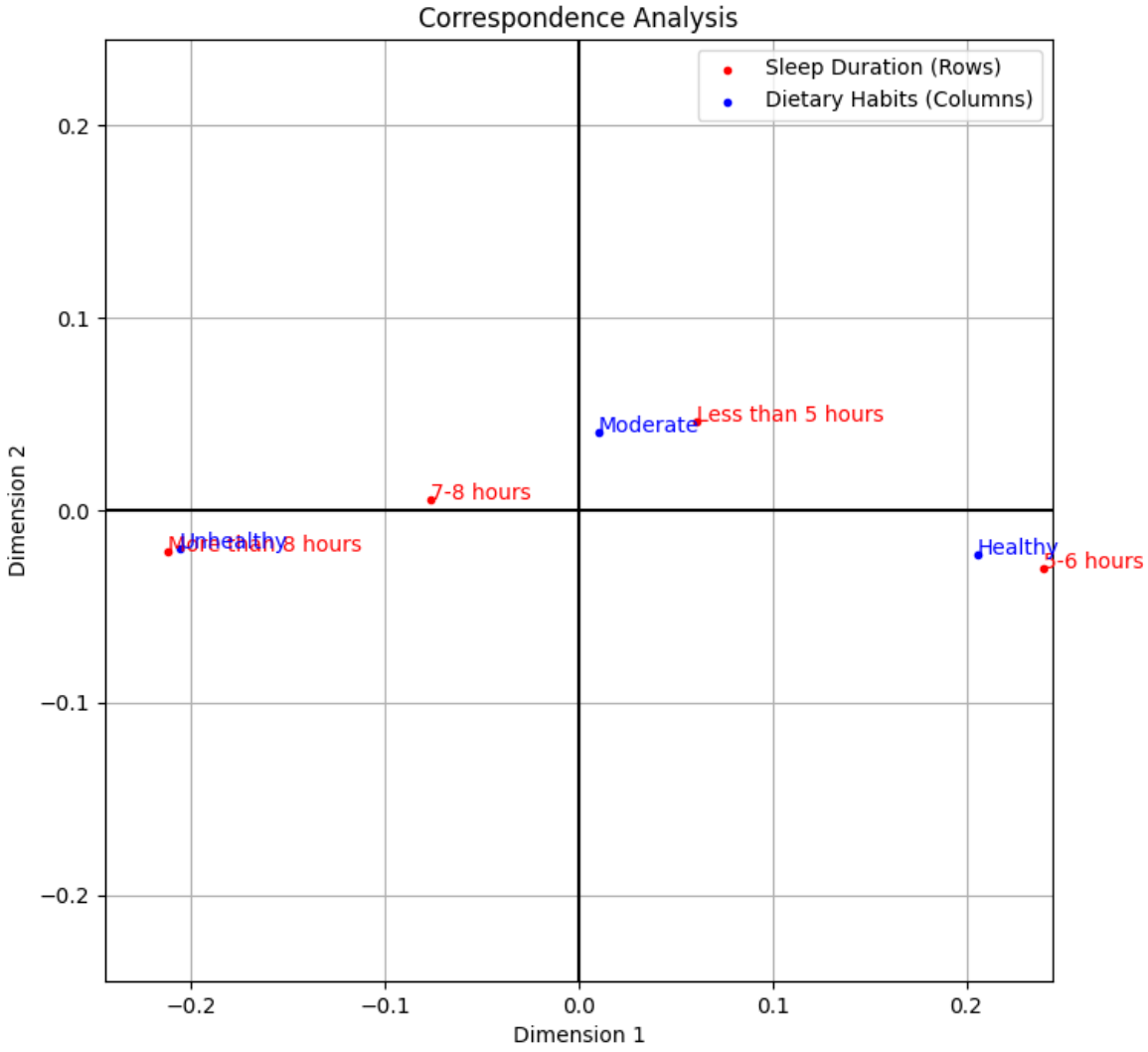
ENSEIRB-MATMECA

Département informatique

EI5IS102 - Traitement de l'information

RAPPORT DE PROJET

Etude d'un jeu de données à l'aide de l'AFC



Bourgouin Raphaël Dumarchat Joan Gouedard Anna

Année universitaire 2023-2024

Table des matières

Table des figures

1 Présentation de la méthode utilisée : AFC

L'Analyse Factorielle de Correspondance est une technique permettant d'analyser des données qualitatives. Plus précisément, elle permet d'analyser les relations entre 2 variables qualitatives catégorielles. Un représentation commune de ces données est la table de contingence, pour lequel un exemple est fourni dans la figure 1.

	Sleep duration				
Study satisafaction	Less than 5 hours	5-6 hours	7-8 hours	More than 8 hours	Total
1.0	23	19	20	24	86
2.0	25	25	25	25	100
3.0	19	25	33	26	103
4.0	29	29	31	27	116
5.0	27	25	19	26	97
Total	123	123	128	128	502

FIGURE 1 – Exemple de table de contingence

1.1 Notation

Dans la suite, on notera n le nombre total d'instances, V_1 la première variable (de taille I), V_2 la seconde (de taille J) et x_{ij} le nombre d'instances étant dans la catégorie i de la variable V_1 et dans la catégorie j de la variable V_2. On définit alors $X = (x_{ij})_{1\leq i\leq I, 1\leq j\leq J}$ la table de contingence. On peut alors parler des valeurs marginales des lignes et colonnes, dont les formules sont :

$$x_{i\bullet} = \sum_{j=1}^{J} x_{ij} \qquad x_{\bullet j} = \sum_{i=1}^{I} x_{ij} \tag{1}$$

Cepandant, on préfèrera toujours travailler sur la table des probabilités définie par $f_{ij} = \frac{x_{ij}}{n}$, pour lesquelles on peut aussi définir les valeurs marginales avec :

$$f_{i\bullet} = \sum_{j=1}^{J} f_{ij} \qquad f_{\bullet j} = \sum_{i=1}^{I} f_{ij} \tag{2}$$

1.2 Test d'indépendance

On souhaite dans un premier temps vérifier si les variables V_1 et V_2 sont indépendantes, ce qui est le cas si $\forall i, j, f_{ij} \approx f_{i,\bullet} f_{\bullet,j}$, on définit alors $\hat{f}_{ij} = f_{i,\bullet} f_{\bullet,j}$ la probabilité théorique sous l'hypothèse d'indépendance des variables. De manière similaire, on définit $\hat{x}_{ij} = n\hat{f}_{ij}$ les données théoriques sous l'hypothèse d'indépendance.

On peut alors procéder au test d'indépendance χ^2, qui consiste à :

- Calculer la distance $\chi^2_{obs} = \sum_{(i,j)} \frac{(x_{ij}-\hat{x}_{ij})^2}{\hat{x}_{ij}}$
- Fixer une p-value (usuellement à 0.05)
- Calculer le degré de liberté $df = (I-1)(J-1)$
- Déterminer $\chi^2_{critical}$ à l'aide d'un table
- Si $\chi^2_{obs} \leq \chi^2_{critical}$ alors les variables sont indépendantes, sinon elles sont corrélées

1.3 Diagonalisation

Maintenant que l'on a confirmé que les variables sont corrélées, nous pouvons utiliser une technique plus précise afin d'obtenir plus d'informations, à savoir ici la diagonaliser la matrice des probabilités.

On suppose ici qu'étudier F reviens à étudier $\tilde{F} = D_I F D_J$ avec $D_I = diag(\frac{1}{\sqrt{f_{1\bullet}}}, \dots, \frac{1}{\sqrt{f_{I\bullet}}})$ et $D_J = diag(\frac{1}{\sqrt{f_{\bullet 1}}}, \dots, \frac{1}{\sqrt{f_{\bullet J}}})$. On réalisera 2 diagonalisations, une sur $\tilde{F}\tilde{F}^T$ pour étudier les lignes et une sur $\tilde{F}^T\tilde{F}$ pour étudier les colonnes. On représente ensuite chaque catégorie de V_1 et V_2 dans un cercle des corrélations à la manière de l'ACP. Une propriété importante ici est le fait que les valeurs propres issues des 2 diagonalisations sont identiques, ce qui va nous permettre de représenter sur le même cercle les 2 variables. Dans ce graphique, si une catégorie de V_1 et de V_2 sont proches tout en étant éloignées de l'origine, cela montrera une corrélation entre cess deux catégories.

On pourra aussi calculer la contribution de chaque ligne/colonne dans les composantes principales afin de mieux interpréter les résultats.