenter les diverses analyses statistiques non supervisées en précisant leurs objectifs et 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-entrant·es dans un IUT, suite aux réformes du bac général en 2019 et du BUT en 2021. Nous avons tout d'abord effectué une Analyse Factorielle Exploratoire pour obtenir une représentation synthétique des relations entre les 41 variables du questionnaire. Nous avons retenu un modèle à cinq facteurs, que nous avons identifié comme étant des dimensions expliquant les difficultés d'adaptation des primo-entrant·es : Adaptation à la Formation, à la Performance Académique, aux Interactions avec les Enseignants, aux Interactions avec les Pairs et à l'Adaptation Personnelle.

Nous avons ensuite effectué un clustering des individus. Dans un premier temps, une Analyse Factorielle Multiple (AFM) a été mise en œuvre en utilisant les groupes de variables liées aux facteurs identifiés lors de l'AFE. L'utilisation des 33 premières composantes de l'AFM, permet d'obtenir un modèle robuste aux fluctuations aléatoires liées à l'échantillon étudié. Nous avons ensuite réalisé un clustering des individus : une CAH a permis de sélectionner le nombre optimal de clusters (trois ici) ; ces clusters ont ensuite été affinés à l'aide de l'algorithme des *k*-means, puis caractérisé en comparant les moyennes des groupes de variables de l'AFE. Les trois clusters obtenus correspondent à des profils différents d'adaptation de primo-entrant·es : un profil d'individus déclarant avoir eu des difficultés d'adaptation dans toutes les dimensions, un profil d'individus déclarant avoir eu des difficultés d'adaptation uniquement pour les dimensions autres qu'académiques et enfin un profil d'individus déclarant n'avoir pas rencontré de difficultés d'adaptation. Si les activités pédagogiques proposées dans les enseignements visent naturellement à améliorer la performance académique des étudiant·es, les résultats de notre étude permettent aux équipes pédagogiques de prendre conscience du fait que les primo-entrant·es en IUT ont également besoin d'activités favorisant leur intégration sociale.

En conclusion, les méthodes d'analyses statistiques non-supervisées se sont montrées efficaces pour identifier les facteurs latents expliquant l'adaptation des primo-entrant·es à l'IUT et mettre à jour divers profils de primo-entrant·es en termes de difficultés d'adaptation. Les résultats contribuent à une compréhension approfondie et complète de l'adaptation de ces primo-entrant·es, permettant ainsi aux équipes pédagogiques de prendre des décisions éclairées et d'élaborer des activités adaptées.

**Apports des analyses non-supervisées aux recherches en sciences de l'éducation**

Les analyses non supervisées constituent un ensemble d'instruments efficaces dans pour identifier et analyser la structure cachée sous-jacente de données multivariées, et permettent d'aborder les problématiques suivantes :

- identifier la structure latente des variables mesurées : les analyses factorielles permettent de quantifier les dimensions latentes de la structure des liaisons des variables et d'identifier les facteurs clés expliquant le phénomène étudié ;
- mettre à jour des profils caractéristiques des individus de l'étude : les méthodes de clustering permettent de segmenter la population étudiée en groupes distincts d'individus similaires. On peut alors construire des indicateurs synthétiques et 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 orienter les chercheurs vers de nouvelles pistes de recherche.

## 6. Références bibliographiques


Bourque, J., Doucet, D., LeBlanc, J., Dupuis, J. et Nadeau, J. (2019). L'alpha de Cronbach est l'un des pires estimateurs de la consistance interne : une étude de simulation. *Revue des sciences de l'éducation*, 45(2), 78-99. https://doi.org/10.7202/1067534ar

Carayon, S. et 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

Escofier, B. et Pagès, J. (1998). *Analyses factorielles simples et multiples*. Dunod, Paris, 284.

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

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., Le, S. et 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

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

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.

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* (pp. 245-276). Louvain-la-Neuve : De Boeck Supérieur

Saporta, G. (2006). *Probabilités, analyse des données et statistique*. Editions technip.