CHAPITRE 9
ANALYSE CANONIQUE

Plan

1. Les données
2. Le principe de l'analyse canonique
3. Les représentations graphiques
4. Une application de l'AC
5. Les cas particuliers de l'AC
6. La critique de l'AC

1. Les données

• 2 tableaux de données :
– le tableau X_1 à n lignes et m_1 colonnes
– le tableau X_2 à n lignes et m_2 colonnes
Pour chacun des 2 tableaux, la ligne i représente l'individu i.

• Les colonnes de X_1 et de X_2 sont constituées soient
– par des variables quantitatives centrées
– par les modalités de variables qualitatives

Remarque : On suppose, dans tout ce chapitre et comme dans les deux précédents, que pour chacun des 2 tableaux, les colonnes sont linéairement indépendantes

2. Le principe de l'analyse canonique

Le principe de l'analyse canonique :

Mettre en évidence des proximités entre deux ensembles de données. Ce qui se traduit par décrire ces proximités entre les deux tableaux de variables que nous avons à notre disposition.

2.1. Les composantes canoniques

L'algorithme de l'analyse canonique :

• **A la première étape :** L'algorithme détermine un couple de variables canoniques (z_1^1, z_2^1) tel que le coefficient de détermination

$$R^2(z_1^1 z_2^1) = \frac{Cov^2(z_1^1, z_2^1)}{\text{Var}[z_1^1]\text{Var}[z_2^1]}$$

ait une valeur maximale sous la contrainte

$$\text{Var}[z_1^1] = \text{Var}[z_2^1] = 1.$$

Remarques :

1. • z_1^1 est donc le premier vecteur propre de P_1P_2
   • z_2^1 est donc le premier vecteur propre de P_2P_1,

   où
   – P_1 désigne le projecteur orthogonal sur l'espace engendré par les colonnes de X_1 et
   – P_2 désigne le projecteur orthogonal sur l'espace engendré par les colonnes de X_2.
2. • z_1^1 est une combinaison linéaire des variables du tableau X_1
   • z_2^1 est une combinaison linéaire des variables du tableau X_2

Mais que valent les projecteurs P_1 et P_2 ?

Rappel d'algèbre linéaire

Définition : Le projecteur orthogonal sur l'espace E engendré par les colonnes de X est l'application linéaire qui fait correspondre à u sa projection orthogonale sur E. Ce projecteur s'écrit

$$P = X(X'X)^{-1}X'.$$

Remarques :

3. Ces deux vecteurs propres z_1^1 et z_2^1 sont associés à la même valeur propre qui est égale au coefficient de détermination

$$R^2(z_1^1, z_2^1).$$

4. • z_1^1 est la première composante canonique du tableau X_1
   • z_2^1 est la première composante canonique du tableau X_2

– **A la k-ième étape :** L'algorithme détermine un k-ième couple de variables canoniques (z_1^k, z_2^k) tel que le coefficient de détermination

$$R^2(z_1^k z_2^k) = \frac{Cov(z_1^k, z_2^k)}{\text{Var}[z_1^k]\text{Var}[z_2^k]}$$

ait une valeur maximale sous la contrainte

$$\text{Var}[z_1^k] = \text{Var}[z_2^k] = 1.$$

et pour $r < k$

$$R(z_1^r, z_1^k) = 0 \quad \text{et} \quad R(z_2^r, z_2^k) = 0$$

Remarques :

1. • z_1^k est donc le k-ième vecteur propre de P_1P_2
   • z_2^1 est donc le k-ième vecteur propre de P_2P_1,

   où
   – P_1 désigne le projecteur orthogonal sur l'espace engendré par les colonnes de X_1 et
   – P_2 désigne le projecteur orthogonal sur l'espace engendré par les colonnes de X_2.
2. • z_1^k est une combinaison linéaire des variables du tableau X_1
   • z_2^k est une combinaison linéaire des variables du tableau X_2

Remarques :

3. Ces deux vecteurs propres z_1^k et z_2^k sont associés à la même valeur propre qui est égale au coefficient de détermination

$$R^2(z_1^k, z_2^k).$$

4. • z_1^k est la k-ième composante canonique du tableau X_1
   • z_2^1 est la k-ième composante canonique du tableau X_2

Deux dernières remarques très importantes :

1. Les composantes canoniques d'un même tableau sont donc deux à deux non corrélées.
2. La composante canonique d'ordre k d'un tableau est non corrélée avec les composantes canoniques d'ordre différent de k de l'autre tableau.

2.2. Les facteurs

A la k-ième étape : z_i^k est une combinaison linéaire des variables du tableau i $(i = 1, 2)$:

$$z_1^k = X_1 a_1^k$$

et

$$z_2^k = X_2 a_2^k,$$

où a_1^k et a_2^k sont les facteurs d'ordre k.

Les facteurs a_1^k et a_2^k possèdent les propriétés suivantes :

• Les facteurs a_1^k sont solutions de

$$V_{11}^{-1}V_{12}V_{22}^{-1}V_{21}a_1^k = R^2(z_1^k, z_2^k)a_1^k,$$

où V_{ij} désigne la matrice

$$V_{ij} = \frac{1}{n}(X_i)'(X_j).$$

• Les facteurs a_1^k sont solutions de

$$V_{22}^{-1}V_{21}V_{11}^{-1}V_{12}a_2^k = R^2(z_1^k, z_2^k)a_2^k,$$

où V_{ij} désigne la matrice

$$V_{ij} = \frac{1}{n}(X_i)'(X_j).$$

Les relations qui existent entre a_1^k et a_2^k sont :

$$V_{11}^{-1}V_{12}a_2^k = R(z_1^k, z_2^k)a_1^k$$

et

$$V_{22}^{-1}V_{21}a_1^k = R(z_1^k, z_2^k)a_2^k$$

• **Dans le cas où les deux variables sont quantitatives**, V_{ij} est la matrices des covariances entre les variables du tableau i et celles du tableau j.

• **Dans le cas où les deux variables sont qualitatives**, V_{ij} est la matrice des fréquences relatives des variables du tableau i et celles du tableau j.

2.3. Le nombre maximal d'étapes

• Une base de l'espace engendré par les colonnes du tableau X_1 est constitué par m_1 vecteurs linéairement indépendants.

• A chaque étape, l'AC détermine dans chaque espace une variable canonique non corrélée aux variables canoniques précédemment déterminées dans cet espace.

• **Le nombre maximal d'étapes** est le plus petit des nombres m_1 et m_2.

2.4. Les proximités entre les individus

• L'AC détermine z_1^k et z_2^k telles qu'en moyenne les 2 variables soient le plus proches possibles pour les n individus, c-à-d de telle manière que

$$\frac{1}{n}\sum_{i=1}^{n}(z_{1i}^k - z_{2i}^k)^2$$

soit le plus petit possible, sous les mêmes contraintes que dans l'espace des variables.

2.5. Et dans la pratique, comment fait-on ?

• On calcule la matrice des corrélations

• On extrait de la matrice des corrélations les matrices dont nous allons avoir besoin par la suite, à savoir :

$$V_{11}, V_{22}, V_{12}, V_{21}.$$

• On calcule alors les facteurs a_1^1 et a_2^1 (qui sont en fait les vecteurs propores associés aux valeurs propres de la matrice $V_{11}^{-1}V_{12}V_{22}^{-1}V_{21}$) en diagonalisant la matrice $V_{11}^{-1}V_{12}V_{22}^{-1}V_{21}$

• Enfin on calcule les premières composantes canoniques z_1^1 et z_2^1.

3. Les représentations graphiques

Rappel : Le but de l'analyse canonique est de mettre en évidence des proximités entre deux ensembles de données (slide 4).

Les représentations graphiques ont pour objectif de décrire ces proximités, aussi bien :

- pour les variables,
- que pour les individus.

3.1. La représentation des variables

• S'intéresser aux résultats de la k-ième étape

$$\Leftrightarrow$$

expliquer pourquoi la corrélation entre z_1^k et z_2^k est élevée

$$\Leftrightarrow$$

expliquer pourquoi la corrélation entre une combinaison linéaire de variables X_1 et une combinaison linéaire de variables X_2 est élevée.

• Il est donc nécessaire de faire figurer sur un même graphique l'ensemble des variables de départ.

• Cette représentation des variables se fait comme en ACP (cf Chapitre 7) à l'aide d'un cercle des corrélations.

L'axe correspondant à la j-ième étape est un compromis entre z_1^j et z_2^j. Soit

$$z^j = \frac{1}{2}(z_1^j + z_2^j).$$

L'axe correspondant à la k-ième étape est un compromis entre z_1^k et z_2^k. Soit

$$z^k = \frac{1}{2}(z_1^k + z_2^k).$$

On vérifie bien entendu que $R(z^j, z^k) = 0$!

3.2. La représentation des individus

Chacun des 2 tableaux de données décrit un nuage pour les mêmes n individus.

• La représentation des individus de l'AC permet de cerner ce qui caractérise le mieux ces nuages d'individus dans les directions pour lesquelles ces nuages sont les plus ressemblants possibles.

• De plus, la représentation des individus de l'AC permet de repérer les individus ayant un comportement particulier.

• A la j-ième étape, il s'agit de comparer la description des individus donnée par la variable canonique z_1^j à la description des individus donnée par la variable canonique z_2^j.

• Enfin la proximité plus ou moins importante entre les deux descriptions des individus peut aussi être mise en évidence en calculant l'écart résiduel qui est défini au sous-paragraphe suivant.

3.3. L'écart résiduel

• L'écart résiduel est égal, pour l'individu i à la k-ième étape, à

$$|z_{1i}^k - z_{2i}^k|.$$

Rappel : Le but de l'AC est la minimisation de la moyenne des carrés des écarts résiduels (paragraphe ?, si vous avez bien suivi depuis le début).

• Si l'écart résiduel est élevé de l'individu i pour la k-ième étape est élevé, cela signifie que cet individu joue un rôle particulier pour le phénomène mis en évidence à la k-ième étape, rôle particulier que vous devez identifier !

4. Les cas particuliers de l'AC

L'AC présente un grand intérêt d'un point de vue théorique car plusieurs techniques statistiques très utilisées en sont des cas particuliers :

Si X_1 décrit une seule variable quantitative, l'AC se ramène à :

- la RLS si X_2 est constitué d'une seule variable quantitative
- la RLM si X_2 est constitué par plusieurs variables quantitatives
- le modèle d'analyse de la variance si X_2 est une ou plusieurs variables qualitatives.
- le modèle d'analyse de la covariance si X_2 est un mélange de variables quantitatives et qualitatives.

Pour ces 4 méthodes citées ci-dessus, le problème est :

maximiser le coefficient de corrélation entre une variable quantitative X_1 et un ensemble de variables X_2.

C'est donc bien un problème d'AC.

• **L'analyse factorielle des correspondances**, étudiée au chapitre 10, est le cas particulier de l'AC pour lequel X_1 et X_2 décrivent chacun les modalités d'une variable qualitative.

• **L'analyse factorielle discriminante**, qui ne sera pas étudiée, est le cas particulier de l'AC pour lequel X_1 décrit un ensemble de variables quantitatives et X_2 une variable qualitative.

5. La critique de l'AC

• L'AC décrit les relations linéaires existant entre 2 ensembles de variables : les premières étapes mettent en évidence les directions de l'espace des variables selon lesquelles les deux ensembles sont le plus proches.

• Mais il est possible que les variables canoniques soient faiblement corrélées aux variables des tableaux X_1 et X_2. Donc elles sont difficilement interprétables.

• En effet, les variables d'origine n'interviennent pas dans les calculs de détermination des composantes canoniques, seuls interviennent les projecteurs sur les espaces engendrés par ces variables.