Kamel **Gana**

L'ESSENTIEL DE L'ANALYSE FACTORIELLE EXPLORATOIRE DES PSYCHOLOGUES

Principes de base et applications pratiques

ellipses

Première partie. **Fondements**

L'analyse factorielle est au cœur de la mesure des construits psychologiques.
Jum C. Nunnally & Ira H. Bernstein (1994, p. 111)

Chapitre 1.
Bases conceptuelles de l'analyse factorielle exploratoire

Pour situer nos réflexions et border le cadre de ce livre qui se propose de démystifier et de rendre intelligible et accessible la méthode d'analyse factorielle, dont l'utilisation est souvent galvaudée et caractérisée par une diversité de pratiques aussi vaste que le nombre d'utilisateurs (Goretzko et al., 2021), il est crucial de présenter quelques évidences.

Tout d'abord, l'analyse factorielle des psychologues, initiée par Spearman (1904), est une méthode statistique, mais elle est de toute évidence une théorie aussi, exclusivement au service de la psychométrie (Nunnally & Bernstein, 1994 ; Tucker & MacCallum, 1997). Une théorie relative aux influences et dynamiques qui sous-tendent les covariations et surtout les variations des mesures/items (scores) d'attributs psychologiques (e.g., traits, aptitudes). Comme le soulignent clairement Nunnally et Bernstein (1994) « l'analyse factorielle est intimement liée à la validation… Elle fournit des éléments de preuve utiles concernant les mesures destinées à avoir une validité de contenu » (p. 111), et d'ajouter quelques lignes plus loin « L'analyse factorielle est au cœur de la mesure des construits psychologiques » (p. 111).

Ensuite, elle s'inscrivait plutôt dans une démarche d'épreuve d'hypothèses sur les structures sous-jacentes des données observées. Enfin, elle s'inscrivait dans une démarche davantage explicative que descriptive. Contrairement à l'analyse en composantes principales, qui offre une description condensée des données en les réduisant à un nombre restreint de composantes principales généralement orthogonales entre elles, l'analyse factorielle n'est pas simplement une méthode de condensation des données. Nous y reviendrons dans le chapitre 4. Mais clarifions d'abord quelques concepts fondamentaux en analyse factorielle.

1. Variables manifestes versus variables latentes

Pour fixer les idées, des clarifications conceptuelles sont nécessaires aussi. Deux termes génériques en particulier méritent d'être explicités, à savoir « variables manifestes » et « variables latentes ».

D'abord, les variables manifestes constituent les données observées qui renvoient aux mesures obtenues sur des variables spécifiques, constituant ainsi les observations que le psychologue analyse. Ce sont des variables qui ont fait l'objet de mesure et pour lesquelles des scores sont disponibles (par exemple le score à un test ou un item d'une échelle de mesure, tel que : « Avez-vous des idées suicidaires ? » allant de très rarement = 1 à très fréquemment = 5 ; « Le patient présente-t-il ce symptôme ? » oui = 1, non = 0 ; « Quel est votre âge ? » [...]). Ces variables mesurées sont appelées variables manifestes, en opposition aux variables latentes, ou encore indicateurs de la variable latente. Ainsi, les termes « variable manifeste », « variable observée », « variable mesurée », « indicateur » et « item » sont souvent employés de manière interchangeable en psychométrie. Nous utiliserons dans les pages qui suivent plutôt « variables mesurées/observées » et items.

En revanche, les variables latentes renvoient, elles, aux facteurs latents désignant les éléments inobservables et sous-jacents qui peuvent expliquer l'organisation et les associations des variables manifestes. Ces facteurs latents ou variables latentes, se rapportent généralement aux construits ou sous-construits (*subconstructs*) psychologiques qui offrent un cadre théorique pour comprendre les structures sous-jacentes des données manifestes/observées. Le lecteur trouvera dans De Boeck et al. (2024) une analyse heuristique approfondie de la notion de « construit psychologique » et des différentes approches permettant son opérationnalisation.

Afin de cerner ce à quoi renvoie un construit psychologique, il suffit de répondre à la question suivante : que mesure-t-on en psychologie en général ? Les aptitudes, les attitudes, les traits et les états/émotions. Ces entités abstraites renvoient toutes à des réalités psychologiques intrinsèques qui échappent à l'observation directe ou qui sont difficiles à observer de manière directe. Les construits psychologiques, tels que les aptitudes, les attitudes, les traits de personnalité et les états émotionnels, sont des abstractions théoriques qui ne peuvent pas être appréhendées directement. Ils nécessitent l'utilisation de mesures, souvent sous la forme de tests, d'échelles, de questionnaires, ou de comportements observables, pour être évalués. Par ces moyens, nous inférons la présence, l'intensité et les variations de ces réalités psychologiques en nous basant sur des comportements observables, des réponses spécifiques à des stimuli ou des performances à des tâches standardisées. Ces réalités psychologiques observées constituent les indicateurs opérationnalisés en items sélectionnés pour capter le construit. Par exemple, les aptitudes mnésiques se rapportent à un construit agissant comme un facteur latent, susceptible d'expliquer

les performances variées à un ensemble d'items destinés à mesurer cette aptitude. Le facteur latent représente ici l'aptitude mnésique, tandis que les réponses aux items constituent les variables manifestes observables. Un item tel que la tâche consistant à faire répéter aux enfants des mouvements de main, utilisé dans le KABC-2, illustre ce principe en cherchant à capter les construits sous-jacents d'empan mnésique et de mémoire visuelle.

Prenons un autre exemple : l'humeur dépressive, qu'elle soit considérée comme un trait ou un état, est un construit psychologique aussi. Agissant en facteur latent, elle détermine les réponses aux items destinés à la mesurer. Ainsi, l'humeur dépressive réfère au facteur latent, tandis que les réponses aux items (e.g., symptômes) constituent les variables manifestes/mesurées/observées.

En analyse factorielle et en psychométrie, les termes « facteur », « facteur latent », « variable latente » et « construit latent » sont souvent employés de manière interchangeable. Un facteur (ou facteur latent) représente une variable latente sous-jacente qui incarne un construit théorique, reflétant ainsi les dimensions abstraites (sous-construits) que le modèle cherche à mesurer.

Au sens statistique, un facteur en analyse factorielle est une variable latente dont la fonction est de capter et expliquer la variance partagée entre un ensemble de variables manifestes/mesurées/items.

La première phrase de l'ouvrage de Reuchlin (1964) intitulé « Méthodes d'analyse factorielle à l'usage des psychologues », qui inspira également le titre de ce livre énonce : « Les observations étudiées par le psychologue prennent presque toujours la forme d'un ensemble de variables associées : la modification de la valeur de l'une permet généralement de prédire, dans une certaine mesure, la modification correspondante de l'autre » (p. 1). Ici, Reuchlin s'inscrit dans la psychologie corrélationnelle de Spearman (1904). Sous cet angle-là, l'association entre deux variables peut résulter de : (a) une relation aléatoire entre ces deux variables, (b) une relation causale où une variable cause l'autre, ou (c) une relation où une troisième variable est la cause commune des deux premières.

En se basant sur cette troisième possibilité, l'analyse factorielle postule que les associations entre les variables observées/mesurées peuvent être expliquées par un nombre réduit de variables latentes appelées facteurs sans lesquels ces variables observées seraient indépendantes les unes aux autres. Il s'agit ici de ce que l'on appelle communément l'indépendance locale (Vermunt & Magidson, 2004). En effet, en analyse factorielle et dans d'autres modèles impliquant des variables latentes (e.g, les classes latentes), l'hypothèse de base est que les variables mesurées n'ont pas d'influence directe les unes sur les autres et que toute relation entre elles doit être expliquée par la variable latente sous-jacente. Autrement dit, le facteur latent est considéré comme la vraie source de covariation (facteur commun) entre les variables mesurées. Ces variables sont donc statistiquement indépendantes

conditionnellement aux facteurs latents (voir Chapitre 2). Il est clair que l'analyse factorielle est une approche centrée sur les variables en corrélation (ce que les variables peuvent avoir en commun, c'est le facteur latent commun) contrairement à l'analyse en classes latentes qui est une approche centrée sur les personnes (ce que les personnes peuvent avoir en commun constitue la classe latente ou le profil latent commun) (voir Gana et al., 2022).

2. Modèle de mesure réflectif versus modèle de mesure formatif

Dans les exemples précédents, les variables mesurées/items sont considérées comme des indicateurs de la variable latente. Toute mesure psychologique (e.g., test) effectuée dans ce contexte suit un modèle de mesure, spécifiquement un modèle de mesure réflectif. Dans un modèle de mesure réflectif les indicateurs/items sont considérés comme des manifestations du construit latent. À l'inverse, dans un modèle de mesure formatif, les indicateurs contribuent à la formation du construit latent.

Les figures 1.1 et 1.2 offrent une représentation diagrammatique de ces modèles. Le modèle réflectif est illustré par la figure 1.1, où les flèches orientées partent des variables latentes (représentées sous forme ronde ou ovale, conformément aux conventions) vers les variables manifestes/mesurées/observées/items (représentées par des rectangles ou des carrés, selon les mêmes conventions). Dans ce modèle, les items/indicateurs/variables mesurées (dep1, dep2, et dep3), conçus pour capter l'humeur dépressive, sont influencés par le construit latent « humeur dépressive ». Ainsi, la position d'un individu sur le facteur latent prédit sa position sur les indicateurs. Supposons que l'item 'dep1' concerne la fréquence à laquelle le participant entretient des idées suicidaires. La réponse à cet item, sur une échelle de 0 (très rarement) à 5 (très souvent), dépend de son humeur dépressive, c'est-à-dire de sa position sur le facteur latent « humeur dépressive ». En partant de l'hypothèse que cet indicateur est un bon représentant du construit latent (on touche ici à la validité de l'item), on peut inférer que plus le participant est déprimé, plus ses idées suicidaires sont fréquentes[1]. Plus sa position sur le facteur latent (« humeur dépressive ») est élevée plus sa réponse à l'item tend vers le score le plus élevé.

1. L'approche psychométrique en réseaux se veut une alternative au modèle réflexif. Dans ce cadre, les réponses aux items sont considérées comme des représentants/indicateurs (proxies) de variables qui interagissent directement entre elles. Par exemple, dans le cas des symptômes/indicateurs de la dépression, ils sont envisagés comme formant des réseaux de variables interconnectées et mutuellement renforcées. Ainsi, les troubles du sommeil peuvent conduire à une perte d'énergie, qui à son tour peut engendrer une faible estime de soi, entraînant des pensées obsédantes qui, en retour, peuvent aggraver les troubles du sommeil (Eskamp et al. 2018).

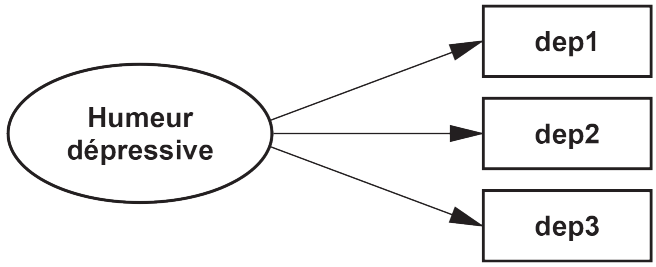

Figure 1.1 : Modèle de mesure refléctif (i.e., reflétant l'effet de variable latente Humeur Dépressive sur les 3 indicateurs, dep1, dep2 et dep3)

D'un point de vue statistique, les flèches orientées représentent les régressions linéaires, appelées saturations factorielles, obtenues lorsqu'une variable observée est régressée sur le facteur latent. L'analyse factorielle s'appuie sur des modèles linéaires pour expliquer les corrélations entre les variables manifestes en les exprimant comme combinaisons linéaires d'un nombre plus réduit de variables latentes, appelées « facteurs » (voir chapitre 2).

Pour autant, il n'est pas illogique — et souvent même plus intuitif — de concevoir que les indicateurs puissent définir un construit latent. Prenons l'exemple du statut socio-économique, qui pourrait être influencée par plusieurs variables tels que le niveau d'éducation, le revenu annuel, la profession. Cette approche est illustrée dans la Figure 1.2, où les indicateurs sont vus comme des déterminants (causaux) du facteur latent, plutôt que comme de simples effets. L'orientation des flèches suggère une relation causale présumée entre les indicateurs mesurés et la variable latente et semble souvent plus plausible que le modèle où les indicateurs sont simplement influencés par le facteur latent. Ce cadre est caractéristique d'un modèle de mesure formatif.

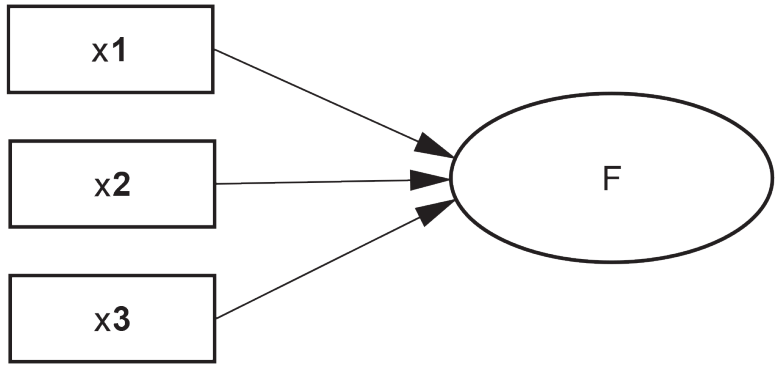

Figure 1.2 : Modèle de mesure formatif avec indicateurs causaux

Bien qu'elle n'en soit pas à l'origine, l'analyse factorielle s'inscrit dans le paradigme du modèle de mesure réflectif, et se situe principalement dans une approche explicative et confirmatoire (épreuve d'hypothèses) plutôt que dans une approche condensative et exploratoire des données observées par le psychologue (voir chapitre 4). L'histoire de l'analyse factorielle témoigne de son épanouissement et de ses avancées au sein du domaine de la métrologie psychologique, où elle a joué un rôle central dans la création et l'amélioration d'outils de mesure.

Le présent ouvrage s'inscrit dans cette même tradition, avec pour objectif principal d'offrir une synthèse rigoureuse et didactique des fondements de l'analyse factorielle exploratoire, en l'occurrence l'analyse en facteurs communs et uniques.

Chapitre 2.
Bases mathématiques et matricielles de l'analyse factorielle exploratoire

L'analyse factorielle transforme une matrice de mesures (i.e., variables observées), correspondant au tableau des données brutes, ainsi que la matrice de corrélations entre ces mesures, en deux nouvelles matrices : celle, la plus importante, des saturations factorielles, qui reflète l'association des mesures avec les facteurs communs extraits, et celle, moins importante car facultative, des scores factoriels, représentant les résultats individuels (e.g., de chaque individu) dans ces facteurs latents. Ces transformations des données en matrice de saturations et en matrice de scores factoriels reposent sur des calculs mathématiques, notamment l'algèbre matricielle.

Bornons-nous ici à la matrice des saturations factorielles. Lorsqu'on applique une analyse factorielle des psychologues, l'objectif principal est de décomposer un vecteur de n variables mesurées, $Y = (Y_1, ..., Y_n)$' en une combinaison linéaire de k facteurs latents communs, $F = (F_1, ..., F_k)$', plus un facteur unique propre à chaque variable observée Y, $u = (u_1,...,u_n)$', où $k < n$. L'apostrophe (') dans cette notation indique la transposition des vecteurs Y, F et u. Ainsi, cette décomposition de Y peut être exprimée par l'équation suivante :

$$Y = \Lambda F + u \tag{2.1}$$

où Λ est une matrice de saturations factorielles de taille $n \times k$, et u représente un vecteur des facteurs uniques propres à chaque variable observée Y, et qui sont supposés être indépendants les uns les autres, mais indépendants aussi des facteurs latents communs F.

Lorsque les facteurs latents communs F sont supposés être non corrélés entre eux (i.e., orthogonaux), la matrice de corrélations entre les variables mesurées Y de l'échantillon est donnée par :

$$R = \Lambda\Lambda' + \Psi \tag{2.2}$$

où Λ' est la transposée de la matrice des saturations factorielles (désignée aussi par Λ^T), et Ψ est la matrice des corrélations entre les facteurs uniques u, qui est supposée être diagonale car, étant indépendants les uns des autres, les corrélations entre ces facteurs uniques sont toutes nulles (voir Tableau 2.1).

En revanche, si les facteurs latents communs F sont corrélés entre eux (i.e., obliques), leur matrice de corrélations Φ est une matrice symétrique de taille $k \times k$, et la matrice

de corrélations entre les variables mesurées Y de l'échantillon se résume alors comme suit :

$$R = \Lambda\Phi\Lambda' + \Psi \tag{2.3}$$

Des méthodes d'extraction des facteurs, également appelées méthode d'estimation des facteurs sont nécessaires pour calculer et obtenir la matrice des saturations factorielles. Ces méthodes d'extraction désignent les techniques statistiques utilisées pour extraire, c'est-à-dire identifier et estimer, un ensemble de facteurs latents susceptibles d'expliquer au mieux les corrélations entre les variables mesurées (la matrice R dans les équations 2.2 et 2.3). Les saturations factorielles, élément central d'une solution factorielle, peuvent être définies, pour simplifier ici notre propos, comme les coefficients qui traduisent la force et la direction du lien entre une variable mesurée et un facteur latent extrait. Ici, on parle de facteurs « extraits » pour désigner ces facteurs latents identifiés par une méthode d'extraction, que nous présenterons dans le chapitre 5. Nous verrons également dans ce même chapitre comment calculer les saturations factorielles, les facteurs uniques ainsi que les scores factoriels en utilisant l'algèbre matricielle.

L'expression matricielle du modèle en facteurs communs et uniques, telle que décrite dans l'équation 2.3, est illustrée dans le Tableau 2.1. Il s'agit ici d'un modèle fictif à deux facteurs communs (F_1 et F_2) corrélés (Φ) à partir d'une matrice de six variables en corrélation (matrice R).

On voit bien que l'équation 2.3 fait appel à des opérations d'algèbre matricielle – multiplication, transposition et addition – dont nous allons présenter brièvement les concepts fondamentaux ainsi que le vocabulaire associé. Le lecteur a la possibilité de passer ce chapitre et d'y revenir ultérieurement pour une meilleure compréhension des détails exposés dans les sections consacrées à l'application pratique de l'analyse factorielle.

Comme le montre le Tableau 2.1, une matrice est une disposition de nombres ou d'autres éléments, organisés en lignes et en colonnes. Les matrices peuvent être, entre autre, rectangulaires lorsque le nombre de lignes diffère du nombre de colonnes, telle que la matrice Λ, de taille 6 x 2. Elles peuvent également être carrées lorsque le nombre de lignes est égal au nombre de colonnes, comme c'est cas de la matrice de corrélations R, qui est, dans notre illustration, de dimension 6 x 6.

1. Les matrices de corrélations en analyse factorielles exploratoire

On voit, par l'équation 2.3, la place de la matrice de corrélations en analyse factorielle, et par le Tableau 2.1, sa forme qui est un tableau qui synthétise les relations bivariées entre les variables d'un ensemble de données. Il s'agit une matrice carrée symétrique où chaque élément représente le coefficient de corrélation entre deux variables. Ce

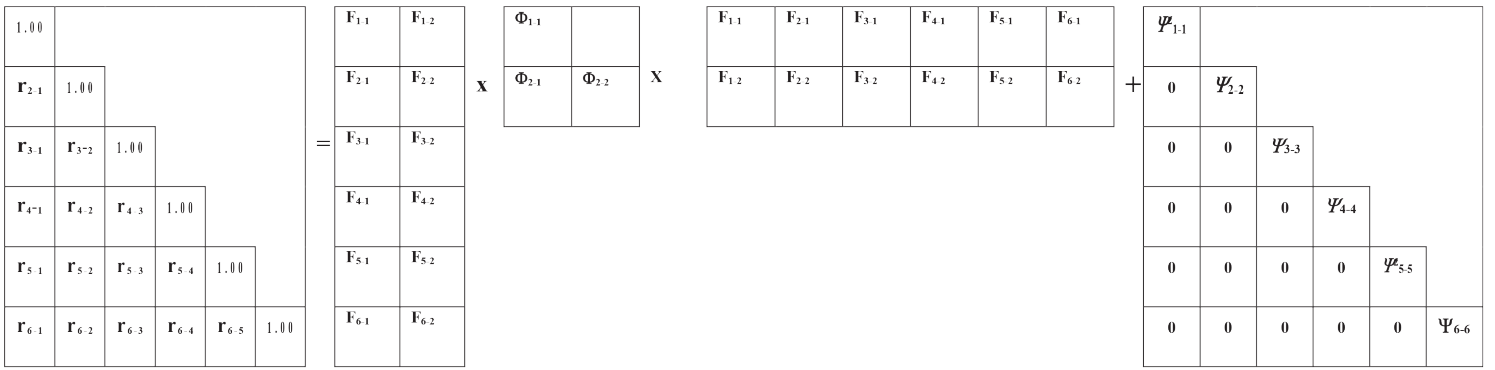

Tableau 2.1 : Le modèle en facteurs communs et uniques représenté sous forme de matrices

Note. La première matrice est la matrice de corrélations R, qui est une matrice triangulaire inférieure où la diagonale reçoit les valeurs 1.00, représentant la variance totale de la variable observée, c'est-à-dire la corrélation parfaite de cette variable avec elle-même. La seconde matrice Λ est la matrice des saturations factorielles, qui présente en colonnes les facteurs latents communs *F* de la solution (ici deux facteurs) ; la troisième matrice Φ est la matrice triangulaire inférieure des corrélations entre les facteurs latents ; la quatrième matrice Λ' est la transposée de la matrice Λ ; la dernière matrice Ψ est la matrice diagonale des facteurs uniques où la diagonale affiche la quantité de variance de la variable observée imputable au facteur unique *u* qui lui est associé.

coefficient varie entre -1 et 1 et mesure l'intensité et le sens de la relation entre les deux variables. La diagonale de cette matrice reçoit des valeurs égales à 1, car chaque variable est parfaitement corrélée avec elle-même. Car la corrélation est calculée à partir d'écarts à la moyenne mesurés en prenant, dans chaque variable, l'écart-type comme unité. Il s'agit d'une standardisation de la covariance. En voici un exemple d'une matrice de corrélations symétrique (R) d'ordre 3 x 3 :

$$R = \begin{pmatrix} 1.00 & 0.30 & 0.60 \\ 0.30 & 1.00 & 0.25 \\ 0.60 & 0.25 & 1.00 \end{pmatrix}$$

Cette matrice est dite symétrique, car la corrélation entre X et Y est identique à celle entre Y et X. Par conséquent, la corrélation pour chaque paire de variables y est affichée deux fois : une fois au-dessus et une fois en dessous de la diagonale. Toutefois, afin d'en faciliter la lecture, il est d'usage, surtout dans les publications, de n'en présenter qu'une version triangulaire (voir Tableau 2.1), où seules les valeurs situées sous la diagonale (matrice triangulaire inférieure) ou au-dessus de celle-ci (matrice triangulaire supérieure) sont affichées. En voici une illustration d'une matrice triangulaire inférieure :

$$R = \begin{pmatrix} 1.00 & & \\ 0.30 & 1.00 & \\ 0.60 & 0.25 & 1.00 \end{pmatrix}$$

En examinant cette matrice, on peut observer que la corrélation entre la première et la deuxième variable est de 0,30, celle entre la première et la troisième est de 0,60, tandis que la corrélation entre la deuxième et la troisième est de 0,25. Le coefficient de corrélation est toujours assortis d'une valeur de p (p-value) permettant d'évaluer sa significativité statistique, tandis que son intensité est appréciée au regard de seuils conventionnels tels que ceux de Cohen (1988). Nous verrons plus loin que l'examen des corrélations, tant en termes de significativité que d'intensité, constitue un préalable essentiel à l'analyse factorielle exploratoire.

Dans le cadre de l'analyse factorielle exploratoire, la matrice de corrélations dérivées des données brutes constitue le point de départ incontournable de l'analyse (voir équations 2.2 et 2.3). En effet, c'est à partir de cette matrice que nous allons chercher à identifier des structures latentes, des dimensions sous-jacentes susceptibles d'expliquer les corrélations observées entre les variables. L'objectif de l'analyse factorielle exploratoire est d'identifier les facteurs latents qui expliquent la plus grande partie de la variance des variables mesurées en corrélation. C'est précisément en analysant la structure des corrélations entre les variables que nous allons pouvoir inférer l'existence de ces facteurs. Cette matrice de corrélations entre les variables mesurées est souvent qualifiée de matrice de

corrélations observées (originales) par opposition à la matrice de corrélations reproduites à partir de la solution factorielle, et dont on parlera dans le Chapitre 5.

Il est désormais connu et admis que le calcul de la corrélation dépend du niveau de mesure des variables observées, conformément aux quatre échelles de mesure décrites par Stevens (1946), qui constituent une référence classique en psychométrie et en statistiques. Ces échelles – nominale, ordinale, d'intervalle et de rapport – ne sont pas donc interchangeables car elles diffèrent fondamentalement par leurs propriétés arithmétiques et les types d'information qu'elles véhiculent. Les échelles nominales et ordinales, dites qualitatives ou non métriques, se limitent à la catégorisation et à l'ordre, respectivement, tandis que les échelles d'intervalle et de rapport, dites quantitatives ou métriques, concernent les variables continues, permettant des opérations arithmétiques plus élaborées, avec ou sans le zéro absolu. Ainsi, ces distinctions épistémologiques imposent l'utilisation de coefficients de corrélation adaptés à la nature des données pour éviter les biais d'interprétation et assurer la validité des inférences statistiques.

Ainsi, trois principaux types de matrices de corrélations peuvent être utilisés dans l'analyse factorielle exploratoire, toutes ayant la même forme carrée et leurs diagonales recevant la valeur de 1,00.

D'abord, la matrice de corrélations de Pearson (également appelée de Bravais-Pearson), la plus courante, est utilisée pour évaluer la relation linéaire entre deux variables continues. Elle suppose une distribution normale bivariée et une relation linéaire.

Ensuite, la matrice de corrélations polychoriques est, quant à elle, spécifiquement conçue pour estimer le lien entre deux variables ordinales (ou polytomiques). Elle part du principe que ces variables ordinales sont des indicateurs d'une variable latente continue sous-jacente. L'utilisation des variables ordinales est particulièrement répandue en psychométrie, notamment à travers les échelles de type Likert, qui sont largement employées pour évaluer des attitudes, des traits et des états psychologiques sur des continuums (par exemple de « pas du tout vrai » à « tout à fait vrai »).

Enfin, la matrice de corrélations tétrachoriques est utilisée pour évaluer la relation entre deux variables binaires (ou dichotomiques). À l'instar de la matrice de corrélations polychoriques, elle suppose que chaque variable binaire est issue d'une variable continue sous-jacente et qu'une coupure arbitraire (seuil) transforme cette variable continue en une variable avec seulement deux catégories (dichotomique). Dans le domaine de la psychométrie, l'utilisation de variables binaires est particulièrement courante, notamment dans les tests cognitifs, où l'on évalue des résultats tels que la réussite (1) ou l'échec (0) à une tâche.

Ainsi, grâce à l'utilisation de matrices de corrélations spécifiques à chaque type de variable (Pearson pour les variables continues, polychoriques pour les variables ordinales, et tétrachoriques pour les variables dichotomiques), l'analyse factorielle exploratoire permet de mieux saisir les relations latentes sous-jacentes aux données. Cette approche augmente non seulement la précision des saturations factorielles, mais

renforce également la robustesse des résultats, en tenant compte des particularités des données analysées.

Cependant, on précisera ici que l'utilisation simultanée de plusieurs types de coefficients de corrélation dans la même analyse factorielle a toujours été possible (Reuchlin, 1964). En effet, lorsque les données destinées à l'analyse factorielle comprennent simultanément des variables continues, ordinales et dichotomiques — formant ainsi un ensemble de données mixtes — il est impératif d'employer plusieurs types de coefficients de corrélation adaptés à la nature des variables. Plus précisément, les coefficients de corrélation suivants doivent impérativement être convoqués : la corrélation de Pearson pour les variables continues, la corrélation polychorique pour les variables polytomiques, la corrélation tétrachorique pour les variables dichotomiques, la corrélation bisériale pour les liens entre variables continues et dichotomiques, ainsi que la corrélation polysériale pour les relations entre variables continues et ordinales.

D'ailleurs, pour faciliter cette approche, le package *psych* offre la fonction 'mixedCor', qui agit essentiellement comme un conteneur (wrapper) pour les différentes fonctions de corrélation disponibles (Pearson, polychoriques, tétrachoriques, polydichotomiques). Cette fonction commence d'abord par identifier les types de variables présentes dans l'ensemble de données, et les organise par type (continu, polytomique, dichotomique). Elle convoque ensuite la fonction de corrélation appropriée pour chaque type de variable, puis combine les matrices de corrélation résultantes en une seule matrice unifiée. Cette matrice constitue une base pour entreprendre une analyse factorielle sur des données mixtes. Le mélange de ces différents types de coefficients dans la même analyse factorielle présente toutefois certaines limites, dont la présentation dépasse le cadre de ce livre.

2. Notions d'algèbre matricielle

D'abord, l'algèbre matricielle est une branche des mathématiques qui étudie les propriétés et les opérations sur les matrices, lesquelles constituent des structures fondamentales dans de nombreux domaines de l'analyse des données, y compris l'analyse factorielle exploratoire. Les matrices sont généralement utilisées pour représenter des systèmes d'équations linéaires, effectuer des transformations linéaires qui facilitent le calcul et l'analyse. Ensuite, en analyse factorielle exploratoire, l'algèbre matricielle permet de manipuler des matrices, comme les matrices de corrélations, et de résoudre les équations nécessaires à l'extraction des facteurs latents. Enfin, l'algèbre matricielle inclut plusieurs opérations fondamentales utilisées pour manipuler et analyser les matrices. Nous présenterons ici celles qui sont indispensables pour expliciter certaines procédures de calculs utilisées par la méthode d'analyse factorielle, afin d'en faciliter la compréhension, et éviter la superficialité à notre exposé.

2.1 Matrice diagonale et matrice d'identité

Une matrice diagonale est une matrice carrée de *n* lignes et *n* colonnes, notée D_n, dans laquelle tous les éléments en dehors de la diagonale principale sont égaux à zéro. Les valeurs sur la diagonale principale peuvent être n'importe quelle valeur, pas forcément 1.00. La matrice Ψ dans le Tableau 2.1 est une matrice diagonale. Voici, en guise d'illustration, une autre matrice diagonale D_3 :

$$D = \begin{pmatrix} 0.8 & 0 & 0 \\ 0 & 0.5 & 0 \\ 0 & 0 & 0.2 \end{pmatrix}$$

Une matrice identité de dimension *n* x *n*, notée I_n, est aussi une matrice carrée de *n* lignes et *n* colonnes. Dans cette matrice la diagonale reçoit des 1.00, tandis que les autres éléments de la matrice sont des valeurs nulles (zéro). Il s'agit d'une matrice diagonale particulière dont la diagonale reçoit obligatoirement des 1.00. En voici, en guise d'exemple, une matrice d'identité I_3 :

$$I = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Il est clair ici que si notre matrice de corrélations s'apparente, même vaguement, à une matrice d'identité, signifiant l'absence de tout lien entre les variables, il est inutile de la soumettre à une analyse factorielle. En effet, il est impossible de dégager des facteurs communs explicatifs à partir de liens inexistants dans cette matrice ; le vide ne peut engendrer que du néant.

2.2 Multiplication de Matrices

La multiplication de matrices est une opération qui combine deux matrices pour en produire une nouvelle. Si *A* est une matrice $m \times n$ (avec *m* lignes et *n* colonnes) et *B* est une matrice $n \times p$ (*n* lignes et *p* colonnes) leur produit $C = AB$ sera une matrice $m \times p$ (*m* lignes *et p* colonnes). Voici une illustration numérique simple.

Si *A* est une matrice 2 × 3 (2 lignes, 3 colonnes) :

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix}$$

Si B est une matrice 3 × 2 (3 lignes, 2 colonnes) :

$$B = \begin{pmatrix} 7 & 8 \\ 9 & 10 \\ 11 & 12 \end{pmatrix}$$

Leur produit $C = AB$ sera une matrice 2 × 2 (2 lignes, 2 colonnes) comme suit :

$$C = \begin{pmatrix} (1.7 + & 2.9 + & 3.11) & (1.8 + & 2.10 + & 3.12) \\ (4.7 + & 5.9 + & 6.11) & (4.8 + & 5.10 + & 6.12) \end{pmatrix} = \begin{pmatrix} 58 & 64 \\ 139 & 154 \end{pmatrix}$$

Le lecteur aura sans doute noté que l'équation 2.3 implique une multiplication matricielle.

2.3 Inversion de Matrices

L'inversion d'une matrice constitue l'une des opérations les plus complexes en algèbre matricielle lorsqu'elle est réalisée manuellement. Contentons-nous donc de savoir que pour inverser une matrice carrée, nous pouvons utiliser plusieurs méthodes, parmi lesquelles la méthode de Gauss-Jordan (également appelée méthode de Pivot de Gauss) qui utilise la matrice d'identité pour calculer l'inverse d'une matrice, et la méthode classique des cofacteurs (ou comatrice), et dont voici la formule :

$$A^{-1} = \frac{1}{det(A)} \cdot (Cof(A))T \tag{2.4}$$

Où A^{-1} est l'inverse de la matrice carrée A ; det (A) est le déterminant de la matrice A ; Cof $(A)^T$ la transposée de la matrice des cofacteurs de A (appelée aussi comatrice). Cette matrice transposée est appelée matrice adjointe.

En empruntant cette méthode des cofacteurs, voici les étapes pour calculer l'inverse :

- Calculer le déterminant det(A).
- Construire la matrice des cofacteurs de A (comatrice).
- Transposer cette matrice des cofacteurs pour obtenir la matrice adjointe.
- Multiplier la matrice adjointe par 1/det(A).

On voit, par ce qui précède, que donner un exemple numérique illustrant le calcul de l'inverse d'une matrice alourdirait inutilement notre présentation. Parce que, d'abord, le lecteur trouvera dans les ouvrages de vulgarisation de l'algèbre matricielle des méthodes détaillées pour effectuer ce calcul ; et qu'ensuite, il pourra utiliser la fonction de base 'solve ()' dans R pour inverser une matrice.

Ici, on se bornera à préciser que l'inversion matricielle est nécessaire pour résoudre des systèmes d'équations linéaires et pour diverses étapes de l'analyse factorielle exploratoire, comme la transformation des données, l'estimation des paramètres du modèle, et le calcul des scores factoriels. Par exemple, lors de l'estimation des scores factoriels (voir section 5.8 plus loin), l'inversion de la matrice de corrélations entre les variables observées permet de tenir compte des relations entre ces variables, afin d'ajuster les scores factoriels calculés pour chaque individu.

2.4 Transposition de Matrices

La transposition d'une matrice A, notée A^{T} ou A' échange ses lignes avec ses colonnes, et vice versa (on parle de la transposée d'une matrice). Voici une illustration numérique.

Prenons la matrice suivante A :

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix}$$

La transposée de cette matrice s'obtient en échangeant les lignes et les colonnes. La première ligne de A devient la première colonne de A^{T}, la deuxième ligne devient la deuxième colonne, et ainsi de suite. La transposée A^{T} est donc :

$$A^{T} = \begin{pmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{pmatrix}$$

Le lecteur aura sans doute remarqué que l'équation 2.3 implique aussi une transposition matricielle. Cette opération de transposition est fréquemment utilisée pour restructurer et manipuler les données, notamment dans le cadre de traitements algébriques en analyse factorielle. Elle permet de préparer les matrices pour des calculs ultérieurs, tels que le calcul de la matrice des corrélations reproduites, un concept que nous aborderons en détail dans le chapitre 5.

2.5 Déterminant d'une matrice

Le déterminant d'une matrice carrée est un scalaire (un nombre) associé à cette matrice, qui fournit des informations importantes sur ses propriétés. Il est noté det (A) où A est la matrice carrée.